OXFORD

Social Cognitive and Affective Neuroscience, 2017, 382–390

doi: 10.1093/scan/nsw135
Advance Access Publication Date: 20 September 2016
Original article

Neural predictors of evaluative attitudes toward celebrities

Keise Izuma,[1,2,3] Kazuhisa Shibata,[4,5] Kenji Matsumoto,[3] and Ralph Adolphs[2]

[1]Department of Psychology, University of York, Heslington, York YO10 5DD, UK, [2]Division of Humanities and Social Sciences, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, USA, [3]Brain Science Institute, Tamagawa University, 6-1-1, Tamagawa-gakuen, Machida, Tokyo 194-8610, Japan, [4]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer Street, Providence, RI 02912, USA and [5]Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Correspondence should be addressed to Keise Izuma, Department of Psychology, University of York, Heslington, York YO10 5DD, UK.
E-mail: keise.izuma@york.ac.uk.

Abstract

Our attitudes toward others influence a wide range of everyday behaviors and have been the most extensively studied concept in the history of social psychology. Yet they remain difficult to measure reliably and objectively, since both explicit and implicit measures are typically confounded by other psychological processes. We here address the feasibility of decoding incidental attitudes based on brain activations. Participants were presented with pictures of members of a Japanese idol group inside an functional magnetic resonance imaging scanner while performing an unrelated detection task, and subsequently (outside the scanner) performed an incentive-compatible choice task that revealed their attitude toward each celebrity. We used a real-world election scheme that exists for this idol group, which confirmed both strongly negative and strongly positive attitudes toward specific individuals. Whole-brain multivariate analyses (searchlight-based support vector regression) showed that activation patterns in the anterior striatum predicted each participant's revealed attitudes (choice behavior) using leave-one-out (as well as 4-fold) cross-validation across participants. In contrast, attitude extremity (unsigned magnitude) could be decoded from a distinct region in the posterior striatum. The findings demonstrate dissociable striatal representations of valenced attitude and attitude extremity and constitute a first step toward an objective and process-pure neural measure of attitudes.

Key words: attitude; attitude extremity; preference; fMRI; MVPA; striatum

Introduction

Attitudes comprise our evaluations of an object, a place, an idea, another person or oneself: which are good or bad, which do we want to approach or avoid? Attitudes influence a wide range of human behaviors ranging from everyday social interactions with other people, health behavior and political behavior to international relations. Although attitude is one of the most extensively studied concepts in the entire history of social psychology (Petty *et al.*, 2009), it has remained a challenge how best to objectively and accurately measure people's attitudes. Problems with self-report measures have been well-documented in the past (in particular, social desirability bias) (DeMaio, 1984; Podsakoff *et al.*, 2003). Although implicit measures of attitudes (Wittenbrink and Schwarz, 2007), such as the implicit association test (IAT; Greenwald *et al.*, 1998), have offered partial solutions to this problem, their validity is also debated (e.g. Bosson *et al.*, 2000; Karpinski and Hilton, 2001; Fazio and Olson, 2003; Arkes and Tetlock, 2004; Conrey *et al.*, 2005; Blanton *et al.*, 2006; Fiedler *et al.*, 2006; Sherman, 2009). For example, scores on the IAT are influenced by multiple cognitive

Received: 30 March 2016; **Revised:** 6 July 2016; **Accepted:** 9 September 2016

processes, not only implicit attitude (e.g. Conrey *et al.*, 2005; Sherman, 2009).

In this study, we aimed to test the feasibility of measuring people's attitudes toward other familiar people based on the power of their brain activations to predict choice behavior, and without requiring any explicit or implicit task in the scanner. A neural measure of incidental attitudes would have significant potential to provide a process-pure metric, and avoid the contamination with many other processes that limits currently available explicit as well as implicit measures of attitude (Conrey *et al.*, 2005; Sherman, 2009). Our approach used multi-voxel pattern analysis (MVPA) together with support vector regression (SVR) on functional magnetic resonance imaging (fMRI) signals. Although conventional univariate fMRI analysis compares the strength of activations in each single voxel independently, MVPA classifies the distributed patterns of activations across multiple voxels in a high-dimensional space and can be more sensitive for detecting and distinguishing different psychological states (e.g. Vickery *et al.*, 2011; Jimura and Poldrack, 2012).

In this study, we particularly focus on people's attitudes toward social objects, familiar people. Although many past social psychological studies as well as social neuroscience studies on attitudes have focused on racial attitudes (e.g. Phelps *et al.*, 2000; Stanley *et al.*, 2011; for reviews, Ito and Bartholow, 2009; Kubota *et al.*, 2012; Amodio, 2014), studies of familiar famous people offer some advantages: unlike racial attitude, the relationship between self-reported attitude and behavior can be more straightforward (e.g. less susceptible to social desirability bias) and thus the relationship among neurally measured attitude, self-report attitude and behavior (revealed attitude) is more easily interpretable.

Furthermore, in addition to testing the feasibility of inferring people's attitudes based on brain activations (a question of psychological interest regardless of the neuroanatomical details), this study also investigates the regional neural representations of attitudes and attitude extremity (i.e. how extreme attitude is regardless of its valence). Although neuroeconomics studies have extensively investigated neural representations of attitude toward non-social objects (e.g. foods, DVDs, etc.) (for review, Levy and Glimcher, 2012; Clithero and Rangel, 2014), only few studies (e.g. Cunningham *et al.*, 2003, 2008; Knutson *et al.*, 2006; Tusche *et al.*, 2013) investigated the neural representation of attitude toward other familiar people, despite the high relevance of this topic to our everyday social behaviors. Furthermore, attitude extremity is known to be one of the important attitude properties (Petty and Krosnick, 1995). For example, attitude extremity affects an individual's information processing (e.g. Powell and Fazio, 1984; Van Boven *et al.*, 2012) and modulates the relationship between explicit and implicit attitudes (Karpinski *et al.*, 2005). However, its neural mechanisms also remain largely unexplored (see Cunningham *et al.*, 2008; Luttrell *et al.*, 2016, for a notable exception).

