# Decoding “us” and “them”:

# Neural representations of generalized group concepts

Cikara, M.,[a] Van Bavel, J. J.,[b] Ingbretsen, Z.,[a] & Lau, T.[a]

[a] Department of Psychology, Harvard University, Cambridge, MA 02138

[b] Department of Psychology, New York University, New York, NY 10003

Corresponding author: Mina Cikara, 33 Kirkland St., Cambridge, MA 02138, 617.495.3819, mcikara@fas.harvard.edu

Keywords: intergroup relations | social categories | functional magnetic resonance imaging | multi-voxel pattern analysis

Word count: 6,594

Humans form social coalitions in every society on earth, yet we know very little about how the general concepts ‘us’ and ‘them’ are represented in the brain. Evolutionary psychologists have argued that the human capacity for group affiliation is a byproduct of adaptations that evolved for detecting more general coalitions. These theories suggest that humans possess a common neural code for the concepts ‘in-group’ and ‘out-group,’ regardless of the category by which group boundaries are instantiated. We used multivoxel pattern analysis to identify the neural substrates of generalized group concept representations. We trained a classifier to encode how people represented the most basic instantiation of a specific social group (i.e., arbitrary teams created in the lab with no history or associated stereotypes) and tested how well the neural data decoded membership along an objectively orthogonal, real-world category (i.e., political parties). The dorsal anterior cingulate cortex/middle cingulate cortex and anterior insula were associated with representing groups across multiple social categories. Restricting the analyses to these regions in a separate sample of participants performing an explicit categorization task, we replicated cross-categorization classification in anterior insula. Classification accuracy across categories was driven predominantly by the correct categorization of in-group targets, consistent with theories indicating in-group preference is more central than out-group derogation to group perception and cognition. These findings highlight the extent to which social group concepts rely on domain-general circuitry associated with encoding stimuli’s functional significance in service of responding adaptively to dynamic environments.

# Decoding “us” and “them”:

# Neural representations of generalized group concepts

Humans reliably carve up the social world into ‘us’ and ‘them.’ This practice is a human universal—present in every society on earth (Brown, 1991). The variety of groups with which humans affiliate is vast, ranging from race and religion to political affiliation and nationality. These groups vary on countless dimensions, from visual cues to group membership to associated stereotype content. Evolutionary psychologists have argued that categorizing people by specific social categories (e.g., race) is a byproduct of adaptations that evolved for detecting more general coalitions (Cosmides, Tooby & Kurzban, 2003; Sidanius & Pratto, 2004; Pietraszewski, Cosmides, & Tooby, 2014). For example, briefly focusing people on adversarial mixed-race groups temporarily diminishes the extent to which they encode race (Kurzban, Tooby, & Cosmides, 2001). Even 5-year olds prefer other-race children to same-race children when the other-race children speak with native rather than foreign accents (Kinzler, Shutts, DeJesus, & Spelke, 2009). Thus, humans may have a flexible, common neural code for representing the concepts ‘in-group’ and ‘out-group,’ invariant to the particular social category by which group boundaries are instantiated.

By some accounts, encountering a new individual activates three “primary’’ social categories—race, sex, and age—which the mind automatically encodes (e.g., Hamilton, Stroessner, & Driscoll, 1994; Fiske & Neuberg, 1990). Each of these social categories is associated with unique stereotypes and social knowledge. Though researchers have discovered a great deal about what regions of the brain track race, gender, and other significant social categories (see Amodio, 2014; Kubota, Banaji, & Phelps, 2012; Ito & Bartholow, 2009 for reviews), far less is known about which brain networks track coalitions—in-groups and out-groups—more generally (see Cikara & Van Bavel, 2014 for a review). Cultural stereotypes, personal experiences, and other category-specific attributes lead to distinct activation patterns

for different social categories (Stolier & Freeman, 2016). However, it is unknown whether the specific patterns of brain activity associated with distinguishing in-group from out-group members along one social category are also associated with distinguishing groups along other, unrelated social categories. If so, it would indicate a generalized representation of 'us' and 'them' abstracted away from the specific features associated with any one category. The current research tests this hypothesis.

Although social group representations appear to be widely distributed across the brain, several candidate brain regions have previously been linked to different representational dimensions theorized to support group representation. For instance, judgments of similar others engage medial prefrontal cortex (mPFC) and draw on overlapping neural populations engaged during self-referential thought (Mitchell, Macrae, & Banaji, 2006; Jenkins, Mitchell, & Macrae, 2008). Thus some have interpreted greater mPFC response to in-group relative to out-group targets as evidence that similarity to one's self is the central dimension along which we group others (e.g., Morrison, Decety, & Molenberghs, 2012). Alternatively, it is possible that generalized social group representations are organized vis-à-vis their functional significance (e.g., good or bad for me?), and therefore draw on domain-general circuitry associated with coordinating behavior in response to motivationally relevant stimuli (Cikara & Van Bavel, 2014). For instance, the anterior insula and dorsal anterior cingulate cortex are part of a broader network that focuses attention on the most relevant among internal and extrapersonal stimuli (both social *and* non-social) in order to select among competing behavioral repertoires (e.g., freeze, fight, flight; Legrain, Iannetti, Plaghki, & Mouraux, 2011; Mennon & Uddin, 2010; Decety, Norman, Berntson, & Cacioppo, 2012). In this view, social categorization is not 'special', but instead co-opts circuitry associated with representing both social and non-social dimensions (e.g., evaluation; see Ruff & Fehr (2014) for a similar discussion regarding the neurobiology of reward and value in social and non-social decision-making).

