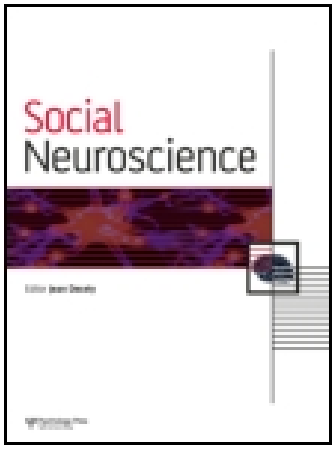

Social Neuroscience
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/psns20

Neurolaw: Differential brain activity for Black and White faces predicts damage awards in hypothetical employment discrimination cases

Harrison A. Korn [a b], Micah A. Johnson [c] & Marvin M. Chun [a d]

[a] Cognitive Science Program, Yale University, New Haven, CT, USA

[b] Yale Law School, New Haven, CT, USA

[c] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA

[d] Department of Psychology and Department of Neurobiology, Yale University, New Haven, CT, USA

Published online: 07 Nov 2011.

To cite this article: Harrison A. Korn, Micah A. Johnson & Marvin M. Chun (2012) Neurolaw: Differential brain activity for Black and White faces predicts damage awards in hypothetical employment discrimination cases, Social Neuroscience, 7:4, 398-409, DOI: 10.1080/17470919.2011.631739

To link to this article: http://dx.doi.org/10.1080/17470919.2011.631739

SOCIAL NEUROSCIENCE, 2012, 7 (4), 398–409

Neurolaw: Differential brain activity for Black and White faces predicts damage awards in hypothetical employment discrimination cases

Harrison A. Korn[1,2], Micah A. Johnson[3], and Marvin M. Chun[1,4]

[1]Cognitive Science Program, Yale University, New Haven, CT, USA
[2]Yale Law School, New Haven, CT, USA
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
[4]Department of Psychology and Department of Neurobiology, Yale University, New Haven, CT, USA

Currently, potential jurors' racial biases are measured by explicit questioning—-a poor measure because people often hide their views to adhere to social norms, and people have implicit views they are not consciously aware of. In this experiment, we investigated whether two alternative methods of measuring racial bias—a standard Black/White, good/bad Implicit Association Test (IAT) and neural activity, measured by fMRI, in response to seeing faces of Black and White individuals—-could predict how much money subjects would award Black victims in hypothetical employment discrimination cases. IAT scores failed to predict how much money subjects awarded victims. However, in right inferior parietal lobule (BA 40) and in right superior/middle frontal gyrus (BA 9/10)—which have both previously been implicated in measuring biases and implicit preferences—-the difference in neural activity between when subjects viewed Black faces paired with neutral adjectives and when subjects viewed White faces paired with neutral adjectives was positively correlated with the amount of money the subjects awarded victims. This suggests that brain activity measures racial bias with more practical validity, at least in this situation and with our sample size, than a common behavioral measure (the IAT).

Keywords: Race; Implicit Association Test (IAT); fMRI; Neurolaw; Bias; Face perception.

The sixth amendment of the US Constitution states that all accused have the right to a trial "by an impartial jury." Therefore, when selecting a jury—especially for cases in which race is a factor—courts should strive to select jurors with the least amount of racial prejudice. Currently, to try to get an impartial jury, potential jurors are sometimes explicitly asked during voir dire whether the race of the victim, plaintiff, or defendant would affect their decision-making (Arterton, 2008). This is not a very effective way to identify impartial jurors, however, as social pressures not to express unpopular or socially unacceptable beliefs may make people unlikely to admit they are racially biased (Fazio, Jackson, Dunton, & Williams, 1995; Nosek & Banaji, 2002).

Further, people have implicit biases that they are not consciously aware of and that differ from their explicit beliefs. In a 1992 dissent, US Supreme Court Justice O'Connor wrote, "[i]t is by now clear that conscious and unconscious racism can affect the way white jurors perceive minority defendants and the facts presented at their trials, perhaps determining the verdict of guilt or innocence" (*Georgia* v. *McCollum)*. Even as explicit biases have declined, the disparity in outcomes in the

Correspondence should be addressed to: Marvin M. Chun, Department of Psychology, Yale University, PO Box 208205, New Haven, CT 06520-8205, USA. E-mail: Harrison.korn@yale.edu

Thanks to Jacqueline Meadow for assisting with data collection and Yi He and Gregory McCarthy for providing the implicit association task script. This study was supported by the Yale University FAS MRI Program funded by the Office of the Provost and the Department of Psychology.

www.psypress.com/socialneuroscience http://dx.doi.org/10.1080/17470919.2011.631739

legal system between Blacks and Whites has persisted (Sniderman & Piazza, 2002). Much of the remaining disparity is likely due to implicit biases that people are unaware they hold (Greenwald & Krieger, 2006). Implicit biases are especially likely to affect decisions in trials since trials are full of ambiguity and uncertainty, conditions that cause people to use heuristics that are often based on stereotypes (Rector & Bagby, 1995).

This suggests that a test known as the Implicit Association Test (IAT), which measures implicit biases, may be a better method for finding the least biased jurors than the explicit questioning method that is currently used. The IAT (Greenwald & Banaji, 1995; Greenwald, McGhee, & Schwartz, 1998) uses response times to test how strongly a person associates two groups (for example, Whites and Blacks) with each of two categories (for example, good and bad). A study of 2.5 million IATs found that implicit biases were pervasive across demographic groups (Nosek et al., 2007).