In the fMRI scanner, participants were presented with 10 members of a Japanese female idol group (the Japanese music performance group 'AKB48'). After the scanning session, they performed an incentive-compatible choice task, which behaviorally quantified each participant's attitude toward each of the 10 members (with a real-world outcome). MVPA was applied to these data to identify those brain regions that could predict a participant's choice behavior toward each member. Since different idol group members were in fact associated with idiosyncratic preferences amongst our participants, the neural data should uniquely encode the attitude for a member that predict the preference choice, unconfounded by the perceptual appearance of the member. We expected to find neural signatures of social attitude within those brain regions previously associated with value representations in general, including striatum, ventromedial prefrontal cortex (vmPFC), insula, amygdala and anterior cingulate cortex (e.g. Buchel *et al.*, 1998; LaBar *et al.*, 1998; Delgado *et al.*, 2000; Knutson *et al.*, 2000; O'Doherty *et al.*, 2001; Knight *et al.*, 2005; Hare *et al.*, 2008; Izuma *et al.*, 2008; for meta-analyses, Etkin *et al.*, 2011; Bartra *et al.*, 2013; Sescousse *et al.*, 2013). Nonetheless, to obtain a data-driven set of results, our primary analysis used a whole-brain approach.

Materials and methods

Participants

A total of 23 college students participated in the study. One participant was excluded from the analysis due to excessive head motion, and the remaining 22 participants were included in the analyses (11 female; mean age = 19.7, s.d. = 1.35). All participants were prescreened so that all of them knew at least 20 members of the idol group and had at least one highly liked member and one highly disliked member. The participants were all right-handed with no history of neurological or psychiatric illness. All participants gave written informed consent for participation, and the study was approved by the Institutional Review Board of Tamagawa University.

Stimuli

Photographs of the faces of individual members of the Japanese female idol group 'AKB48' (https://en.wikipedia.org/wiki/AKB48) were used in this study. All stimuli were obtained from the Internet. AKB48 consists of more than 100 members and is popular especially among young Japanese people. We decided to use this idol group for our source of experimental stimuli for two reasons: first, and most importantly, their unique annual election contest allows us to have an incentive-compatible choice task with real-world validity (see later for more detail on the choice task). Every year, the idol group has a unique annual election contest as a marketing strategy. Before the election contest, their new album is released, which includes a voting code. Each fan can vote for his/her favorite member using the code, and this election contest has a significant influence on each member's media exposure. The more votes a member receives, the more heavily she will be promoted. This election system allows each fan to vote more than once by purchasing many albums. Second, related to the first point, because of the competitive relationship among members of the group created by the election contest, each fan (participant) typically has a positive attitude toward some members and a negative attitude toward other members. This large variance in attitudes provides an ideal real-world background for the purpose of this study.

During the fMRI scanning, each participant was presented with pictures of 10 different members. Since we were interested in the neural representation of already established social attitudes (as opposed to first impressions toward people never seen before), we ensured that each participant was familiar with 10 members presented during the fMRI scanning with the following procedure. Before an fMRI experiment, each participant was asked to provide the experimenter with names of his/her most and least favorite members (at least one for each) by email. Based on the names each participant gave, the experimenter picked 10 members for the participant including his/her favorite and least favorite members. If participants named fewer than 10 members, the experimenter selected other members based on general popularity. On the day of the fMRI experiment, before they entered an fMRI scanner, each participant was shown pictures of the 10 members and asked if they could identify all of them. If there were any members that could not be identified, participants were asked to pick other members they knew from a list of all members' names and pictures. Accordingly, picture stimuli used in the fMRI experiment were different for each participant, although there was overlap. More specifically, a total of 43 different members of the idol group were used in the experiment, and each of them was presented to at least one participant. Fifteen out of the 43 members (34.9%) were presented to only one participant. The most consistently used member was presented to 15 different participants. However, these 15 participants had idiosyncratic attitudes toward the member ranging from revealed attitude scores of 3–17 (possible range = 0–18; see later for more information on the revealed attitude score). Thus, the results of the across-participant MVPA reported later are highly unlikely to be explained by decoding of person identity or face features (e.g. hair length).

Experimental tasks

In each trial, a picture of a member of the idol group was presented for 4 s, and participants were asked to perform a simple button press task during the fMRI session. At a random point between 1.5 and 3 s after the onset of the picture presentation, the picture became darker for 0.2 s (Figure 1a). Participants were asked to press the button as soon as they detected the luminance change. The intertrial interval was set to 4, 6 or 8 s (pseudorandomly determined). In each fMRI run, each of 10 members was presented three times (30 trials). Each run lasted 5 min, and there were eight fMRI runs in total. Importantly, before the scanning, participants were not told that the experiment was about attitudes, and they were not explicitly asked to think about their attitude toward each member during the scanning.