To examine whether people possess a common neural code associated with the

representation of the concepts of ‘us’ and ‘them,’ we employ cross-categorization multi-voxel pattern analysis (MVPA). This approach was inspired by neuroimaging analyses of supramodal representations of numbers (Eger Sterzer, Russ, Giraud, & Kleinschmidt, 2003), objects (Pietrini et al., 2004; Tanaka, 1993), and emotions (Peelen, Atkinson, & Vuilleumier, 2010; Skerry & Saxe, 2014). MVPA uses the information carried by fine-grained patterns of blood oxygenation level-dependent (BOLD) activity within different brain regions to ‘decode’ the representation of different categories of stimuli or visual features (Haynes and Rees, 2006; Mur, Bandettini, & Kriegeskorte, 2009; Norman, Polyn, Detre, & Haxby, 2006). Unlike traditional univariate analysis, MVPA uses pattern classification algorithms to map categories of stimuli or psychological states to brain activity. In short, MVPA allows investigators to examine whether different neural patterns of activation within specific brain regions—which may have the same mean-level of activation, and would therefore go undetected by traditional univariate analysis—distinguish between different psychological representations. Furthermore, we probe the specific constellation of hits and errors made by the classifier to determine whether in-group and out-group members are represented uniquely or whether a domain-general dimension (e.g., salience, threat) drives generalized social group representation (see below).

*Overview of the current experiments*

On one hand, brain networks previously associated with representing social groups along one social category be used to represent groups along other, unrelated social categories. On the other hand, cultural stereotypes, personal experiences, and other attributes specifically associated with certain categories might lead to distinct activation patterns for every category a person encounters. Here we explore a third possibility: social group representation, like object or emotion recognition, may rely on the integration of features along a hierarchy of increasingly abstract “feature-detectors” (Martin, 1994). In this framework, people should exhibit neural substrates associated with representing specific categories that are distinct from those associated with representing generalized group concepts: ‘us’ and ‘them.’

We assigned participants to arbitrary groups and had them complete a task in which social categorization was incidental to the task at hand. We then examined the whole brain for regions involved in decoding group membership across arbitrary and real political groups. This approach allowed us to distinguish the neural substrates of category-specific representations from the neural substrates of generalized social group concept representations. In the *within-category classification*, we trained the classifier to encode how people represented in-group and out-group targets of one kind (e.g., Democrats versus Republicans) and then tested how well the neural data decoded group membership within that same category but in data from a run that had been excluded from training (i.e., the leave-one-run-out cross-validation approach). Conversely, in the *cross-category classification* we trained a classifier to encode how people represented the most basic instantiation of a single social category (i.e., arbitrary groups created in the lab that have no history or associated stereotypes) and tested how well the neural data decoded membership along an objectively orthogonal, real-world social category boundary (i.e., political party membership).[1]

We focused on political affiliation as the test-case for cross-category classification because, unlike race and many other significant social categories, it is not confounded by visual cues to group membership. Moreover, recent evidence suggests that implicit bias and behavioral discrimination along political boundaries is now as potent as bias against racial out-groups in some domains (Iyengar, Sood, & Lelkes, 2012; Iyengar & Westwood, 2015; Motyl, Iyer, Oishi, Trawalter, & Nosek, 2014).

Of course our social worlds are more complicated than just ‘us’ and ‘them.’ Although it is safe to ignore or neglect many out-groups, there are some out-groups who pose an active threat, to whom we must attend in order to protect the in-group. To better model these conditions, we included three classes of targets within each social category: in-group, a neutral out-group, and a competitive, and therefore threatening out-group. We examined the precise pattern of the classifier’s hits and errors to glean insight into whether friends, not-friends, and

foes are represented uniquely or whether a more general process underlies generalized social group representation. For example, if the classifier mistakes in-group and competitive out-group members more often than in-group and neutral out-group members, this would suggest group representation is driven by identifying salient targets, irrespective of their allegiances. If, however, the classifier mistakes in-group and neutral out-group members more often than in-group and competitive out-group members, this would suggest that group representation is driven by identifying active threats. Finally, if the classifier mistakes neutral and competitive out-group members most often, this would suggest that group representation is driven by a binary "us" versus "them" distinction.

Finally, we ran a second confirmatory experiment with a separate sample of participants, restricting the analyses to the regions identified in Experiment 1. In a modified task, in which participants explicitly categorized targets as arbitrary or political in-group and out-group members, we examined whether cross-categorization classification remained significantly above chance.

## Methods

*Participants*. Forty-eight participants (27 female; $M_{age}$ = 22.5) were recruited via flyers and word-of-mouth from the university and paid $40 (base pay plus bonus: see below) for their time. The final sample for Experiment 1 included 19 Democrats and 6 Republicans. For Experiment 2 we recruited only self-identified Democrats. All were right-handed, native English speakers with normal or corrected vision, with no history of psychiatric or neurological problems. We obtained written informed consent; procedures complied with the university's institutional review board's guidelines.

Because both experiments were entirely within-subject designs we aimed for a minimum of 20 participants within each experiment to attain 80% power to detect a moderate-sized effect. For Experiment 1, one participant was excluded because she was not able to clearly see the stimuli and one participant did not complete the entire scan session, leaving 23 participants (17

Democrats and 6 Republicans). For Experiment 2, two participants were excluded from analysis because of excessive head movement (greater than 2mm) while in the scanner, and one participant was excluded because of technical problems during scanning, leaving 20 participants (all Democrats). We found no gender differences on any of our outcome variables. We report all data exclusions, manipulations, conditions, and measures in both experiments. Summary data, analysis code, and materials are available at: https://osf.io/g9rth/

*Experiment 1 pre-test measures: Team assignment, group affiliation, demographic information*. Approximately one to two weeks prior to scanning, each participant completed a series of online questionnaires. First, participants were told that they would be assigned to a team for the experiment. Second, participants indicated the strength of their agreement with five personality items, ostensibly for the purposes of team assignment. In reality, each participant was randomly assigned to one of two competitive teams, the Eagles or the Rattlers. Third, participants answered 12 questions assessing their propensity to value and join groups (e.g., "The social groups we belong to are one of the most important things in our lives" and "We are defined, at least in part, by the social groups that we belong to"; *strongly disagree* (1) to *strongly agree* (7), Cronbach's $\alpha = .69$; Dunham & Van Bavel, in prep). Finally, participants answered two manipulation check questions ("What team are you on?" and "Against which team will you be competing?") and completed demographic information (i.e., age, gender, ethnicity, college year, political party affiliation, and extent to which they were socially and fiscally *liberal* (1) to *conservative* (7)).

