Reduced self-referential neural response during intergroup competition predicts competitor harm

Running Head: REDUCED SELF-REFERENTIAL NEURAL RESPONSE IN INTERGROUP COMPETITION

Reduced self-referential neural response during intergroup competition predicts competitor harm

Cikara, M.[1], Jenkins, A. C.[2], Dufour, N.[3], & Saxe, R.[3]

[1]Carnegie Mellon University; [2]University of California, Berkeley; [3]Massachusetts Institute of Technology

Cite as: Cikara, M., A.C. Jenkins, N. Dufour, and R. Saxe. 2014. "Reduced Self-Referential Neural Response During Intergroup Competition Predicts Competitor Harm." NeuroImage 96 (August): 36–43. doi:10.1016/j.neuroimage.2014.03.080.

Abstract

Why do interactions become more hostile when social relations shift from “me versus you” to “us versus them"? One possibility is that acting with a group can reduce spontaneous self-referential processing in the moral domain and, in turn, facilitate competitor harm. We tested this hypothesis in an fMRI experiment in which (i) participants performed a competitive task once alone and once with a group; (ii) spontaneous self-referential processing during competition was indexed unobtrusively by activation in an independently localized region of the medial prefrontal cortex (mPFC) associated with self-reference; and (iii) we assessed participants’ willingness to harm competitors versus teammates. As predicted, participants who showed reduced mPFC activation in response to descriptions of their own moral behaviors while competing in a group were more willing to harm competitors. These results suggest that intergroup competition (above and beyond inter-personal competition) can reduce self-referential processing of moral information, enabling harmful behaviors towards members of a competitive group.

KEYWORDS: self, intergroup competition, medial prefrontal cortex, fMRI, social cognition

Reduced self-referential neural response during intergroup competition predicts competitor harm

1. Introduction

A group of people will often engage in actions that are contrary to the private moral standards of each individual member of that group. Otherwise decent individuals can be swept up into "mobs" that commit looting, vandalism, even physical brutality. In experimental contexts, individuals acting together will act more ruthlessly than when acting alone, for example defecting more often in Prisoner's Dilemma Games and assigning other people to drink more painfully hot hot-sauce (Cohen, Montoya, & Insko, 2006; Meier & Hinsz, 2004). Explicit competition between groups amplifies these tendencies: competition makes group membership more salient (Hamilton, Sherman, & Lickel, 1998; Tajfel, 1982), which strengthens intergroup bias and hostility (Hogg, 1992, 1993; Mullen, Brown, & Smith, 1992).

Increased hostility in competitive intergroup interactions has many psychological explanations. Acting in a group (especially in group competition; Hogg, 1993) creates many of the conditions that facilitate immoral behavior for individuals (Bandura, 1999). Individuals are most likely act contrary to their own moral standards when (i) it is possible to reframe and/or justify the action as serving a greater good, (ii) their sense of personal responsibility is mitigated by anonymity, or diffusion/displacement of responsibility, and (iii) the salience of their own moral standards is low. For example, individuals are more likely to cheat on behalf of another (Gino, Ayal, Ariely, 2013) or in a dark room (Zhong, Bohns, & Gino, 2010), more likely to deliver electric shock to a victim when wearing a hood (Zimbardo, 1995) or when the victim was described as an "animal" (Bandura, 1999), and less likely to help the victim of a crime if there

are other people present (Darley & Latane, 1968). On the other hand, people are less likely to cheat on a test, and even their taxes, if they have first explicitly reflected on their own moral standards (Mazar, Amir, & Ariely, 2008; Shu et al., 2012).

Intergroup interactions seem to provide all three conditions for immoral behavior. First, harming the out-group can often be justified as a rational means to serve one's own group's "greater" good. Intergroup competition increases the salience of the in-group's interests, which allows individuals to reframe harmful behaviors as critical for achieving the in-group's goals (Pinter & Wildschut, 2012). The conflict of interest between groups can mean that harm to the out-group often improves the outcome for the in-group. For example, in sports, failures of the opposing team are necessary for victory of the home team; and in war, the moral requirement to defend one's own nation can create permission or even an obligation to cause harm and suffering to the enemy.

Second, violence is facilitated by acting with a group, even when the violence does not instrumentally serve the in-group, as in mob violence. This may occur partly because acting in a group provides anonymity (Diener, 1979; Festinger, Pepitone, & Newcomb, 1952; Schopler et al., 1995), and allows for displacement/diffusion of responsibility for harmful outcomes (Bandura, 1999; Milgram, 1965; Zimbardo, 1995). For example, in the Milgram (1965) obedience research paradigm, groups of "teacher" participants delivered significantly more severe shocks to "learners" than individual "teacher" participants; they also reported feeling less personal responsibility (Jaffe, Shapir, & Yinon, 1981). The potential for violence increases in highly-structured contexts in which a group is working toward a common goal; individual

agency is diminished, as is one's ability to take responsibility for one's actions (e.g., military or police brutality; Kelman, 1973).