After the fMRI session, participants were asked to perform a choice task outside of the scanner (they were not told before the scanning that there would be the choice task). In each trial, two members from the 10 members used in the fMRI task were presented on the computer screen (Figure 1b), and participants were asked to select the one they want to vote for at the next election event by pressing one of two keys on the keyboard. There were 45 unique choice pairs, and each pair was presented twice (with switched positions) so that participants made 90 binary choices in total. During the choice task, each of the 10 members was presented 18 times. Therefore, each member could be selected a minimum of 0 times and maximum of 18 times. Since any attitude measure is considered to be good so long as it can predict relevant behaviors, this 'revealed attitude score' for each member was used as labels in the subsequent MVPA analysis. Thus, in this project, we aim to predict individual's choice behavior (i.e. revealed attitude score) based on brain activations that would encode attitude but without relying on any psychological measures of attitude (e.g. self-report).

Importantly, before they started the choice task, all participants were instructed that after the choice task, one trial would be selected randomly and we would actually vote for the member the participant selected in that trial. All participants were shown a new music album including a voting code, and using the code, the experimenter actually voted for a member selected by the participant in a randomly selected trial.

Finally, participants were asked to rate each of 10 members on attitude (how much do you like this member?) and attractiveness (how attractive do you think this member is) using a 9-point scale. At the end of the experiment, all participants received the album in addition to the monetary compensation (8000 Japanese yen) for their participation.

fMRI data acquisition

All fMRI data were acquired using a Siemens 3.0 Tesla Trio 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 34 contiguous 3.5-mm-thick transaxial slices covering nearly the entire

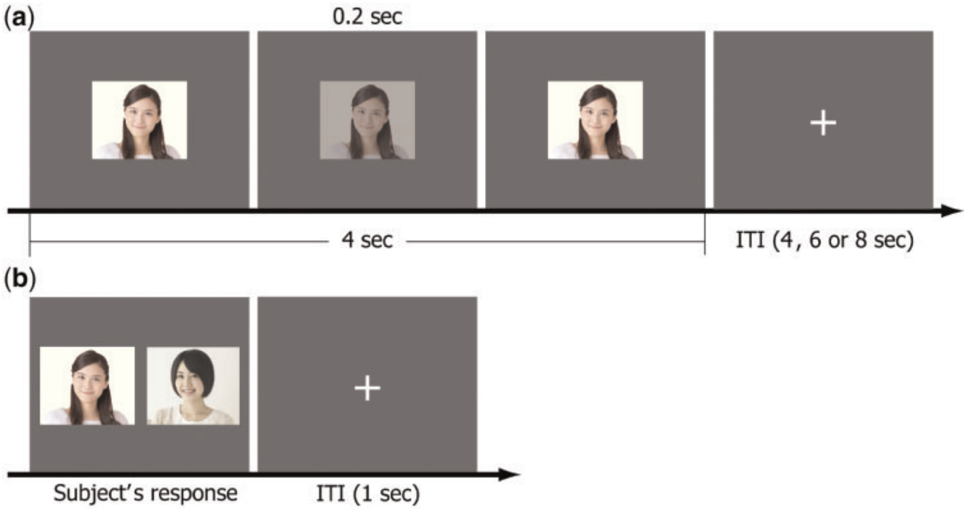

Fig. 1. Experimental tasks. (a) A single trial during the fMRI scanning. Each participant was asked to press a key when the luminance of a picture changed. (b) A single trial of the choice task. After the fMRI session, each participant performed the choice task. In each trial, two members of the idol group were presented on the screen, and participants were asked to choose the member they want to vote for at the next election event. Note that due to copyright restrictions, the two individuals depicted in the pictures in this figure are not actual members of the idol group.

cerebrum (repetition time = 2000 ms; echo time = 25 ms; flip angle = 90°; field of view = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.5 mm). A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

fMRI data preprocessing

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 image volume, head motion was corrected. Following motion correction, 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 (resliced to a voxel size of 3.0 × 3.0 × 3.5 mm). These normalized data were used for the MVPA data analyses. For the univariate analysis, the normalized fMRI data were spatially smoothed with an isotropic Gaussian kernel of 4 mm (full-width at half-maximum).

fMRI data analysis: searchlight MVPA

In the across-participant MVPA analysis, we attempt to predict attitudes toward each of 10 idol group members based on the data obtained from all other participants. We first ran a conventional general linear model (GLM) analysis. In the GLM, each of 10 members was separately modeled (duration = 4 s). Button presses (duration = 0 s) and head motions were also included in the model as nuisance regressors. Contrast images for each member were created by using the data from all of the eight fMRI runs. These contrast images were used as input, and revealed attitude scores (how many times each member was chosen during the choice task; 0–18) were used as labels in the MVPA analysis.

MVPA was performed by using custom-made MATLAB in combination with LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). To predict the parametric variable of revealed attitude scores, we employed SVR (Drucker *et al.*, 1997), as implemented in LIBSVM, with a linear kernel and a cost parameter of $c = 0.01$. This cost parameter was selected a priori following a previous study investigating value and salience signals in the brain (Kahnt *et al.*, 2014). The regression MVPA was performed using a whole-brain searchlight procedure (Kriegeskorte *et al.*, 2006) with a radius of three voxels (maximum of 123 voxels, and less at the boundaries of the brain). In each searchlight, accuracy at predicting revealed attitude scores for the 10 members was computed using leave-one-participant-out cross validation. In each cross-validation, one participant was left out, and the SVR was performed using the data from all other participants and then tested on the left-out participant. This procedure was repeated for each participant (a total of 22 times; to test the robustness of our findings, we also ran a 4-fold cross-validation which replicated the main findings [Supplementary Results, Table S2]). In each searchlight analysis, Spearman's rank-order correlations were computed between a participant's revealed attitude scores and their predicted attitudes; this correlation value was then assigned to the center voxel of the searchlight, resulting in an anatomical correlation map for each participant. The correlation values were Fisher z-transformed, spatially smoothed with an isotropic Gaussian kernel of 4 mm (full-width at half-maximum) and then submitted to the second level analysis (i.e. one sample *t*-tests across all participants).