*Experiment 1 procedure.* We told participants they were taking part in a functional magnetic resonance imaging (fMRI) study of spatial location and information processing and that they would compete in a problem-solving challenge against a member of the competitive out-group afterward (i.e., a previous participant who was returning to play on behalf of the opposing group): "We are going to investigate how your neural responses during the scan relate

to your performance during the competition after the scan." Figure 1 provides an overview of the procedure.

Participants then completed a series of pre-scan tasks. Participants read the following introduction: "As we've told you before, you are going to participate in an ongoing problem solving challenge between two teams: the EAGLES team and the RATTLERS team. First you will complete the fMRI scan and then you will compete in the challenge. Depending on the performance of you and your team, you may win money in addition to the basic participation payment. Specifically you will receive $30 for participating, but you could increase your earnings to $40 if your team performs well. Please note: Some people do not fit the profile of either the Eagles or the Rattlers. We're assigning these people to the BEARS team. You will not meet or compete with anyone from this team." Everyone received the $10 bonus at the end.

As a manipulation check, participants were asked to report the name of their team; all but one participant correctly recalled their team assignment from the pre-test survey. We also showed participants a social network diagram illustrating that they were much more similar to their teammates (and that the competing players were much more similar to one another) than the groups were to each other because greater group cohesion increases intergroup bias (manipulation was identical to that in Experiment 4 in Cikara, Bruneau, Van Bavel, & Saxe, 2014). We explained that the participants' own team had accumulated 82 points whereas the other team had earned 84 points indicating that it was a tight race. Whichever team had the higher score at the end of the experiment would win the bonus. Participants then indicated how much they agreed with the following statements: "I [like/value/feel] connected the [Eagles/Rattlers] group" (*Strongly Disagree* (1) to *Strongly Agree* (100) see (Cikara et al., 2014)). Both team identification scales showed good reliability (Cronbach's α for in-group = .75, out-group = .68). We averaged across the three items to generate arbitrary in-group and out-group identification scores, respectively (see Cikara et al., 2014).

Next participants completed a series of practice trials for the main task. Participants' task was to identify in which of four quadrants a statement appeared (upper left/right, lower left/right corresponding to the 1, 2, 3, and 4 buttons). Each statement took the structure of "[name] is a [group]" (e.g., "Don is a Rattler.") and stayed on the screen for 4s. The names were gender-matched to the participant to avoid activating gender as another social category. There were six group labels: Eagle, Rattler, Bear, Democrat, Republican, Constitutional, creating a 3 (Group: in-group, out-group, neutral group) × 2 (Social Category: arbitrary team, political party) design. Participants saw two example trials and were quizzed on the correct answer for another two sample trials.

At the end of the pre-scan task we explained: "You may have noticed that there is reference to a team called the Bears. These people do not fit the profile of either a Rattler or an Eagle, but we give them an opportunity to participate and earn money anyway. They are not a part of the problem challenge. You are not in competition with them. Remember, you have been assigned to your team based on your personality item responses. Your teammates are people, who share your traits and who have already been scanned or will be scanned in the next few weeks. Depending on the performance of you and your team, you and your teammates may win money in addition to the basic participation payment. Specifically you will receive $30 for participating today, but we will send you a $10 bonus if at the end of the study your team has outperformed the other team!" We also clarified that participants would have an opportunity to be the opponent for a future participant for further compensation if they were interested in signing up.

Participants first underwent an anatomical scan. Then they started the main task, which included 10 runs. Each run included 4 trials of each of the 6 conditions, resulting in 24 trials per run (~5 minutes). Trials were interleaved with a jittered ITI, which ranged from 6 to 20 seconds. Each trial within each condition appeared in a different quadrant to ensure condition was not correlated with location. At the end of each run, we asked participants the contents of the last

statement to confirm they were actually encoding the content of the statement in addition to its location. On average, participants reported the correct quadrant on 97% of trials (95% when we coded trials in which participants did not respond within 4 s as errors), indicating they were paying attention to the task. After the scan, participants were told they would not actually have to participate in a challenge against another player and were fully debriefed about the purpose of the experiment.

*Experiment 2 pre-test measures.* Team assignment followed the same procedure as in Experiment 1, except that all participants were assigned to the Eagles team. Participants reported their party affiliation and the extent to which they "like/value/feel connected to" their political party (α = .83, *M* = 62.96).

*Experiment 2 procedure.* The procedure for Experiment 2 differed from Experiment 1 in only a few ways. First, we omitted the spatial location and information processing cover story. Instead, we explained, "this is an fMRI study examining the effect of cooperation and competition on the mind and brain. We are going to investigate how your neural responses during the scan relate to your performance during the competition after the scan." The problem-solving challenge structure remained the same in order to set up a competitive relationship between the Eagles and Rattlers.

As in Experiment 1, participants reported their team membership, saw the social network diagram, and indicated how much they agreed with the following statements: "I [like/value/feel] connected the [Eagles/Rattlers] group" (*Strongly Disagree* (1) to *Strongly Agree* (100) see (Cikara et al., 2014)). Both team identification scales showed good reliability (α for in-group = .89, out-group = .76). Participants then completed a practice round of the main fMRI task. We instructed them: " you will be reading descriptions of people. Please tell us as quickly as possible whether or not the person described is a member of your group. The descriptions will vary in terms of personality or political membership." Participants were instructed to press the

"1" key if it was an in-group member, and the "2" if it was an out-group member. There was no third neutral party in this experiment (i.e., no references to Bears or Constitutionals).

The explicit categorization task utilized an event-related design. Participants completed two runs of the task. Each run included 13 trials of each of the 4 conditions, resulting in 52 trials per run (~8m). On each trial, participants read statements of the form "[X] is a [group]" and indicated, using a button box, whether the person described was an in-group or an out-group member. Statements appeared for 2s followed by a fixation cross, which appeared for 2-16s (jittered). Participants were required to respond within the 2s window during which the statement was presented.