In a review of studies involving interracial behavior, the IAT was a significantly better predictor of behavior than more explicit, self-report measures (mean $r_{IAT} = .24$, mean $r_{self\text{-}report} = .12$) (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). The behaviors that IAT scores have been found to predict are numerous. For example, scores on a racial IAT have been found to predict the likelihood that a physician will give a hypothetical White patient but not a hypothetical Black patient the appropriate treatment for a myocardial infarction (Green et al., 2007), the likelihood a hiring manager will give an interview to a Swedish applicant and not an Arab-Muslim applicant (Rooth, 2010), and the likelihood a person would vote for John McCain in the 2008 election, even taking into account conservatism (Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009). However, in the legal field, very little research has been done on the IAT (Levinson, Cai, & Young, 2010). One study shows that IAT scores can predict actual judges' verdicts in certain hypothetical scenarios. Actual judges read vignettes about a juvenile shoplifter, a juvenile robber, and a battery. When the defendant's race was subliminally primed (but not when it was explicitly stated), the judges' IAT scores predicted the sentences they gave the Black defendant (Rachlinski, Johnson, Wistrich, & Guthrie, 2009).

But if the IAT is a better predictor of certain racially discriminatory behaviors than explicit measures, then perhaps going one step further back and looking at neural activity would be an even more sensitive measure. Implicit behavioral measures like the IAT still require some conscious cognition, but fMRI measures can tap unconscious processes even further along the implicit–explicit spectrum. Since minimal need for introspective access is desired when studying discrimination and prejudice, fMRI appears to be an attractive tool.

IAT scores have been correlated with subjects' neural activity when viewing Black and White faces. The brain areas that are correlated with IAT scores unfortunately vary on parametric factors, such as how long the faces were presented. When faces are presented for short periods of time (about 30 ms) or as part of a cognitively demanding task, the difference in neural activity for Black versus White faces in emotional regions like the amygdala predicts IAT scores (Cunningham et al., 2004; Phelps et al., 2000). When faces are presented for longer periods of time with a less cognitively demanding task—allowing for more controlled processing—dorsal lateral prefrontal cortex (DLPFC) activity in response to Black faces can predict IAT scores (Richeson et al., 2003), reflecting the executive control functions of these regions (right anterior cingulate continuing into the right medial frontal gyrus and right middle frontal gyrus). Greater activity in the executive control regions in response to Black faces was correlated with more implicit bias, suggesting that more biased individuals must put forth more effort in order to suppress these biases and behave according to their explicit, non-prejudicial values (Richeson et al., 2003). Cunningham et al. (2004) found evidence that when people have time to engage in controlled processing, DLPFC activity in response to Black faces may work to override people's automatic emotional response in response to Black faces and suppress amygdala activity. As a result, amygdala activity does not correlate with IAT scores when subjects can engage in controlled processing.

Because racial bias has neural correlates, it is reasonable to hypothesize that neural activity when viewing Black and White faces might be able to predict legal judgments. We know of only one study where general legal judgments were predicted by neural activity. Buckholtz et al. (2008) found that, for a range of criminal scenarios, neural activity in regions related to affective processing predicted the magnitude of punishment the subject believed the defendant deserved, and neural activity in the right DLPFC predicted whether the subject believed the defendant was criminally responsible. This study, however, measured subjects' neural activity while they made decisions about specific cases, not neural activity designed to measure a baseline bias that would be reflected in the decisions they made outside the scanner.

In the present study, we sought to show that the difference in a subject's neural activity in response to Black and White faces can predict not only the person's score on an IAT, but also his or her decisions in hypothetical legal cases where race is salient.

Specifically, activity in DLPFC regions which have previously been shown to predict IAT scores were expected to also predict subjects' judgments in the legal cases. The fact that DLPFC regions have recently been shown to be active when people make legal decisions (Schleim, Spranger, Erk, & Walter, 2010), and that Buckholtz et al. (2008) found that DLPFC was involved in determing whether punishment was warranted, further suggests that activity for Black faces minus White faces in this region may be able to predict people's decisions in legal cases where race is salient.

We asked subjects to read short vignettes based on actual cases and say how much money, if any, they would award the victim. We chose employment discrimination cases because race is a salient factor in the cases and they are very common, accounting for over 40% of all civil, civil rights cases in US District Courts in 2010 (Administrative Office of the United States Courts, 2011). Subjects were then scanned in an fMRI while they saw Black and White faces. We paired these faces with positive, negative, and neutral adjectives to see whether pairing a White face (in contrast to a Black face) with a negative adjective or a Black face (in contrast to a White face) with a positive adjective would evoke a greater DLPFC response in more racially biased subjects that in turn would predict a smaller award. We also included neutral adjectives and planned to also collapse across all adjective conditions so that we could perform contrasts that were similar to the contrasts performed in the previously mentioned studies. However, we decided against choosing regions of interest from prior studies a priori because the IAT fMRI literature is inconsistent on what specific regions correlate with measures of racial bias. That is, given that stimulus duration and task factors influence the neural correlates and IAT, and considering that our stimuli and task had novel features, we used whole-brain analyses for a more neutral exploration of our data.

METHODS

Participants

Twenty-five right-handed subjects with normal or corrected-to-normal vision were scanned in exchange for monetary compensation. The Yale University Human Investigations Committee and Human Subjects Committee approved the experimental protocol, and informed consent was obtained from all subjects. The first three subjects scanned were excluded because of technical difficulties with the button box. One additional subject was excluded due to errors with projection equipment, and two additional subjects were excluded due to excessive motion (greater than 4 mm in one run) or low accuracy on the arrow task (less than 75% correct) in more than one run. The remaining 19 subjects (12 male, mean age = 19.7 years, age range = 18–26) were used in all data analysis. All subjects' native language was English, and all self-identified as white, non-Hispanic.

Behavioral measures

Vignettes survey

Before scanning, all subjects completed a five-question, paper-based survey to determine how they would rule in a variety of employment discrimination cases. For each question, the subject read a short vignette (see online Appendix) based on a real case and then chose one of five evenly spaced options of monetary awards, measured in months of salary, that he or she felt the person in the vignette should receive as compensation for the discrimination (0 months was always a choice if the subject felt the person was not a victim of discrimination). The questions were presented in a random order for each subject. In two of the vignettes Black females were the alleged victims, and in the remaining three Black males were the alleged victims.