Finally, acting in a group may cause individuals to lose touch with the moral standards that would otherwise guide their behavior. A number of researchers have proposed that acting in a group facilitates a loss of private self-awareness (Deiner, 1979; Duval & Wicklund, 1972; Prentice-Dunn & Rogers, 1989) and increases sensitivity to group identity relative to personal identity (Reicher, Spears, & Postmes, 1995). People may also get swept up in the excitement of acting in a group (Postmes & Spears, 1998), undermining individuals' ability to evaluate on-line whether their behavior coheres with their privately held standards (Diener, 1979).

So far, research examining this third mechanism remains scant due to the difficulty of measuring the accessibility of individuals' moral standards (Postmes & Spears, 1998). Although there is clear evidence that increasing the accessibility of personal moral standards can increase moral behavior in individuals (Mazar et al., 2008; Shu et al., 2012), there is little evidence as to whether acting in a group directly reduces the accessibility of those standards, and whether this facilitates aggression. In part, the absence of evidence is due to the difficulty of quantifying the immediate accessibility of personal standards of morality. For example, a recent study on physical aggression among soccer fans (Van Hiel et al., 2007) measured private self-awareness, though it did so via self-report ("If my team scores a goal I really lose myself completely"). The usefulness of this dependent measure hinges on participants' ability to reflect explicitly, retrospectively, and accurately on their own reduced self-reflection.

This methodological challenge affords an opportunity for cognitive neuroscience (Ellemers, 2012): functional neuroimaging can provide an online, unobtrusive measure of

ongoing psychological processes. In particular, self-referential processing can be measured by activity in a specific and easily localized brain region: medial prefrontal cortex (mPFC). In dozens of studies, a region of mPFC is engaged more when participants reflect on their own (compared to another's) personality traits, mental states, or physical characteristics (e.g., Jenkins & Mitchell, 2011; Kelley et al., 2002; Macrae et al., 2004) or access self-knowledge (Jenkins, Macrae, & Mitchell, 2008; Mitchell, Banaji, & Macrae, 2006). The mPFC response is higher for trait descriptions that are true versus false of the participant (Moran et al., 2006), and higher for self-relevant facts and words (e.g., the participants' own name) versus irrelevant ones (Moran et al. 2009). Response in the mPFC is also correlated with the "self-reference effect" (better memory for words encoded with reference to oneself than others; Symons & Johnson, 1997): that is, greater mPFC response at encoding predicts better subsequent memory for words encoded with respect to oneself (Mitchell, Macrae, & Banaji, 2004). Accordingly, we use spontaneous mPFC activation in response to statements about one's own behaviors as an on-line index of self-referential processing.

1.1 Current investigation

The central hypothesis of the current study is that acting with an in-group can reduce spontaneous self-referential processing in the moral domain and that, when this occurs, it facilitates out-group harm. We scanned participants using fMRI while they took part in a competitive task under two experimental conditions: acting as part of a team and acting alone (see Figure 1). Ostensible distractor stimuli (actually of primary interest) were sentences that described participants' and other individuals' morally-relevant behaviors. In a region of MPFC identified by an independent self-referential processing localizer, we measured spontaneous

activation in response to one's own (versus others') moral behaviors during individual versus group competition. Willingness to harm members of the opposing team was assessed after scanning.

Participants were assigned to one of two competing teams, ostensibly based on personality characteristics. All participants competed in two conditions sequentially (order was counterbalanced across participants). In the "group" condition, participants were told that the nine other members of their group were present, that points accumulated during the task reflected the combined performance of all group members, and that an additional prize for best performance would be split among all ten members of the winning team. These conditions were designed to maximize the experience of acting with, and for, a group: the outcomes of group members were interdependent and visible to all, and in-group success depended on out-group failure. In the "alone" condition, participants were told that team members were not present, that points reflected only the participant's own performance, and that a bonus was available for the top-performing 50% of individual participants.

The questions of primary interest were (i) to what extent performing the task in a group context would reduce spontaneous self-referential moral processing in some participants (indexed by reduced mPFC response to self-relevant moral statements), compared to performing the same task alone, and (ii) whether participants who experienced such a reduction would be more willing to harm members of the competing group. As a measure of willingness to harm, we asked participants to choose, for public distribution, a photograph of each of two in-group and two out-group members. We predicted that participants who exhibited reduced mPFC activation during group competition would choose relatively less

flattering photographs for the out-group (vs. in-group) targets than those who did not exhibit such a reduction. Note that the harm had no bearing on the outcome of the group competition (i.e., harm was not instrumental to advancing the in-group's interests).