*fMRI Acquisition*. At the beginning of each scan session, we acquired a high-resolution T-1 weighted anatomical image (T1-MPRAGE, 1 × 1 × 1 mm) for use in registering activity to each participant's anatomy and spatially normalizing data across participants. Echo-planar images were acquired using a Siemens Magnetom Verio 3T System (Siemens Solutions, Erlangen, Germany) at the Scientific Imaging & Brain Research Center (TR = 2000 ms, TE = 29 ms, field of view = 192 mm, matrix size = 64 × 64). Near whole-brain coverage was achieved with 36 interleaved 3.0 mm near-axial slices.

*fMRI preprocessing and data analysis.* SPM8 (www.fil.ion.ucl.ac.uk/spm/software/spm8/) was used to analyze each participant's MRI data, which were motion corrected and then normalized and resliced to 2mm x 2mm x 2mm voxels onto a common brain space (Montreal Neurological Institute, EPI Template). Functional images were motion-corrected within-run to the first image of each run, then coregistered to the anatomical image. Normalization warp was produced by SPM combined segmentation and normalization and then applied to the anatomical image and the coregistered functionals.

We first built a general linear model of the experimental design, and used this model to analyze the BOLD response in each voxel. In Experiment 1, the model included the 6 regressors of interest (in-group/out-group/neutral × arbitrary team/political party) as well as

nuisance regressors (run effects and time and dispersion derivatives). In Experiment 2, the model included 4 regressors of interest (in-group/out-group × arbitrary team/political party) as well as nuisance regressors. Each event consisted of the TRs during which each statement was presented on the screen. We modeled the conditions as a boxcar (matching the onset and duration of each event) convolved with a standard hemodynamic response function (HRF). This process generated 10 betas per regressor (1 beta image per run) in Experiment 1, and two betas per regressor in Experiment 2.

*Classification*. The MVPA analyses were conducted in PyMVPA, using a LIBSVM classifier. For Experiment 1, we used an n-fold cross-trainer partitioner, which allowed the linear CSVMC classifier to train on 9 runs of one class of groups and decode the other class of groups from the run that was left out. For example, the classifier would train on arbitrary team in-group/out-group/neutral betas from runs 1−9 and then test on political in-group/out-group/neutral betas from run 10. The partitioner iterated through all 10 runs and averaged across the folds to generate accuracy scores. We ran the cross-category classifier twice—once training on arbitrary and testing on political groups and once training on political and testing on arbitrary groups. We also ran two within-category classifiers: training on arbitrary, then testing on arbitrary; training on political, then testing on political.

A searchlight (3-voxel radius, (Kriegeskorte, Goebel, & Bandettini, 2006)) analysis across the whole brain assigned a classification accuracy value to each voxel, from which we subtracted chance accuracy (33.33% or 1/3). We then smoothed the accuracy maps to 6 mm FWHM (note that there was no smoothing applied during pre-processing). To reduce the number of comparisons across the whole brain, we generated a mask using FSL's MNI structural atlas that masked out the cerebellum, brain stem, ventricles, occipital lobe, and white matter. We chose to exclude occipital lobe because the stimuli were text-based and we controlled for the number of characters in each statement string across conditions. Finally, we conducted a group-level one-tailed t-test against zero to determine which voxels exhibited

classification accuracy significantly greater than chance. AFNI's 3dClustSim Monte Carlo simulation determined a minimum cluster size of 87 contiguous voxels to achieve corrected $p < 0.05$ given a voxelwise threshold of $p < 0.005$.

Experiment 2 restricted analyses to the regions identified by the searchlight classifier in Experiment 1. The linear CSVMC classifier trained on one run of one class of groups and tested the other class of groups from the run that was left out. The partitioner iterated through both runs and averaged across the folds to generate accuracy scores. We ran the classifier twice—once training on arbitrary and testing on political groups and once training on political and testing on arbitrary groups—and averaged across both before testing against chance in each ROI identified in Experiment 1.

## Results

*Behavior*. In a pre-scan survey, Experiment 1 participants evaluated their own arbitrary team ($M$ = 70.80, $SE$ = 3.06) much more positively than the other team ($M$ = 30.04, $SE$ = 3.04), 95% CI [30.95, 50.56], $t(22) = 8.62$, $p<.0001$, $d = 2.79$. Additionally, Republicans ($M$ = 5.00, $SE$ = 0.68) were more socially conservative than Democrats ($M$ = 2.53, $SE$ = 0.21), 95% CI [1.37, 3.57], $t(21) = 4.65$, $p = .0001$, $d = 2.21$. Republicans (M = 5.83, SE = 0.40) were also more fiscally conservative than Democrats ($M$ = 3.53, $SE$ = 0.29), 95% CI [1.18, 3.43], $t(21) = 4.27$, $p = .0003$, $d = 2.03$. These analyses served as a manipulation check on participants' arbitrary group and political preferences.

Participants' self-reported propensity to join and value groups did not differ by team or political party, $p$s > .46. Furthermore, A 3 (in-group/out-group/neutral) × 2 (arbitrary team/political party) multilevel model (treating participant as a random effect) on response times (indicating where on the screen each statement appeared) indicated no main effects of group, $F(2, 5384) = 1.284$, $p = 0.277$, or category, $F(1, 5384) = 0.003$, $p = 0.955$, nor a group × category interaction, $F(2, 5384) = 0.154$, $p = 0.858$.

In a pre-scan survey, Experiment 2 participants also evaluated their own arbitrary team ($M$ = 62.22, $SE$ = 3.92) more positively than the other team ($M$ = 32.40, $SE$ = 3.20), 95% CI [17.31, 42.33], $t(19) = 4.99$, $p < .0001$, $d = 1.86$. Thus, participants exhibited a strong degree of arbitrary in-group favoritism across both experiments. In contrast to Experiment 1, a 2 (in-group/out-group) × 2 (arbitrary team/political party) multilevel model treating participant as a random effect on response times indicated a significant main effect of group, $F(1, 2006) = 9.307$, $p = 0.002$), but not of category, $F(1, 2006) = 0.431$, $p = 0.512$, and a marginal group × category interaction, $F(1, 2006) = 3.168$, $p = 0.075$. Specifically, participants were faster to identify in-group ($M$ = 0.84, $SE$ = 0.21) than out-group trials ($M$ = 0.87, $SE$ = 0.22). Because Experiment 2 was purely confirmatory, and analyses were restricted to the regions identified in Experiment 1 (in which there was no response time difference across groups or categories), we do not discuss these findings further.