The surveys were scored by assigning a point value (0–4) to each of the five choices for each question (with 0 representing an award of 0 months' salary and 4 representing the largest possible award). The mean verdict award for each vignette ranged from 2.00 to 2.47 (SD = 1.16–1.52), and the survey showed good inter-item reliability (Cronbach's alpha =.75). The subject's responses across all five questions were averaged to produce an "average verdict" variable. A large range of average verdicts was observed across subjects (range = 0.6–3.6, mean = 2.24, SD = 0.94).

Implicit Association Test (IAT)

After scanning, all subjects completed a Black-White IAT following the procedure suggested by Greenwald, Nosek, and Banaji (2003). Color pictures of five Black and five White (three male and two female in both cases) individuals from the FERET database (Phillips, Moon, Rizvi, & Rauss, 2000; Phillips, Wechsler, Huang, & Rauss, 1998) were used. In order to prolong the processing time and get a larger range of response times, pictures were cropped to show only the areas of the eyes and nose (He, Johnson, Dovidio, & McCarthy, 2009). The words and procedure used were the same as in He et al. (positive words: *happy*, *joy*, *love*, *lucky*, *peace*; negative words:

death, *devil*, *pain*, *terrible*, *war*; subjects did not need to correct errors).

The IAT was scored by the algorithm suggested by Greenwald et al., 2003 (600 ms added to error trials, trials with response times less than 400 ms removed, and block standard deviation performed including error trials). A positive IAT score means the subject more strongly associates Whites with positive words and Blacks with negative words (as measured by a lower response time when they are paired) than Whites with negative words and Blacks with positive words. In other words, a more positive IAT score indicates a pro-White bias while a more negative IAT score indicates a pro-Black bias.

Experimental task

Subjects completed four functional runs in the scanner with 30 total trials in each run. (One subject completed only three runs due to button-box difficulties, and three subjects had runs that did not contain 30 trials due to button-box difficulties. All subjects had at least 75 usable trials across at least three runs.) In each trial, a Black or White face was displayed against a black background for 500 ms followed by a text phrase (white Arial font on a black background) reading "This person is <adj>" for 1500 ms. "<adj>" was replaced by a randomly selected positive, negative, or neutral adjective. The two variables were crossed for a total of six conditions (2 races of faces $\times$ 3 types of adjectives). In each functional run, five trials of each of the six conditions were presented in a random order. Subjects were instructed to remember the pairings of each face and adjective to the best of their ability. Each face and adjective was used only once for each subject. Trials were spaced 12 s (stimulus onset was synched to scan acquisition) apart, and each run lasted 6 min 20 s (including time before the first trial and after the last trial).

The 60 Black and 60 White faces were selected from the color FERET database. All faces were head-on, had a neutral expression, and were cropped to tightly frame the face. Participants ($n = 29$) in a pilot experiment were asked to think of a person described by each of 250 adjectives and rate on a Likert scale (1 = least favorable, 7 = most favorable) how much they would like a person described by that adjective. The 250 adjectives were adjectives that Anderson (1968) found people felt very positively, very neutrally, or very negatively about. The adjectives used in the fMRI experiment were the 40 adjectives that were rated the most positively (mean rating = 5.94), the 40 adjectives that were rated the most neutrally (mean = 4.04), and the 40 adjectives that were rated the most negatively (mean = 1.81).

To ensure the subject was paying attention, subjects were asked to perform a simple task between each trial. Beginning 700 ms after the sentence was removed from the screen, an arrow pointing to the right or the left was presented (each direction was equally probable). Subjects were instructed to press the right button on the button box if the arrow was pointing to the right and the left button if the arrow was pointing to the left. Subjects were instructed to respond as quickly and accurately as possible. The arrow stayed on screen until the subject responded or until the end of the response window (800 ms), whichever came first. No feedback was presented. A new arrow was presented every 1.2 s with a total of six arrows between each trial. The experiment was programmed in Matlab (Mathworks, Natick, Mass), using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997).

fMRI data acquisition

Neuroimaging data were collected on a 3T Siemens Trio (Siemens Medical Solutions, Erlangen, Germany) scanner at Yale University, using a 12-channel head coil. After an initial anatomical localizer, a high-resolution, T1-weighted anatomical image was acquired. Next, 34 axial slices (4-mm thickness, no gap), parallel to the anterior commissure-posterior commissure line, covering the whole brain were defined, and a second T1-weighted anatomical image was acquired. Functional data were acquired with a T2*-weighted gradient-echo, echo-planer imaging sequence (TE = 25 ms; TR = 2000 ms; FA = 90^0; matrix = 64 x 64; FOV 224 mm), using the same 34-slice orientation. For all functional runs, 190 volumes were acquired. Images were projected onto a screen at the rear of the scanner, which the subject could view via a mirror on the head coil.

fMRI data analysis

All data analysis was done by BrainVoyager QX 1.10 (Brain Innovation, Maastricht, The Netherlands). The first two volumes of each functional run were discarded. The data were then corrected for head motion, specially smoothed (6-mm FWHM kernel), corrected for slice acquisition time, detrended, normalized into Talairach space (Talairach & Tournoux, 1988), and interpolated to 3-mm isotropic voxels. Subjects' functional data were coregistered with their T1-weighted anatomical volumes, which were also transformed into Talairach space.

All functional data were z-normalized, and the volume time courses were corrected for serial

correlations. Predictors for a general linear model (GLM) with six regressors (2 faces x 3 adjective conditions) were obtained by convolving a model hemodynamic response function with the time course of the stimuli. Separate regressors were created for each condition per subject (random effects analysis). A second GLM with two regressors (Black faces and White faces, collapsed across adjective type) was created in the same manner. For all analyses, a mask was applied, covering the whole brain (excluding the cerebellum), restricting the number of voxels to 62,056.

A whole-brain random-effects analysis ANCOVA was performed with the contrast "Black faces paired with neutral adjectives > White faces paired with neutral adjectives" with IAT scores as a covariate to show regions where the β weights of the GLM-contrast significantly correlated with subjects' IAT scores. The Brodmann's area (BA) that corresponded to the most significant voxel in each significant cluster was identified by Talairach Daemon (Research Imaging Centre, University of Texas Health Science Centre at San Antonio) (Lancaster et al., 2000). Using the cluster threshold size estimator plug-in BrainVoyager QX, Monte Carlo simulations (10,000 iterations) were performed—taking into account the number of significant voxels, the functional voxel size, and the smoothness of the map—to calculate the probability that 10 contiguous voxels would be significant by chance (Forman et al., 1995). This information was used to calculate the cluster-corrected false probability rate.