To identify the neural representations of attitude extremity, we also ran the same across-participant MVPA analysis using the attitude extremity score, which is computed by calculating the absolute value of the difference between a revealed attitude score for each member and the midpoint of the revealed attitude score (i.e. 9). We further ran three control MVPA analyses using (i) self-report attractiveness ratings, (ii) self-report attitude ratings and (iii) reaction times (RTs) (see Supplementary Information).

fMRI data analysis: univariate analysis

We also ran a standard univariate fMRI analysis to see whether univariate activations might be correlated with revealed attitude score. The GLM included three main regressors; (i) all idol group member presentations (duration = 4 s), (ii) member presentations parametrically modulated by participant's revealed attitude score and (iii) member presentations parametrically modulated by participant's attractiveness rating. Button presses (duration = 0 s) and head motions were also included in the model as nuisance regressors.

For both the MVPA and univariate analyses, the statistical threshold was set at $P < 0.001$ voxelwise (uncorrected) and cluster $P < 0.05$ (FWE corrected for multiple comparisons).

Results

Behavioral results

Not surprisingly, revealed attitude scores (choice behavior) were highly correlated with self-reported attitudes (average $r = 0.91$, $t[21] = 18.11$, $P < 0.001$). They were also correlated with attractiveness ratings (average $r = 0.76$, $t[21] = 10.80$, $P < 0.001$). The revealed attitude scores were more strongly correlated with the attitude ratings than the attractiveness ratings ($t[21] = 3.33$, $P = 0.003$). Attitude ratings and attractiveness ratings were also correlated with each other (average $r = 0.76$, $t[21] = 10.69$, $P < 0.001$).

Inside the fMRI scanner, participants were asked to press the button as soon as a picture gets dimmed, and their performance for this simple button press task was nearly perfect (98.6%) and average RT across participants was 297 ms (s.d. = 52), indicating that participants paid attention to each picture stimulus. The analyses also revealed that RTs were significantly negatively correlated with participants' revealed attitude scores (average $r = -0.20$, $t[21] = -3.14$, $P = 0.005$) and attitude ratings (average $r = -0.16$, $t[21] = 2.27$, $P = 0.034$). Thus, the more favorable their attitudes were toward members, the faster (smaller) the RT. This result may suggest that pictures of their favorite members captured attention relative to less favorite members, which in turn enhanced their RTs. RTs were not correlated with the attractiveness ratings (average $r = -0.08$, $t[21] = 1.24$, $P = 0.23$, n.s.) or the attitude extremity scores (average $r = -0.12$, $t[21] = 1.30$, $P = 0.21$, n.s.).

After the scanning, participants were asked to make binary choices between two members of the idol group for whom they wanted to vote in the next election event. They were presented with the same choice pair twice, and their choices were largely consistent across two presentations of the same pair (choice consistency = 86.7%). Inconsistent choices were more likely to happen when the difference in attitude ratings between two members was small (e.g. when two persons were similarly liked). When the within-pair ratings difference was equal to or less than 2, choice consistency was 79.3%, while it was 94.0% when the difference was more than 2. This difference between the two choice consistency values was significant (*t*[21] = 5.32, $P < 0.001$). Not surprisingly, participants' choices were highly accurately predicted by their self-report attitudes (mean choice prediction accuracy = 88.5%), and the accuracy was significantly higher than chance (50%) (*t*[21] = 20.1, $P < 0.001$; note that choice accuracy was computed after excluding inconsistent choice pairs).

fMRI results: searchlight MVPA

The searchlight MVPA analysis revealed that spatial activation patterns in the anterior part of the right striatum significantly predicted participants' revealed attitude scores (Figure 2a). The average Spearman's correlation coefficient in the peak of this anterior striatum cluster was 0.26 (*t*[21] = 7.24, $P < 0.001$). Activation patterns in right inferior frontal gyrus (IFG) also significantly predicted participants' attitudes (Spearman's Rho = 0.23, *t*[21] = 6.49, $P < 0.001$; Figure 2a, Table 1). No other region significantly predicted participants' revealed attitude scores. The average correlation of 0.26 found in the anterior striatum is equivalent to 59.6% in terms of accuracy for predicting participants' binary choices (see Supplementary Results for more details about the choice prediction accuracy results and the direct comparison of the choice prediction accuracy between neural and self-report measures).

The MVPA analysis with the attitude extremity score revealed that activation patterns in a more posterior part of the right striatum significantly predicted attitude extremity (Figure 2a). Activation patterns in other brain regions including inferior orbitofrontal cortex (OFC), supplementary motor area (SMA), posterior insula, precentral gyrus, inferior parietal lobule and occipital pole also significantly predicted attitude extremity (Table 1).