*fMRI*. First, we ran the two within-category classifiers on the data from Experiment 1: train and test on arbitrary groups, and train and test on political parties, respectively (see Table 1 for results). Both the arbitrary team classifier and the political affiliation classifier identified regions of the medial prefrontal cortex (mPFC)—ventral and dorsal, respectively—and inferior temporal gyrus (ITG) as regions in which group membership could be decoded above chance. The arbitrary group classifier additionally identified left anterior insula/inferior frontal gyrus (AI/IFG) and pregenual anterior cingulate cortex (pgACC). The political affiliation classifier additionally identified two distinct clusters in right middle frontal gyrus (MFG) and left superior temporal gyrus (STG). Accuracy in these clusters did not differ by participant gender, team, or political party.

Next, we examined whether multi-voxel patterns associated with distinguishing arbitrary teams could successfully decode political party membership. A three-way classifier trained on arbitrary teams and tested on political parties identified left anterior insula (AI) and dorsal anterior cingulate cortex/middle cingulate cortex (dACC/MCC; Table1, Figures 2a,b) as brain

regions supporting cross-category group representation. Accuracy in these two clusters did not differ by participant gender, team, or political party.

The classifier trained on political parties and tested on arbitrary teams identified only left rostrolateral prefrontal cortex (RLPFC; Table1, Figure 2c). This cluster was directly anterior to, but did not overlap with, the left AI cluster identified in the arbitrary-to-political classifier. (At a voxel-wise threshold $p<.05$, 15.80% of voxels overlap.) Again, accuracy in this cluster did not differ by gender or team ($p$s>0.54), but did vary as a function of political party: Democrats were 4.12% above chance whereas Republicans were only 0.13%, 95% CI [1.19%, 6.79%], $t(21) = 2.97$, $p = .0074$, $d = 1.41$. Accuracies across these three regions are plotted in Figure 2d.

We then extracted the confusion matrix for each cross-category cluster to examine the precise pattern of hits and errors made by the classifier (Figure 2e; note that these results are purely descriptive because it is inappropriate to run another set of statistical analyses on data from voxels *selected by* the pattern of interest). Note that in this first experiment there were three groups so chance classification = 33.3%. Across the ACC and AI, the success of the classifier was driven mostly by the correct classification of in-group targets (see first column of each matrix). That is, when the target was an in-group member, the classifier guessed 'in-group' correctly 42.9% (dACC/MCC) and 41.0% (left AI) of the time: the highest hit rate in either of these region's confusion matrices. In contrast, classification success in the left RLPFC was driven by correctly guessing 'in-group' for in-group targets (38.0%) as well as 'neutral' for neutral targets (39.1%).

We then ran a confirmatory test of the same cross-category classification accuracy in a separate sample of participants, who explicitly categorized Rattler/Eagle/Democrat/Republican targets as in-group and out-group members (again, there was no third neutral out-group in this experiment). We restricted the analysis to the MCC and left AI regions identified in Experiment 1 and averaged across both cross-category classification accuracies (i.e., train on arbitrary, test on political as well as train on political, test on arbitrary). In this second experiment there were

only two groups so chance classification = 50%. We found that the patterns in AI could distinguish in-group from out-group targets significantly better than chance, accuracy = 59.38%, $t(19) = 2.32$, $p = .032$, $d = 0.52$. In-group and out-group accuracies in MCC, however, were not significantly different from chance in this new sample, accuracy = 47.50%, $t(19) = -0.59$, $p = .56$. Note that accuracies in the two clusters were also (marginally) significantly different from one another, $t(19) = 2.1391$, $p = 0.046$, $d = 0.64$.

## Discussion

Coalition-based cognition appears common across all human cultures. Preference for own group members has also been documented across species ranging from rats (e.g., Bartal, Decety, & Mason, 2011) to chimpanzees (e.g., Wrangham, 1996). Several candidate brain regions have previously been linked to different dimensions along which in-group and out-group may be represented (e.g., self-similarity). In contrast to those dimensions that are uniquely social, we found that that general social group concept representation appears to rely on circuitry that encodes external stimuli's functional or evaluative significance. The current experiment allowed us to distinguish the neural substrates associated with lower-level group-specific representations (e.g., Democrats versus Republicans) from the neural substrates of higher-level conceptual representations of "us" and "them." Strikingly, we found that abstract group representations in left AI generalize to other samples of participants engaged in an explicit social categorization task.

*Group-specific representations*. We used within-category classification (e.g., train and test on arbitrary groups) to identify the neural substrates of group-specific representation. The arbitrary team classifier identified mPFC and pgACC as well as left IFG and ITG. The mPFC/pgACC finding is particularly noteworthy because it is precisely these regions that have been identified in research examining self/similar-other representation overlap (e.g., Jenkins et al., 2008) as well as categorization of in-group relative to out-group labels (Morrison et al.,

2012). Note, however, that the political party classifier identified a different, non-overlapping network of regions including, dmPFC, two distinct regions of right MFG, left STG, and right IFG.

Why is there no overlap between these two maps? During the training phase, the machine-learning algorithm weighs most heavily the voxels that most strongly distinguish the groups in calculating the decision hyperplane, which is then used in the testing phase. Because people have richer social knowledge about political parties, the particular voxels that maximally distinguish among Democrats, Republicans, and Constitutionals are different than the voxels that distinguish among Eagles, Rattlers, and Bears.

These results may indicate that different psychological dimensions maximally distinguish in-group, neutral out-group, and threatening out-group target representations depending on the social category under consideration. In other words, we cannot conclusively rule out the possibility of category-dependent task discrepancies. Despite the fact that the task instructions were the same across categories, participants may have recruited different features or exemplars in the arbitrary versus political category conditions. For example, because participants were told they were assigned to their arbitrary teams based on their personality profiles, similarity to the self may have been the most diagnostic dimension along which to differentiate the targets in the arbitrary group classifier. This was not the case in the political party classifier, which selected a different network (perhaps because the stimuli activated representations of specific Democrats and Republicans). Nevertheless, these data suggest that at group-specific levels of representation, participants possess unique associated neural codes, which are distinct from one another and distinct from generalized group concept representations.