To show which regions predict award size, another whole-brain random-effects ANCOVA with the contrast "Black faces > White faces" (using the two-condition GLM) and the covariate "average verdict" was computed to show regions where the β weights of the GLM-contrast significantly correlated with subjects' survey results. Another whole-brain random-effects ANCOVA was performed with the contrast "Black faces paired with neutral adjectives > White faces paired with neutral adjectives" (using the six-condition GLM) and the covariate "average verdict." The same random-effects ANCOVA was performed two more times for positive and negative adjectives. A threshold of $p < .0025$ uncorrected, extent 10 voxels, was used to identify significant regions (based on Baird, Silver, & Veague, 2010, who performed a similar analysis).

RESULTS

Behavioral results

A large range of IAT scores was observed (range = −0.79 − 0.99, mean = 0.44, $SD = 0.44$). The mean IAT score was significantly greater than 0, $t(18) = 4.37$, $p < .001$, meaning that, overall, subjects showed a pro-White/anti-Black bias. However, no significant correlation was found between subjects' IAT scores and their average verdict, $r(17) = -.10$, *ns* (see Figure 1).[1]

fMRI results

Correlation between IAT scores and neural activity

As in previous studies, we performed a whole-brain, random-effects ANCOVA for the contrast "Black faces paired with neutral adjectives > White faces paired with neutral adjectives" with subjects' IAT scores as a covariate (threshold: uncorrected $p < .005$, extent 10, cluster-corrected $p < .001$). Neural activity in two regions was significantly negatively correlated with subjects' IAT scores: the right anterior cingulate continuing into the right medial frontal gyrus (BA 32/10; *x*, *y*, *z*: 15, 50, 7; $r = -.77$; $p < .001$) and the left superior frontal gyrus (BA 10; *x*, *y*, *z*: −30, 56, −2; $r = -.73$; $p < .001$) (see Figure 2).

Correlation between verdicts and neural activity

Results revealed that greater average verdicts were correlated with greater activity for Black faces paired with neutral adjectives than for White faces paired with neutral adjectives in the right inferior parietal lobule (BA 40; *x*, *y*, *z*: 45, −52, 46; $r = .78$; $p < .001$) and the right superior frontal gyrus (BA 9; *x*, *y*, *z*: 36, 47, 28; $r = .77$; $p < .001$) (see Table 1, Figure 3).

None of the other contrasts found regions that predicted verdict amounts at our strict threshold ($p < .0025$ uncorrected, extent 10 voxels). However, at a slightly looser threshold ($p < .005$ uncorrected, extent 10 voxels, $p < .001$ corrected), greater activity for Black faces than for White faces (collapsed across adjective types) predicted greater average verdicts (see Table 1, Figure 3) in the right brainstem near the subthalamic nucleus/substantia nigra (*x*, *y*, *z*: 12, −16, −5; $r = .74$; $p < .001$) and a comparable area in the left midbrain near the subthalamic nucleus/substantia nigra bordering the parahippocampal gyrus (BA 28; *x*, *y*, *z*: −18, −16, −11; $r = .74$; $p < .001$).

In addition, at a looser threshold ($p < .005$ uncorrected, extent 9 voxels, $p < .002$ corrected) activity

[1] These results are consistent with the results of a pilot study that used the same IAT and a very similar survey of employment discrimination vignettes. The pilot study used undergraduates of many races. The pilot study also found no significant correlation, $r(27) = .01$, *ns*. For just the White subjects in the pilot study, there was still no significant correlation, $r(14) = .21$, *ns*.

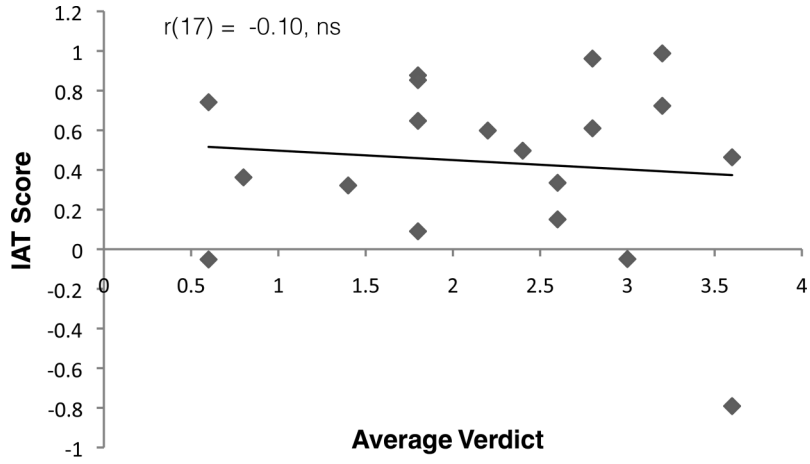

Figure 1. Relationship between subjects' IAT scores and the average amount they awarded. $r(17) = -.10$, *ns*.

for Black faces paired with negative adjectives minus activity for White faces paired with negative adjectives in right medial frontal gyrus was found to negatively correlate with verdict size (BA 9; *x*, *y*, *z*: 6, 53, 37, $r = -.73$; $p < .001$).

No regions in the final a priori contrasts (Black faces paired with positive adjectives minus White faces paired with positive adjectives) predicted verdict size even at the looser threshold.