To further quantify the functional dissociation within the striatum, we extracted each participant's correlation coefficients for both the revealed attitude score and the attitude extremity score from the voxels within a 4 mm sphere surrounding the peaks of these two striatum subregions (note that the following statistical analysis is of course not statistically independent of our earlier discovery; it is intended merely to provide further detail for the dissociation). After Fisher's *z* transforming these correlation coefficients, we performed a 2 (anterior *vs* posterior striatum) × 2 (revealed attitude score *vs* attitude extremity score) repeated-measure analysis-of-variance (ANOVA). It revealed a significant interaction (*F*[1, 21] = 26.4, $P < 0.001$) (Figure 2b). We further performed a paired *t*-test within each of the anterior and posterior striatum regions. In the anterior striatum, prediction

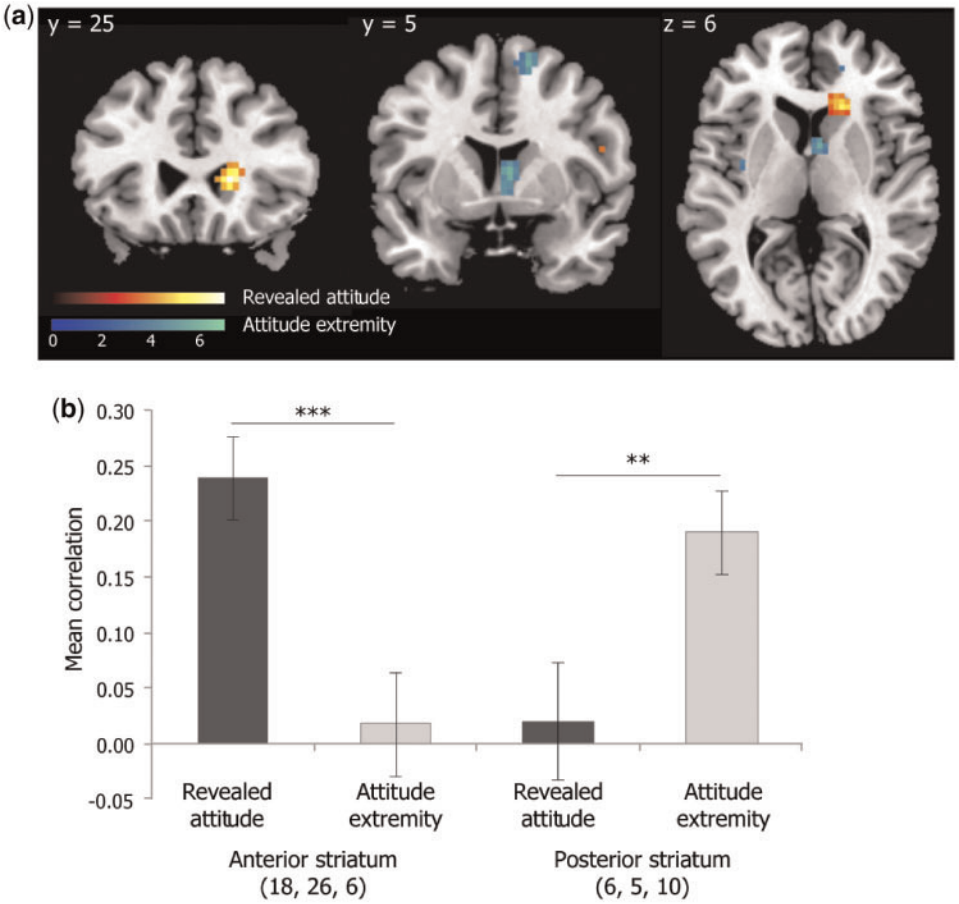

Fig. 2. fMRI results. (a) Activation patterns in the anterior striatum significantly predicted participants' revealed attitude scores, whereas activation patterns in the posterior striatum (blue) predicted attitude extremity. (b) Mean within-participant Spearman's rank correlation between predicted values and actual values (revealed attitude or attitude extremity) from the anterior and posterior striatum. **$P < 0.01$, ***$P < 0.001$. Error bars represent SEM.

Table 1. Regions where activity patterns significantly predicted revealed attitude score and attitude extremity score

		MNI coordinates				Number of Voxel
Location	Side	x	y	z	z value	
Revealed attitude score						
Anterior striatum	R	18	26	6	5.07	37
IFG	R	54	14	20	4.76	56
Attitude extremity score						
Posterior striatum	R	6	5	10	4.56	31
Inferior OFC	R	30	38	−4	3.90	25
SMA	R	15	2	66	4.82	26
Posterior insula	L	−48	−7	2	4.00	46
Precentral gyrus	L	−33	−19	62	5.07	63
Inferior parietal lobule	L	−51	−31	44	4.38	48
Occipital pole	R	30	−95	−15	4.48	45

Note that the size of a voxel is 3 × 3 × 3.5 mm.

performance (mean correlation) was significantly higher for revealed attitude compared with attitude extremity ($t[21] = 4.24$, $P < 0.001$). In contrast, in the posterior striatum, prediction performance was significantly higher for attitude extremity than for revealed attitude ($t[21] = 3.48$, $P = 0.002$) (Figure 2b). Furthermore, correlations were not significantly different from zero for attitude extremity in the anterior striatum and for revealed attitude in the posterior striatum. These results indicate a clear functional dissociation within the striatum such that the anterior striatum represents an individual's evaluation of each celebrity, whereas the posterior striatum represents how extreme his/her attitude is regardless of valence. Main effects of region and attitude property were both not significant ($P > 0.50$). Finally, a mixed ANOVA with gender as an additional between-subject factor revealed no main or interaction effect involving gender (all $P > 0.35$).