*Generalized group concept representations*. Across dACC/MCC and AI, patterns of brain activity distinguishing between arbitrary teams, created in the lab, successfully discriminated targets' political parties. We also replicated cross-categorization classification in left AI in a separate sample of participants completing an explicit social categorization task.

Classification accuracy in these regions was driven by sensitivity to in-group targets. This result is consistent with the social psychological research indicating that in-group preference is more central than out-group derogation to social identification processes (Balliet, Wu, & De Dreu, 2014). This evaluative preference is highly predictive of discrimination, stereotype activation, trust, and empathy, among many other psychological and behavioral manifestations of intergroup bias (Hewstone, Rubin, & Willis, 2002). Furthermore, many experiments indicate in-group preference is activated automatically (Nosek, Greenwald, & Banaji, 2007) suggesting that evaluation or valence is the key dimension along which generalized social group representations are organized.

The salience of group identities is context-dependent and alliances are dynamic. As such, humans have to be immediately responsive to re-categorization of any given target. Anatomically, dACC/MCC and AI are reasonable candidates to support social, motivationally-relevant representations and to regulate subsequent behavior given their anatomical connectivity with sensory and motor systems (Seeley et al., 2007; Uddin, 2015; Vogt, 2005). Future experiments could examine how these networks respond to flexible re-categorization. For instance, if a player is traded to the other team, are the patterns of activation used to represent out-group members now applied to the former in-group members?

In contrast, the classifier trained on political parties and tested on arbitrary teams identified only left RLPFC, which shared 16% overlapping voxels with the left AI cluster at a lower significance threshold. By some accounts left RLPFC is central to relational integration (Christoff et al., 2001; Kroger et al., 2002). Specifically, left RLPFC appears to play a domain-general role in integrating the higher-order relationships between task-relevant knowledge representations (Bunge, Helskog, & Wendelken, 2009; Westphal, Reggente, Ito, & Rissman, 2015). For example, left RLPFC is engaged more when participants evaluate the concordance of an analogy (e.g., “shoe is to foot as glove is to hand?”) as compared to when they complete an analogy (Wendelken et al., 2008). One highly speculative explanation is that participants

were spontaneously evaluating the concordance of the groups across categories (e.g., Eagles:Democrats?); however, we have no means of testing whether this is the case in the current dataset.

Why did the classifier trained on political party and tested on arbitrary groups identify a different (albeit partially overlapping) region of frontal cortex relative to the other cross-category classifier? As we note above, the dimensions of social knowledge that maximally distinguish among Democrats, Republicans, and Constitutionals may be different than the dimensions that distinguish among Eagles, Rattlers, and Bears. That said, the party-to-team cross-category classifier was nevertheless successful, indicating that even when different attributes are driving classification, those attributes are represented similarly across different kinds of in-group, neutral out-group, and threatening out-group targets.

*Fundamental psychological dimensions distinguishing generalized 'us' from 'them.'* Though these experiments were not designed to identify which specific psychological dimensions distinguish representations of the generalized concepts of 'us' and 'them,' the present findings aid in narrowing the hypothesis space. Classification success was primarily driven by the correct identification of in-group targets, such that threatening and neutral out-groups were most often confused for one another. These results do not support a threat-driven classification scheme, which predicted that in-group and neutral out-groups would be most often confused for one another, or a familiarity-driven classification scheme, which predicted that in-group and threatening out-groups would be most often confused. If the fundamental dimension distinguishing the concepts of 'us' and 'them' were similarity to one's self, we would predict a cross-categorization map that included mPFC/pgACC (as we observed in the within-arbitrary team classifier).

The behavioral results speak to the likelihood of two further variables as the fundamental dimensions driving classification success. The response time data from Experiment 1 indicated that the in-group trials (in either category) did not engender significantly faster or slower

responding than out-group trials. These results indicate that accessibility is unlikely to be a good candidate dimension driving the distinction between in-group and out-group concepts. Further, the self-reported team evaluation scores revealed that participants did *not* feel more positively about the in-group than they felt negatively about the out-group. Specifically, in Experiment 1 the average in-group team evaluation was 21 points above the midpoint of the scale, and the average out-group team evaluation was 20 points below the midpoint of the scale—making them roughly equivalent in terms of affective significance. Thus evaluative asymmetry (feeling more strongly about the in-group than the threatening out-group, irrespective of valence) is an unlikely candidate dimension as well.

Our results suggest that valence, specifically functional significance or evaluation (i.e., will this stimulus help me or not?), is a plausible candidate for the dimension distinguishing representations of 'us' and 'them.' This hypothesis is consistent with our results as well as decades of theorizing that emphasizes the priority of functional relations (i.e., whether people or groups are cooperative, competitive, or independent in their outcomes) as an organizing principle for group-related perception and cognition (Campbell, 1958; Fiske, Cuddy, & Glick, 2007; Sherif, Harvey, White, Hood, Sherif, 1961; Tajfel & Turner, 1979; see Cikara & Van Bavel, 2014 for a review).

*Future directions*. While our results suggest that social group concepts rely on domain-general circuitry associated with encoding stimuli's valence or functional significance, a more precise test would include both social and non-social stimuli that varied along these dimensions. For example, future experiments could include trials of positively and negatively valenced stimuli that hold little instrumental value as well as stimuli that have high versus low instrumental value to determine whether the same regions that classify the general concepts of 'us' and 'them' also classify non-social targets that vary along these dimensions. It is possible that these social and non-social representations overlap in space but are still associated with discriminable patterns of activation.