Finally, motivated by Buckholtz et al.'s (2008) finding that total brain activity compared to baseline (although using a different a different task and a different type of stimuli) could predict punishment magnitude, we conducted a post-hoc, exploratory

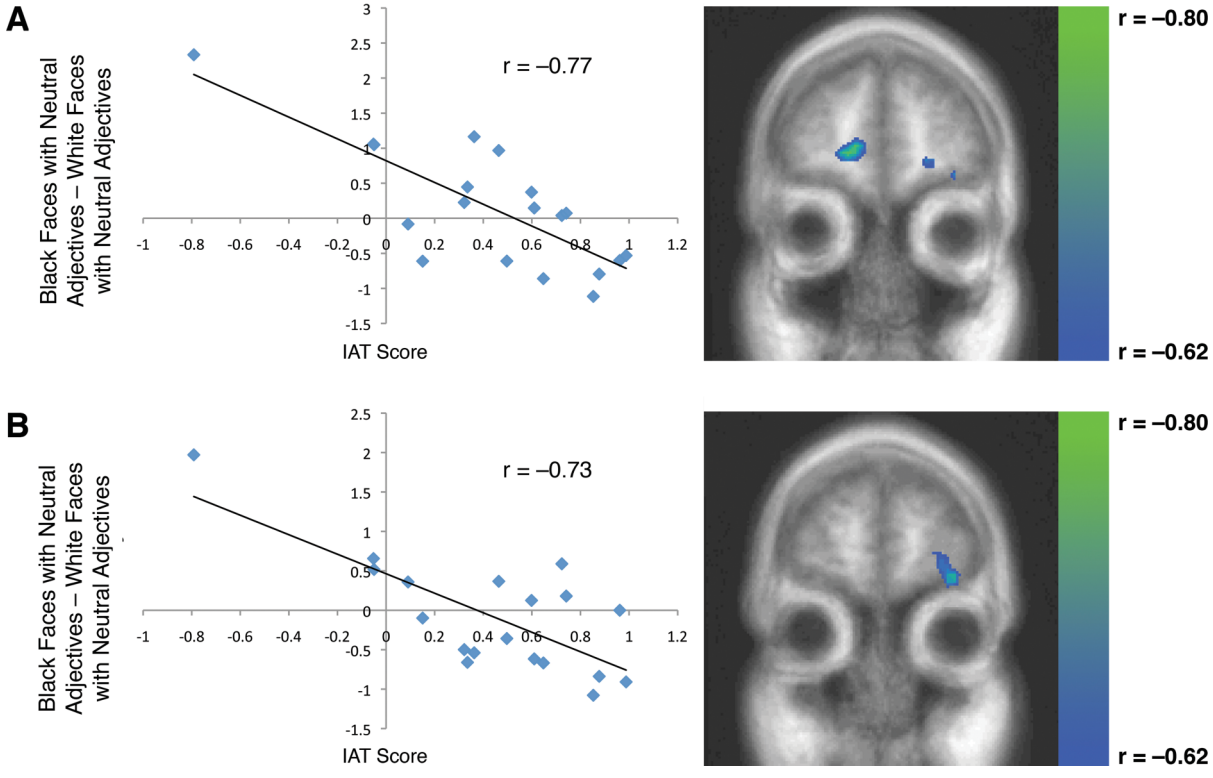

Figure 2. Scatterplots (left) and statistical activation maps (right) of significant correlations between average IAT scores (higher is more pro-White bias) and neural activity for Black faces paired with neutral adjectives minus neural activity for White faces paired with neutral adjectives. (A) right medial frontal gyrus (BA 32/10; 15, 50, 7); (B) left superior frontal gyrus (BA 10; −30, 56, −2). Scatterplots are for most significant voxel in the cluster.

TABLE 1
Regions showing significant correlation with average verdicts

Region	Brodmann area	Talairach coordinates			Region size (mm^3)	r value
		x	y	z		
A. Black (neutral adj.) – White (neutral adj)						
R. inferior parietal lobule	40	45	−52	46	355	.78*
R. superior/middle frontal gyrus	9/10	36	47	28	419	.77*
B. Black (all adj.) – White (all adj.)						
R. midbrain (subthalamic nucleus/ substantia nigra)		12	−16	−5	357	.74*
L. midbrain (subthalamic nucleus/ substantia nigra)[1]		−18	−16	−11	418	.74*
C. Black (negative adj.) – White (negative adj.)						
R. medial frontal gyrus	9	6	53	37	247	−.73*

Notes: For (A) statistical threshold: $p < .0025$, extent 10 voxels; for (B) statistical threshold: $p < .005$, extent 10 voxels (cluster-corrected $p < .001$); for (C) statistical threshold: $p < .005$, extent 9 voxels (cluster-corrected $p < .002$). *$p < .001$. Talairach coordinates and r values are of the most signficant voxel. [1]Peak voxel falls within the parahippocampal gyrus (BA 28) in a cluster that is contiguous with other significant voxels within the left midbrain region bilateral to the right midbrain region.

whole-brain ANCOVA with total activity collapsed across all six conditions compared to baseline and the covariate "average verdict." At a threshold of $p < .0025$ uncorrected, extent 10 voxels, four regions were found where total activity compared to baseline, collapsed across all conditions without a Black–White contrast, predicted verdict size: right anterior cingulate (BA 32; x, y, z: 3, 38, 10; $r = .85$; $p < .001$), right middle frontal gyrus (BA 8; x, y, z: 21, 23, 40; $r = .81$; $p < .001$), right inferior parietal lobule (BA 40; x, y, z: 42, −43, 43; $r = .77$; $p < .001$), and left middle frontal gyrus (BA 47; x, y, z: −42, 38, −5; $r = -.75$; $p < .001$).

DISCUSSION

In this study, we looked to see whether the amount of money people awarded victims of employment discrimination could be predicted by IAT scores or neural activity in response to Black and White faces paired with adjectives. We found the IAT scores did not predict verdict size. However, neural activity for Black faces paired with neutral adjectives minus neural activity for White faces paired with neutral adjectives in right inferior parietal lobule (BA 40) and in right superior/middle frontal gyrus of DLPFC (BA 9) did predict verdict size. This suggests that brain activity could be a more valid measure of racial bias than the IAT, at least when sample size is limited.