Our control analyses further confirmed that the MVPA results reported earlier cannot be explained by perceived attractiveness of faces or RTs (i.e. attention) (see Supplementary Results, Figure S1, Table S1). Furthermore, additional searchlight MVPA analysis restricted to an anatomical mask of the striatum confirmed that activation patterns within the striatum are responsible for the findings reported earlier and could not be ascribed to partial volume effects from nearby regions (Supplemental Results).

fMRI results: univariate analysis

The results of the univariate fMRI data analysis revealed that the activity only in the left posterior fusiform gyrus ($x = -27$, $y = -91$, $z = -15$; 60 voxels) was significantly positively correlated with the revealed attitude scores. Only when the threshold was lowered to $P < 0.005$, did we find activations in the vmPFC ($x = -6$, $y = 53$, $z = -12$), one of the areas commonly associated with preference and valuation (Bartra *et al.*, 2013; Clithero and Rangel, 2014). We did not find any activation in the striatum even with this lowered threshold. In addition, no area was significantly negatively correlated with the revealed attitude scores. Furthermore, no area was significantly correlated (either positively or negatively) with the attitude extremity scores.

Discussion

This study investigated the possibility of measuring people's social attitude toward familiar others based on multivariate neural activations from fMRI. The across-participant MVPA revealed that activation patterns in the anterior striatum can significantly predict the choices made based on one's attitude toward members of an idol group. This result indicates that spatial patterns of activations in the anterior striatum contain reliable information about an individual's attitudes, and these neural representations of attitudes in the anterior striatum are sufficiently similar across different individuals so that it is possible to infer attitudes of an individual based on the association between revealed attitudes and brain activation patterns found in other individuals. In contrast, our univariate analysis failed to find any significant activations related to participants' attitudes (and attitude extremity) in reward-related areas, indicating that average amplitude of activity is an insufficiently sensitive measure to represent attitudes in this study. Although previous studies have demonstrated that univariate activations in reward-related brain regions such as vmPFC and striatum are correlated with people's attitudes or preference for various items, especially when individuals were asked to report their preference for each stimulus inside the scanner (Lebreton *et al.*, 2009; Izuma *et al.*, 2010; for a meta-analysis, see Bartra *et al.*, 2013), such univariate (mean) activation does not appear to robustly track preference for stimuli during passive viewing (but see also Levy *et al.*, 2011; Tusche *et al.*, 2013). In general, our results add to a growing body of evidence showing the higher sensitivity of MVPA compared with univariate analyses (e.g. Jimura and Poldrack, 2012; Kohler *et al.*, 2013).

Although we found that activation patterns in the anterior striatum and right IFG can predict participants' attitudes revealed in the choice task (i.e. behavior) significantly better than would be expected by chance, prediction accuracy was considerably lower than their self-reported attitude ratings. Thus, at this point, self-report measure outperforms neural measures of attitudes. This result is not surprising because self-report attitude toward celebrities should be much less susceptible to social desirability bias compared with attitudes toward socially sensitive issues (e.g. racial prejudice), and there is no apparent reason for participants to regulate their answers during attitude ratings of the celebrities. Nonetheless, our present study provides an important reference point to which future social neuroscience studies can be compared. For example, it might well be the case that neurally measured attitude can outperform self-report or implicit measures of attitudes in predicting racially discriminatory behavior for which prediction accuracy of behavioral measures (both implicit and explicit measures) are known to be limited (Greenwald *et al.*, 2009; Oswald *et al.*, 2013), an important direction for future studies.

The greatest potential of a neural measure of attitude comes from the fact that we could successfully predict participants' choice behavior (i.e. revealed attitude score) from their brain activations alone without asking them to engage in any attitude-related task during scanning, and without incorporating information about attitude judgments into our analysis of the fMRI data. Although implicit attitude measures circumvent some of the problems with self-reports, they cannot be 'process-pure', so that a score on an implicit attitude measure generally reflects some factors other than implicit attitude toward a stimulus (Conrey *et al.*, 2005; Sherman, 2009). In contrast, in this study, participants were never asked to report their attitudes (i.e. self-report) or make any response based on attributes of attitude objects (i.e. implicit measure) during the scanning—yet the activation patterns could robustly predict later choices based on attitudes. This suggests that it may be possible to infer people's spontaneous attitudes using incidental brain imaging methods. As neurally measured incidental attitudes are unlikely to be influenced by any automatic or controlled processes, there is no translational gap between the construct (i.e. attitude) and the way it is measured (Sherman, 2009). Our findings suggest that it may be possible to measure people's attitudes even when they are unable and/or unwilling to report them truthfully (although ethical concerns for such an approach need to be carefully considered). Thus, a neural measure of incidental attitudes has the potential to be highly, if not completely, process-pure as well as relatively effort-free, offering some distinct advantages over any implicit (and explicit) behavioral measures.

It should be noted that while the neural measure was outperformed by self-report in predicting individual's choice behaviors, the choice prediction accuracy found in this study (59.6%) is comparable to the three past neuroeconomics studies (Tusche *et al.*, 2010; Levy *et al.*, 2011; Smith *et al.*, 2014), which attempted to predict individual's choices from neural responses while passively viewing stimuli (Levy *et al.*, 2011; Smith *et al.*, 2014) or actively engaging an demanding attention task (Tusche *et al.*, 2010). In contrast to this study, all of the three studies focused on within-participant predictions (i.e. predicting a participant's choice between two items based on his/her brain activations in response to other items) and reported the choice prediction accuracy ranging from 56 to 83%. Smith *et al.* (2014) also tested across-group predictions, which is conceptually similar to our across-participant predictions and reported the prediction accuracy of 61.2%. Taken together with the findings of Smith *et al.* (2014), the present results indicate that neural activation patterns associated with attitudes are sufficiently similar across different individuals, suggesting a great potential for an objective measure of attitudes using a neuroimaging method.