Perhaps the most interesting aspect of these results is that the classifier exhibited an over-inclusion bias (i.e., guesses “in-group” more often than any other label). On one hand this could be interpreted as a maladaptive tendency: surely it would be safer to err on the side of assuming everyone is an out-group member until one is certain of a target’s in-group membership. However, in their daily lives, people tend to interact more often with individuals who are similar to themselves along numerous demographics and dimensions (McPherson, Smith-Lovin, & Cook, 2001). Given the statistical regularities of individuals’ self-selected environments, it makes sense that their priors would bias them to assume other novel targets are in-group members until they receive information that indicates otherwise. Note also that even though the groups in question are competitive, they are not associated with threats to individuals’ physical well-being. Future experiments should examine whether running the same experiment with physically threatening groups, or in social contexts characterized by threats to physical safety (e.g., Gaza) would yield the opposite effect: an over-*exclusion* bias. Finally, we only used political and arbitrary groups—and our Experiment 2 did not include Republican participants. Liberals and conservatives exhibit differences in negativity bias (e.g., Hibbing, Smith, & Alford, 2014) and the extent to which they see themselves as similar to in-group members (e.g., Stern, West, & Schmitt, 2013). Thus participants could be weighing target features differently when representing in-group and out-group members depending on their political orientation (though note that we did not observe differences in classification accuracy by political party). To put the generalizability property to an even more stringent test, future experiments should test other orthogonal coalitional boundaries including demographic (e.g., race, nationality) and self-selected groups (e.g., religion, sports team affiliation).

*Conclusion*. The current work suggests that humans possess a common neural code for the concepts ‘in-group’ and ‘out-group,’ regardless of the category by which group boundaries are instantiated. These findings both shed light on neural circuits that have been proposed by

evolutionary psychology to support coalitional concepts and generate novel hypotheses at the intersection of social cognition, biological anthropology, and cognitive neuroscience.

**References**

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670-682.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556.

Bartal, I. B. A., Decety, J., & Mason, P. (2011). Empathy and pro-social behavior in rats. *Science*, *334*(6061), 1427-1430.

Brown, D. E. (1991). Human universals. New York, NY: McGraw-Hill.

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *Neuroimage*, *46*(1), 338-342.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3*, 14–25.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., et al. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage, 14(5)*, 1136–1149.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of experimental social psychology*, *55*, 110-125.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations an integrative review. *Perspectives on Psychological Science*, *9*(3), 245-274.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in cognitive sciences*, *7*(4), 173-179.

Decety, J., Norman, G. J., Berntson, G. G., & Cacioppo, J. T. (2012). A neurobehavioral evolutionary perspective on the mechanisms underlying empathy. *Progress in neurobiology*, *98*(1), 38-48.

Eger, E., Sterzer, P., Russ, M. O., Giraud, A. L., & Kleinschmidt, A. (2003). A supramodal number representation in human intraparietal cortex. *Neuron*,*37*(4), 719-726.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology*, *23*, 1-74.

Hamilton, D. L., Stroessner, S. J., Driscoll, D. M. (1994). Social cognition and the study of stereotyping. In E. G. Devine, D. L. Hamilton, T. M. Ostrom (eds), *Social Cognition: Impact on social psychology* (pp. 291-321). New York: Academic.

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

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology*, *53*(1), 575-604.

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *The Behavioral and brain sciences*, *37*(3), 297-307.

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in cognitive sciences*, *13*(12), 524-531.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology a social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405-431.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*,*59*(3), 690-707.

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition Suppression of Ventromedial Prefrontal Activity during Judgments of Self and Others. *Proceedings of the National Academy of Sciences*, *105*(11), 4507-4512.

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social cognition*, *27*(4), 623.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863-3868.

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral Cortex, 12(5)*, 477–485.

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature neuroscience*, *15*(7), 940-948.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387-15392.

Legrain, V., Iannetti, G. D., Plaghki, L., & Mouraux, A. (2011). The pain matrix reloaded: a salience detection system for the body. *Progress in neurobiology*,*93*(1), 111-124.

Martin, K. A. (1994). A brief history of the "feature detector". *Cerebral Cortex*, *4*(1), 1-7.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*(5-6), 655-667.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306-1315.

Morrison, S., Decety, J., & Molenberghs, P. (2012). The neuroscience of group membership. *Neuropsychologia, 50*, 2114–2120.

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1-14.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social cognitive and affective neuroscience*, 4(1), 101-109.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, *10*(9), 424-430.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), Social psychology and the unconscious. the automaticity of higher mental processes (pp. 265–292).

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

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PloS one*, *9*(2), e88534.

Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W. H. C., Cohen, L., ... & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences*, *101*(15), 5658-5663.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549-562.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., ... & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *The Journal of neuroscience*, *27*(9), 2349-2356.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup cooperation and competition: The Robbers Cave experiment*. Norman, OK: University Book Exchange.

Sidanius, J., Pratto, F. (2004). Social Dominance Theory: A New Synthesis. In J. T. Jost, J. Sidanius (eds), *Political Psychology: Key readings* (pp. 31-48). Ann Arbor, MI: Psychology.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of Neuroscience*, *34*(48), 15997-16008.

Stern, C., West, T. V., & Schmitt, P. G. (2014). The liberal illusion of uniqueness. *Psychological Science*, *25*(1), 137-144.

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature neuroscience*, *19*, 795-797.

Tajfel, H., & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences*, *107*(24), 10827-10832.

Tanaka, K. (1993). Neuronal Mechanisms of Object Recognition. *Science*,*262*, 685-68

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*(1), 55-61.

Vogt, B. A. (2005). Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience*, *6*(7), 533-544.

Wendelken, C., Nakhabenko, D., Donohue, S. E., Carter, C. S., Bunge, S. A. (2008). “Brain is to thought as stomach is to ??”: Investigating the role of rostrolateral prefrontal cortex in relational reasoning. *Journal of Cognitive Neuroscience, 20(4)*, 682–693.

Westphal, A. J., Reggente, N., Ito, K. L., & Rissman, J. (2015). Shared and distinct contributions of rostrolateral prefrontal cortex to analogical reasoning and episodic memory retrieval. *Human Brain Mapping, 37,* 896-912.

Wrangham, R. W. (1996). *Chimpanzee cultures*. Cambridge, MA: Harvard University Press.

## Footnotes

[1] Note, however, that assignment to arbitrary groups does not preclude participants from making inferences about the correlation between, for example, being an Eagle and being a Republican.