Correlation between brain activity and verdicts

We found that larger verdicts were associated with greater activity for Black faces paired with neutral adjectives than for White faces paired with neutral adjectives in right inferior parietal lobule (BA 40). The involvement of the inferior parietal lobule in social perception is consistent with prior work in the literature. One study found that the difference in bilateral activity in inferior parietal lobule between when subjects viewed pictures of John Kerry and when they viewed pictures of George W. Bush correlated with how much they liked Bush (Kaplan, Freedman, & Iacoboni, 2007). More generally, the inferior parietal lobule has been found to be more active when people think about people whom they find likable than when they think about people whom they find unlikable (Marsh et al., 2010). Therefore, as we found, the people who gave the largest verdicts (which likely means they have the least anti-Black or pro-White bias) should be expected to show more activity in the inferior parietal lobule when they view Black faces paired with neutral adjectives than when they view White faces paired with neutral adjectives. Likewise, as we also found, the people who gave the smallest verdicts should be expected to show more activity in the inferior parietal lobule when they view Whites faces paired with neutral adjectives than when they view Black faces paired with neutral adjectives.

Larger verdicts were also associated with greater activity for Black faces paired with neutral adjectives than for White faces paired with neutral adjectives in right DLPFC (BA 9/10). Prior studies have shown that this region is involved in storing stereotypic associative information (Milne & Grafman, 2001) and therefore is less active in incongruent blocks of the IAT than in congruent blocks, and the difference correlated with subjects' IAT scores (Knutson, Mah, Manly, & Grafman, 2007). For subjects with a large pro-White bias, pairing a Black face with a neutral word could be similar to the incongruent condition during an IAT.

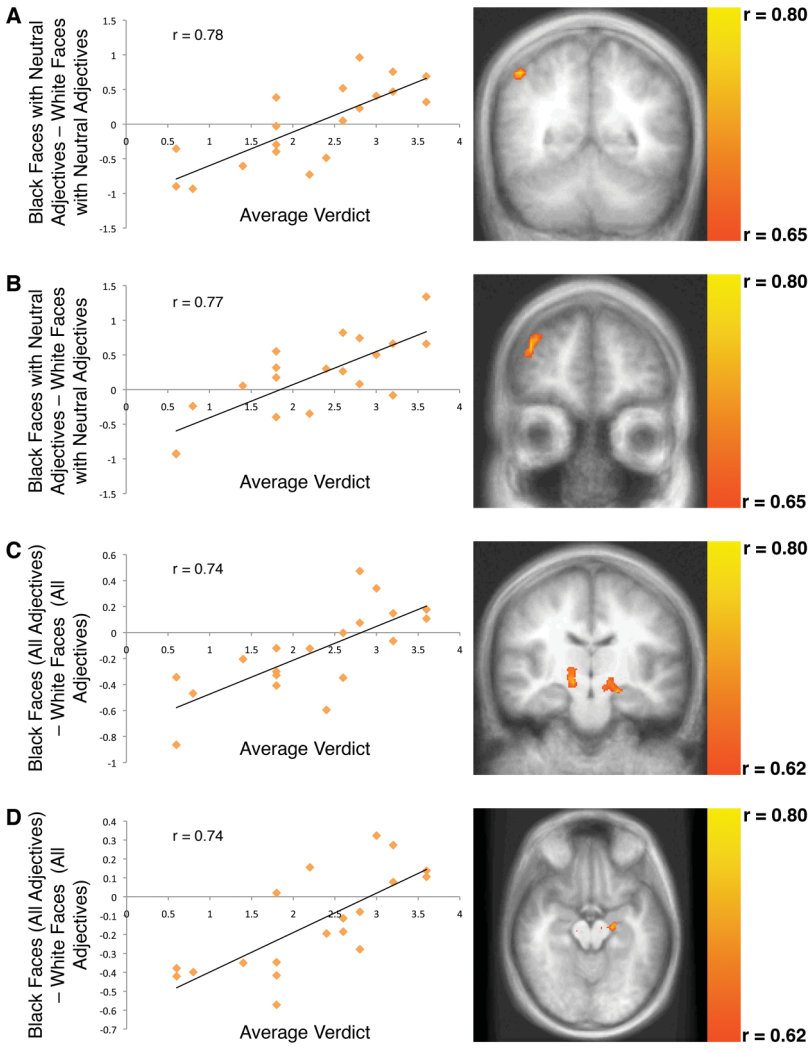

Figure 3. Scatterplots (left) and statistical activation maps (right) of significant correlations between average verdicts (higher is more money) and neural activity. (A) right inferior parietal lobule (BA 40; 45, −52, 46); (B) right superior/middle frontal gyrus of DLPFC (BA 9/10; 36, 47, 28); (C) right brainstem (near subthalamic nucleus/substantia nigra) (12, −16, −5); (D) left brainstem (near subthalamic nucleus/substantia nigra) bordering the parahippocampal gyrus (BA 28; −18, −16, −11). Scatterplots are for most significant voxel in the cluster.

Subjects might expect that Black faces would be paired with negative words, so pairing them with neutral or positive words would be incongruent. For these subjects, pairing White faces with neutral words would be congruent (pairing them with negative words would be incongruent). Therefore, for subjects who gave small average verdicts (possibly indicating large pro-White bias), neural activity for Black + neutral (incongruent) minus neural activity for White + neutral (congruent) should be expected to be negative, and this is consistent with what we found. For subjects with larger average verdicts (who presumably have less pro-White bias), Black + neutral and White + neutral should be neither congruent nor incongruent, so there should be no difference in neural activity for Black + neutral and neural activity for White + neutral, and this is also consistent with what we found. Finally, for subjects with the largest average verdicts (indicating possibly a pro-Black bias), the conditions might be reversed and Black + neutral would be congruent and White + neutral would be incongruent. In this case, neural activity for Black + neutral minus neural activity for White + neutral should be expected to positive, and this is what we found in subjects who gave the largest verdicts.