We also found that attitude extremity can be predicted by activation patterns in more posterior parts of the striatum among other regions. This result indicates that both highly liked and highly disliked members induce similar patterns of activations in this region. Although the striatum is widely implicated in reward processing (Delgado, 2007), it is also known to respond to saliency (e.g. Zink *et al.*, 2004; Jensen *et al.*, 2007). As stimuli with high attitude extremity scores in this study are also likely to be highly salient, our results are consistent with the role of the striatum in processing saliency. In addition, the pattern of results we obtained (Figure 2b) clearly indicates a functional dissociation within the striatum; whereas the anterior striatum represent attitudes, the posterior striatum represents saliency or attitude extremity. Functional dissociation within the striatum found in this study is further supported by past studies which identified several subregions within the striatum based on intrinsic functional connectivity (Choi *et al.*, 2012) and patterns of coactivation with other cortical areas (Pauli *et al.*, 2016). Although the reports on how the striatum is organized differ slightly across studies, it seems clear that the two striatal regions found in this study lie in different subregions within the striatum. Furthermore, the anterior striatum's role in representing attitudes is consistent with the findings from large-scale coactivation data that this subregion is the most strongly involved in representing stimulus value (Pauli *et al.*, 2016) (note that the anterior striatum cluster found in this study seems to be the closest to the cluster labeled 'ventral striatum' in Pauli *et al.*, 2016). In contrast, however, posterior striatum identified in this study seems to be the striatal subregion most strongly associated with executive function (Pauli *et al.*, 2016), which is not necessarily consistent with our finding that this region is involved in attitude extremity or salience.

Nonetheless, our findings may be explained by two types of dopamine neurons systematically located in the midbrain and projections from the midbrain to the striatum. Matsumoto and Hikosaka (2009) found two types of dopamine neurons in the monkey midbrain; those that are excited by positive stimuli (juice) and inhibited by negative stimuli (airpuff to the eyes), and those that are excited by both the positive and negative stimuli (Matsumoto and Hikosaka, 2009). Consistently across two monkeys, those neurons which respond to both positive and negative stimuli (i.e. motivationally salient stimuli) are located in the dorsolateral substantia nigra pars compacta which projects mainly to the dorsal striatum, whereas those neurons which predominantly respond to positive stimuli are located in the ventromedial substantia nigra pars compacta (also ventral tegmental area) which send projections mainly to the ventral striatum (Haber and Knutson, 2010). Thus, striatal functional dissociation found in this study might reflect the activations of these two types of neurons in the midbrain. Thus, although speculative, the across-participant MVPA results may suggest that different populations of neurons encode reward (attitude) and saliency (attitude extremity), each of which is localized in a different subregion of the striatum and receives projection from dopamine neurons in a different subregion of the midbrain.

Conclusion

This study investigated the potential of a neuroimaging method to predict people's incidental social attitudes toward others. We found that patterns of activations in the anterior striatum can reliably predict an individual's attitudes toward famous people, suggesting the feasibility of such an approach. Although the prediction accuracy was higher for the self-report measure than the neural measure, this study represents an essential first step toward neural measures of social attitudes and demonstrated that we could successfully predict attitudes without asking participants to engage in any attitude-related task. Although we focused on explicit attitudes toward familiar people, it will be important in future research to see how accurately similar neural measures might predict implicit attitudes such as prejudice toward racial or social outgroups.

Supplementary data

Supplementary data are available at *SCAN* online.

Funding

This work was supported by the Japan Society for the Promotion of Science Fellows (to K.I. and K.S.).

Acknowledgement

We thank Atsushi Miyazaki for help with fMRI data collection.

Conflict of interest. None declared.

References

Amodio, D.M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience* **15**, 670–82.

Arkes, H.R., Tetlock, P.E. (2004). Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the implicit association test?". *Psychological Inquiry* **15**, 257–78.

Bartra, O., McGuire, J.T., Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76**, 412–27.

Blanton, H., Jaccard, J., Gonzales, P.M., Christie, C. (2006). Decoding the implicit association test: implications for criterion prediction. *Journal of Experimental Social Psychology* **42**, 192–212.

Bosson, J.K., Swann, W.B., Pennebaker, J.W. (2000). Stalking the perfect measure of implicit self-esteem: the blind men and the elephant revisited?. *Journal of Personality and Social Psychology* **79**, 631–43.

Buchel, C., Morris, J., Dolan, R.J., Friston, K.J. (1998). Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* **20**, 947–57.

Choi, E.Y., Yeo, B.T., Buckner, R.L. (2012). The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology* **108**, 2242–63.

Clithero, J.A., Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience* **9**, 1289–302.

Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., Groom, C.J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of Personality and Social Psychology* **89**, 469–87.

Cunningham, W.A., Johnson, M.K., Gatenby, J.C., Gore, J.C., Banaji, M.R. (2003). Neural components of social evaluation. *Journal of Personality and Social Psychology* **85**, 639–49.

Cunningham, W.A., Van Bavel, J.J., Johnsen, I.R. (2008). Affective flexibility: evaluative processing goals shape amygdala activity. *Psychological Science* **19**, 152–60.

Delgado, M.R. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences* **1104**, 70–88.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology* **84**, 3072–7.

DeMaio, T.J. (1984). Social desirability and survey measurement: a review. In: Turner, C.F., Martin, E., editors. *Surveying Subjective Phenomena*, 257–82, New York: Russell Sage.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems* **9**, 155–61.

Etkin, A., Egner, T., Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences* **15**, 85–93.

Fazio, R.H., Olson, M.A. (2003). Implicit measures in social cognition research: their meaning and use. *Annual Review of Psychology* **54**, 297–327.