## Author Note

We would like to thank Jordyn Greenberg, Deborah Viszlay, and Scott Kurdilla for assistance with data collection, John Pyles for assistance with experiment implementation, and Julian Zhou for preliminary data analysis. We thank the Intergroup Neuroscience Lab and Social Perception and Evaluation Lab for thoughtful comments on an earlier version of this manuscript. This work was supported by a Berkman Faculty Development Grant (awarded to MC), the Pershing Square Venture Fund for Research on the Foundations of Human Behavior (awarded to MC), and NSF grant #1349089 (awarded to JVB). We presented these results at the Social Brain Sciences Symposium meeting in Boston, MA (February, 2016).

Table 1. Classification results for 4 classifiers

| Classifier & Region | *x* | *y* | *z* | Cluster size | Accuracy % |
|---|---|---|---|---|---|
| *CROSS-CATEGORY* | | | | | |
| Train on arbitrary, test on political | | | | | |
| dACC/MCC | 10 | 16 | 40 | 92 | 36.28 |
| L AI | -34 | 24 | -2 | 89 | 36.49 |
| Train on political, test on arbitrary | | | | | |
| L RLPFC | -40 | 49 | -1 | 89 | 36.41 |
| *WITHIN-CATEGORY* | | | | | |
| Train on arbitrary, test on arbitrary | | | | | |
| mPFC | -2 | 50 | -1 | 301 | 37.56 |
| L AI/IFG | -46 | 21 | 1 | 151 | 38.16 |
| pgACC | -2 | 34 | 0 | 106 | 37.47 |
| L ITG | -67 | -22 | -15 | 96 | 37.06 |
| Train on political, test on political | | | | | |
| dmPFC | -8 | 37 | 44 | 402 | 37.15 |
| R MFG | 22 | 32 | 46 | 179 | 37.37 |
| L STG | -50 | -44 | 17 | 136 | 37.07 |
| R MFG | 44 | 10 | 44 | 127 | 37.11 |
| R ITG | 57 | -15 | -21 | 97 | 36.84 |

Cluster size reported in voxels ($2mm^3$)
Coordinates are in MNI space; indicate center of cluster
Chance accuracy = 33.3%
dACC/MCC: dorsal anterior cingulate cortex/middle cingulate cortex; AI: anterior insula; RLPFC: rostrolateral prefrontal cortex; mPFC: medial prefrontal cortex; IFG: inferior frontal gyrus; pgACC: pregenual anterior cingulate cortex; ITG: inferior temporal gyrus; dmPFC: dorsomedial prefrontal cortex; MFG: middle frontal gyrus; sTG: superior temporal gyrus.

Figure Legends

Figure 1. (A) Schematic overview of the procedure. Participants completed personality questionnaires (to receive assignment to competitive, arbitrary teams) and demographic information (including political party); ~1-2 weeks later in the lab, participants completed manipulation check measures and the main experiment in the scanner. (B) Main experimental trials examples. Participants identified the corner in which target descriptions appeared.

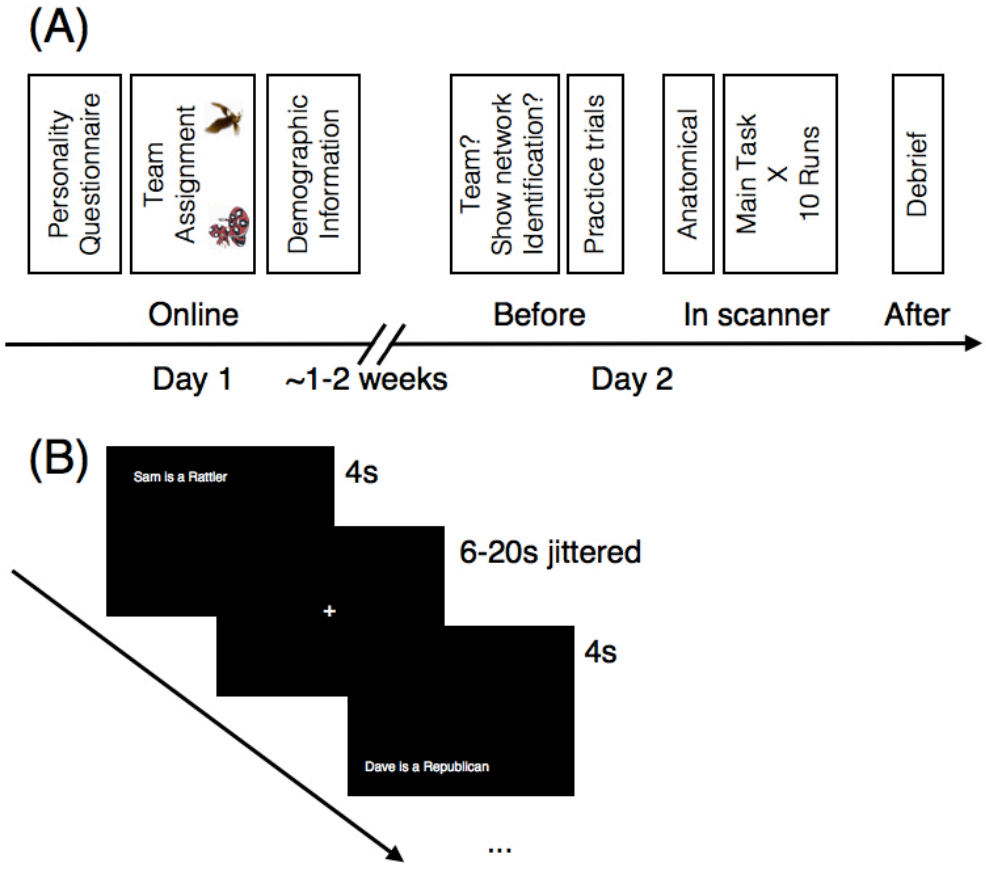

Figure 2. (A) dACC/MCC (x=8) and (B) left AI (z=-3) identified by the classifier that trained on arbitrary teams and tested on political parties. (C) Left RLPFC (z=0) identified by the classifier that trained on political parties and tested on arbitrary teams. (D) Average accuracy scores by cluster. (E) Confusion matrices by cluster. With three groups, chance classification = 33.3%.