With a looser threshold, neural activity in the right medial frontal gyrus (BA 9) for Black faces paired with negative adjectives minus White faces paired with negative adjectives was negatively correlated with verdict size: The people who gave the smallest awards, and therefore probably had the largest

pro-White/anti-Black bias, had the most activity in response to Black faces. This corroborates Richeson et al. (2003), who found that activity in response to Black faces in right medial frontal gyrus was positively correlated with IAT scores (the people with the largest pro-White/anti-Black bias showed the most activity). Additionally, right medial frontal gyrus has also been found to be less active in incongruent blocks of the IAT than in congruent blocks (Knutson et al., 2007). Subjects who gave the smallest average awards (suggesting a large anti-black bias/pro-White bias) should find Black faces paired with negative adjectives the most congruent and white faces paired with negative adjectives the most incongruent, resulting in a large positive difference in activity for Black–White contrast just as we found. Subjects who gave larger average verdicts should show no difference in activity or even the reverse, resulting in the negative correlation between verdict size and activity that we found.

At the looser threshold, neural activity for Black faces minus White faces (collapsed across all adjectives) bilaterally in the brainstem near the subthalamic nucleus/substantia nigra was also found to predict award amounts. There is less literature to support an explanation of why this region would predict awards, but there is some evidence that the substantia nigra is involved in the processing of race: It has been found to be more active when people view Black faces then when they view White faces (Liberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005).

In a post-hoc analysis, total brain activity compared to baseline in response to all conditions correlated with verdicts in right anterior cingulate (BA 32), right middle frontal gyrus (BA 8), right inferior parietal lobule (BA 40), and left middle frontal gyrus (BA 47). The results are interesting with respect to a prior study that found that total brain activity compared to baseline could predict punishment magnitude (Buckholtz et al., 2008). A possible explanation is that subjects who give larger awards tend to process information in a general way that is especially attentive to our face-adjective stimuli (or in the case of BA 47 where the correlation is negative, in a way that is especially inattentive to it). However, these post-hoc correlations were not anticipated, and further investigation is needed, as our study design does not allow us to properly test any specific hypotheses. It is important to note that these results do not affect the Black–White results; the analyses are orthogonal.

It should also be noted that subjects always read the vignettes and made judgments prior to being scanned, so it is possible that the neural activity could reflect some sort of guilt pattern relating to their judgment rather than a prejudice pattern. However since there were 15–20 min between when subjects read the vignettes and when the functional scanning began—during which they competed the MRI safety checklist and completed the structural scans—we think this is less likely.

Correlations between neural activity and IAT

IAT scores correlated negatively with activity for Black faces paired with neutral adjectives minus activity for White faces paired with neutral adjectives in right medial frontal gyrus (BA 10) and left superior frontal gyrus (BA 10). Right medial frontal gyrus (BA 10) has previously been found to be less active in incongruent blocks of the IAT than in congruent blocks (Knutson et al., 2007). Therefore, the same logic that applies to why right BA 9 positively correlates with average award size (where higher numbers represent less racial bias) applies to why right BA 10 negatively correlates with IAT scores (where higher numbers represent more racial bias).

Richeson et al. (2003) also found that neural activity in response to Black faces in the right medial frontal gyrus (x, y, z: 6, 39, 33) predicted subjects' race-IAT scores (although the region is more superior/dorsal than the region we found). However, our correlation was negative (subjects with greater pro-White bias showed more activity in the prefrontal regions for White faces compared to Black faces), whereas the correlation found by Richeson et al. was positive. The correlation we performed was slightly different from the one done by Richeson et al., who correlated IAT scores with neural activity in response to Black faces (without subtracting neural activity in response to White faces). However, if we perform that same contrast in the same region, our results are still negative ($r = -.58$, $p < .01$), so that cannot explain the difference.

The key difference between the paradigm we used and the one Richeson et al. used was our use of adjectives, and this is the most likely reason our results differ. In our study, the adjectives—which were not used in Richeson et al.—caused the black + neutral adjective condition to be similar to an incongruent condition in an IAT, possibly causing a different neural response. Furthermore, in our study, in the scanner subjects were asked to remember the adjective-face pairing, whereas in Richeson et al.'s study subjects just observed the pictures and had to report whether they were on the right or left side of the screen. Tasks that force subjects to think of each face as an individual—as in our task where each photo was followed by an

adjective phrase stating "this person is"—have been shown to produce different responses in the amygdala when people view Black and White faces than when subjects view the faces as part of a social categorization or visual search task. In particular, when it is viewed as part of an individuation task, people show more activity in the amygdala when viewing White faces than when viewing Black faces, but when it is viewed as part of a categorization task, they show the opposite (Wheeler & Fiske, 2005). Since the prefrontal cortex (PFC) activity in response to Black and White faces is thought to be an attempt to regulate the amygdala response (Stanley, Phelps, & Banaji, 2008), the opposing responses in the amygdala between the two tasks might also be reflected in the PFC, explaining our results.

Additionally, in another study, medial PFC (mPFC)—near the region we found to significantly correlate with IAT scores—was found to be more active when subjects made judgments of people they implicitly felt more similar to (Mitchell, Macrae, & Banaji, 2006). White subjects who have high IAT scores likely feel they are much more similar to other Whites than Blacks (more so than people with lower IAT scores), so it is logical that subjects with higher IAT scores show more activity in mPFC for White faces than for Black faces.

Unlike Phelps et al. (2000) we found no correlation between amygdala activity, but this is to be expected, as in this study the faces were presented for a relatively long period of time, 500 ms, allowing for more controlled processing (Cunningham et al., 2004).

Although we did not find that behavioral IAT predicted verdicts, this could be because we had a small sample and correlations between behavioral measures and racial IATs are often weak (Greenwald et al., 2009). However, if IAT correlated as strongly with verdicts as fMRI did, we would expect to see some correlation between IAT and verdicts even with this sample size. Since our IAT/verdict correlation was far from significant, we can conclude that fMRI is a stronger predictor of verdicts than IAT. Further, prior literature suggests that even with a larger sample it is possible that IAT scores would not predict verdicts: A study of 133 judges found that judges' IAT scores did not predict the sentences they gave Black defendants when race was made explicit, as it was in our study (Rachlinski et al., 2009).