Fiedler, K., Messner, C., Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": a logical and psychometric critique of the implicit association test (IAT). *European Review of Social Psychology* **17**, 74–147.

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**, 1464–80.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., 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.

Haber, S.N., Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* **35**, 4–26.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience* **28**, 5623–30.

Ito, T.A., Bartholow, B.D. (2009). The neural correlates of race. *Trends in Cognitive Sciences* **13**, 524–31.

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of National Academy of Sciences of the United States of America* **107**, 22014–9.

Izuma, K., Saito, D.N., Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron* **58**, 284–94.

Jensen, J., Smith, A.J., Willeit, M., *et al.* (2007). Separate brain regions code for salience vs. valence during reward prediction in humans. *Human Brain Mapping* **28**, 294–302.

Jimura, K., Poldrack, R.A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia* **50**, 544–52.

Kahnt, T., Park, S.Q., Haynes, J.D., Tobler, P.N. (2014). Disentangling neural representations of value and salience in the human brain. *Proceedings of National Academy of Sciences of the United States of America* **111**, 5000–5.

Karpinski, A., Hilton, J.L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology* **81**, 774–88.

Karpinski, A., Steinman, R.B., Hilton, J.L. (2005). Attitude importance as a moderator of the relationship between implicit and explicit attitude measures. *Personality and Social Psychology Bulletin* **31**, 949–62.

Knight, D.C., Nguyen, H.T., Bandettini, P.A. (2005). The role of the human amygdala in the production of conditioned fear responses. *NeuroImage* **26**, 1193–200.

Knutson, B., Westdorp, A., Kaiser, E., Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage* **12**, 20–7.

Knutson, K.M., Wood, J.N., Spampinato, M.V., Grafman, J. (2006). Politics on the brain: an FMRI investigation. *Social Neuroscience* **1**, 25–40.

Kohler, P.J., Fogelson, S.V., Reavis, E.A., *et al.* (2013). Pattern classification precedes region-average hemodynamic response in early visual cortex. *NeuroImage* **78**, 249–60.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of National Academy of Sciences of the United States of America* **103**, 3863–8.

Kubota, J.T., Banaji, M.R., Phelps, E.A. (2012). The neuroscience of race. *Nature Neuroscience* **15**, 940–8.

LaBar, K.S., Gatenby, J.C., Gore, J.C., LeDoux, J.E., Phelps, E.A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**, 937–45.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* **64**, 431–9.

Levy, D.J., Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology* **22**, 1027–38.

Levy, I., Lazzaro, S.C., Rutledge, R.B., Glimcher, P.W. (2011). Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *Journal of Neuroscience* **31**, 118–25.

Luttrell, A., Stillman, P.E., Hasinski, A.E., Cunningham, W.A. (2016). Neural dissociations in attitude strength: distinct regions of cingulate cortex track ambivalence and certainty. *Journal of Experimental Psychology: General* **145**, 419–33.

Matsumoto, M., Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–U834.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* **4**, 95–102.

Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., Tetlock, P.E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* **105**, 171–92.

Pauli, W.M., O'Reilly, R.C., Yarkoni, T., Wager, T.D. (2016). Regional specialization within the human striatum for diverse psychological functions. *Proceedings of National Academy of Sciences of the United States of America* **113**, 1907–12.

Petty R.E., Fazio R.H., Brinol P., editors. (2009). *Attitudes: Insights from the New Implicit Measures*. New York: Psychology Press.

Petty, R.E., Krosnick, J.A., editors. (1995). *Attitude Strength: Antecedents and Consequences*. *Mahwah*, NJ: Erlbaum.

Phelps, E.A., O'Connor, K.J., Cunningham, W.A., *et al.* (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience* **12**, 729–38.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., Podsakoff, N.P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology* **88**, 879–903.

Powell, M.C., Fazio, R.H. (1984). Attitude accessibility as a function of repeated attitudinal expression. *Personality and Social Psychology Bulletin* **10**, 139–48.

Sescousse, G., Caldu, X., Segura, B., Dreher, J.C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews* **37**, 681–96.

Sherman, J.W. (2009). Controlled influences on implicit measures: confronting the myth of process-purity and taming the cognitive monster. In: Petty R.E., Fazio R.H.,, Brinol P., editors. *Attitudes: Insights from the New Implicit Measures*, 391–426, New York: Psychology Press.

Smith, A., Bernheim, B.D., Camerer, C.F., Rangel, A. (2014). Neural activity reveals preferences without choices. *American Economic Journal: Microeconomics* **6**, 1–36.

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 National Academy of Sciences of the United States of America* **108**, 7710–5.

Tusche, A., Bode, S., Haynes, J.D. (2010). Neural responses to unattended products predict later consumer choices. *Journal of Neuroscience* **30**, 8024–31.

Tusche, A., Kahnt, T., Wisniewski, D., Haynes, J.D. (2013). Automatic processing of political preferences in the human brain. *NeuroImage* **72**, 174–82.

Van Boven, L., Judd, C.M., Sherman, D.K. (2012). Political polarization projection: social projection of partisan attitude extremity and attitudinal processes. *Journal of Personality and Social Psychology* **103**, 84–100.

Vickery, T.J., Chun, M.M., Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron* **72**, 166–77.

Wittenbrink, B., Schwarz, N., editors. (2007). *Implicit Measures of Attitudes: Procedures and Controversies*. New York: Guilford.

Zink, C.F., Pagnoni, G., Martin-Skurski, M.E., Chappelow, J.C., Berns, G.S. (2004). Human striatal responses to monetary reward depend on saliency. *Neuron* **42**, 509–17.