It is not a problem that IAT scores and average awards do not correlate with each other even though they each correlate with prefrontal cortex activity. There is no overlap between the specific brain regions with which IAT and verdicts each correlate (see Figures 2 and 3). The DLPFC region where activity in response to faces paired with neutral adjectives correlates with average verdict amounts is in right BA 9 while the region that correlates with IAT scores is more medial and inferior in right BA 10. To confirm that the regions do not overlap, we correlated IAT scores with the activity contrasts in each of the five brain regions that we found correlated with verdicts. For four of the five regions, the correlation between IAT scores and brain activity was less than .05 ($p > .8$). For one region (right superior frontal gyrus; BA 9) $r = -.35$, but that was still not significant ($p = .14$). It would be interesting to further explore the apparent functional specificity within prefrontal cortex for correlations with IAT and for correlations with award damages.

Conclusions

We found that activity in right DLFPC and right inferior parietal lobule in response to Black and White faces can predict how much subjects will give victims in hypothetical employment discrimination cases. Further studies should seek to corroborate these results by testing different employment discrimination vignettes and should seek to expand these results by testing whether these regions predict verdicts in other types of cases where race is salient. Additionally, our study used only White subjects. Further studies should look to see whether the same results can be reproduced in subjects of other races.

We are not suggesting that potential jurors be put in an MRI machine during jury selection for cases where race is salient. The cost of doing so would be prohibitive, many people might feel it is overly invasive, and— because of lack of data—courts have been hesitant to allow neuroimaging data to be used in trials, although the acceptance of neuroimaging data by the courts has steadily been increasing (Moreno, 2009). Further, it is unclear from our data which people are the least biased. The people whose neural activity in response to White faces subtracted from their neural activity in response to Black faces is greatest in the regions we identified—the people who would be predicted to award the victim the most money—are not necessarily the least racially biased. It is possible that these individuals have a pro-Black/anti-White bias. Still, this study is notable in that it shows that neuroimaging data can measure a racial bias that is reflected in juror decisions more effectively than a common behavioral measure—the IAT.

Original manuscript received 8 April 2011
Revised manuscript accepted 4 October 2011
First published online 7 November 2011

REFERENCES

Administrative Office of the United States Courts (2011). *2010 annual report of the director: Judicial business of the United States courts*. Washington, DC: US Government Printing Office.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*, 272–279.

Arterton, J. B. (2008). Unconscious bias and the impartial jury. *Connecticut Law Review*, *40*, 1023–1034.

Baird, A. A., Silver, S. H., & Veague, H. B. (2010). Cognitive control reduces sensitivity to relational aggression among adolescent girls. *Social Neuroscience*, *5*, 519–532.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron*, *60*, 930–940.

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. *Psychological Science*, *15*, 806–813.

Fazio, R., Jackson, J., Dunton, B., & Williams, C. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*, 636–647.

Green, A., Carney, D., Pallin, D., Ngo, L., Raymond, K., Iezzoni, L., et al. (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., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*, 945–967.

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.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test. III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. (2009). Implicit race attitudes predict vote in the 2008 presidential election. *Analyses of Social Issues and Public Policy*, *9*, 241–253.

He, Y., Johnson, M. K., Dovidio, J. F., & McCarthy, G. (2009). The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. *Social Neuroscience*, *4*, 426–442.

Kaplan, J. T., Freedman, J., & Iacoboni, M. (2007). Us versus them: Political attitudes and party affiliation influence neural response to faces of presidential candidates. *Neuropsychologia*, *45*, 55–64.

Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping*, *28*, 915–930.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, *10*, 120–131.

Levinson, J. D., Cai, H., & Young, D. (2010). Guilty by implicit racial bias: The Guilty/Not Guilty Implicit Association Test. *Ohio State Journal of Criminal Law*, *8*, 187–208.

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. *Nature Neuroscience*, *8*, 720–722.

Marsh, A. A., Kozak, M. N., Wegner, D. M., Reid, M. E., Yu, H. H., & Blair, R. J. R. (2010). The neural substrates of action identification. *Social Cognitive and Affective Neuroscience*, *5*(4), 392–403.

Milne, E., & Grafman, J. (2001). Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. *Journal of Neuroscience*, *21*, 1–6.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*, 655–663.

Moreno, J. (2009). The future of neuroimaged lie detection and the law. *Akron Law Review*, *42*, 717–734.

Nosek, B. A., & Banaji, M. R. (2002). (At least) two factors moderate the relationship between implicit and explicit attitudes. In R. K. Ohme, & M. Jarymowicz (Eds.), *Natura Automatyzmow* (pp. 49–56). Warszawa,Poland: WIP PAN and SWPS.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88.

O'Connor, S. D. (1992). *Georgia* v. *McCollum* Dissenting Opinion. Supreme Court of the United States.

Pelli, D. G. (1997), The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 729–738.

Phillips, P., Moon, H., Rizvi, S., & Rauss, P. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *22*, 1090–1104.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, *16*, 295–306.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review*, *84*, 1195–1246.

Rector, N. A., & Bagby, R. M. (1995). Criminal sentence recommendations in a simulated rape trial: Examining juror prejudice in Canada. *Behavioral Sciences & the Law*, *13*, 113–121.

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

Rooth, D. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*, 523–534.

Schleim, S., Spranger, T. M., Erk, S., & Walter, H. (2010). From moral to legal judgment: The influence of normative context in lawyers and other academics. *Social Cognitive and Affective Neuroscience*, *6*, 48–57.

Sniderman, P. M., & Piazza, T. (2002). *Black pride and Black prejudice*. Princeton, NJ: Princeton University Press.

Stanley, D., Phelps, E., & Banaji, M. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science*, *17*, 164–170.

Talairach, J., & Tournoux, P. (1988). *Co-planar Stereotaxic atlas of the human brain: 3-Dimensional proportional system: An approach to cerebral imaging*. Stuttgart, Germany: Thieme Medical Publishers.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*, 56–63.