Our selection of the mPFC as a region of interest (ROI) does not reflect an assumption that there is a one-to-one mapping between mPFC response and self-referential processing; mPFC is engaged across a wide variety of tasks and cognitive processes. Instead, we focus on this region a priori (i) because it is reliably correlated with self-referential processing, and (ii) to avoid the practice of post-hoc theorizing about activations that are identified in a whole brain contrast. Furthermore, we include a within-experiment manipulation check—a surprise memory task for self- versus other-relevant items—to confirm that activity in our ROI is related to self-referential processing in our participants.

2. Methods

2.1 Participants

Twenty-three volunteers (11 female; M_{age}=23.1) were recruited from the university's participant pool. 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. We excluded 1 participant for excessive movement and another due to technical issues; final N=21. We found no gender differences on any of our outcome variables.

2.2 Stimuli and measures: Team assignment, identification, and pretest

See Figure 1 for overview. Approximately 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 a series of five personality items, ostensibly for the purposes of team assignment; in actuality, each was randomly assigned to either the Eagles or the Rattlers. Third, participants answered 3 questions about their identification with each group on unmarked slider scales ranging from 0 (strongly disagree) to 1.00 (strongly agree): "I [value/like/feel connected to] the [Eagles/Rattlers]."

In the last part of the questionnaire, we assessed the self-relevance of 60 statements involving remote communication ("I have more than 600 Facebook friends," "I never look at friends' Twitter feeds.") and 60 statements involving positive and negative moral behaviors ("I have stolen food from shared refrigerators," "I always apologize after bumping into someone"). We asked, "To what extent is each of the following true of you?" (1 not at all, 4 extremely true); 1s and 2s were coded as false of the participant, 3s and 4s were coded as true.[1] These responses were used to create unique stimulus sets for each participant that included 80 sentences (40 communication/40 moral) in the first person that were true for the participant and 80 complementary sentences in the third person (using both "He" and "She") that were false for the participant. We used the following procedure to generate 160 items from 120 ratings: First, we took all of the sentences in each condition (out of 60) that a participant said was true of her. If there were fewer than 40 such items, we then took items the participant said were *never* true of her, and negated them (e.g. if the participant rejected as false the sentence "I have cheated on an exam", then we included the item "I have *never* cheated on an exam.")

Conversely, if the participant rejected fewer than 40 items, we took an item that the participant did endorse as true, and negated it to create a third person item (e.g. if the participant said it was true that "I have skipped class to do something fun", then we included the statement "*She* has *never* skipped class to do something fun" (emphasis added) in the third-person condition, bringing the total number of stimuli in that condition up to 40 items. Negated items never appeared in the same run as the original item. Half of the statements in each set were randomly assigned to be displayed in the group condition, and half in the alone condition.

2.3 Procedure

We told participants that the study investigated sensitivity to "remote communication" (i.e., information related to communication that does not happen face-to-face: social networks, texting, messaging), and that we were further interested in whether sensitivity to these cues changed in the context of groups. To that end, participants would complete two runs of a remote-communication detection task: once with their group (ostensibly other team members who had already been scanned, and who were returning to the lab to play in real time with the participant), and once alone. In both conditions, the task—go/no-go—was to push a button as quickly as possible when the statement on the screen was related to remote communication and withhold a response otherwise. Note that the difference in motor responses across the go/no-go conditions is not problematic for our analyses because our critical comparison lies within the no-go condition (across alone vs. in a group).

To set-up a competitive structure and incentivize participants, we offered monetary bonuses: in the group condition, the ten members of the best performing team would equally divide an extra $100; in the alone condition, the top 50% of performers, irrespective of team

membership, would receive an additional $10 (keeping individual incentives equivalent across the two conditions). Team scores would be computed as the average speed/accuracy score of all teammates in the group condition. Thus participants would receive the basic participation fee with a possibility of receiving an additional $20.

Prior to entering the scanner, participants were asked to name their team: all participants correctly recalled their team. 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 (increased group cohesion increases intergroup bias; Gaertner & Schopler, 1998).

Participants first underwent an anatomical scan while the experimenters ostensibly set up the participants' teammates at computer stations in a lab across the hall so that they could perform the task in real time with the participant. The run order (group/alone) was counterbalanced between participants. Immediately before the "group" run, participants saw a 3X3 matrix of video feeds showing 9 "teammates" getting ready to compete alongside them. (In actuality, these were pre-recorded videos of 9 people "logging in" to the task; videos of 11 total individuals were pre-recorded, from which 9 were selected to be "teammates" and two were selected to be "competitors" for each participant for purposes of the photo harm measure).

Participants then underwent functional magnetic resonance imaging (fMRI) while doing the go/no-go and localizer tasks. Each run of the main task—an event-related design optimized by optseq for condition order and trial timing—included 40 communication and 40 moral items (though we only referred to these items as "other," not "moral"), respectively. We included a scoreboard, in the lower right hand of the screen: "Eagles [Rattlers] score" in the group

condition and “Your score” in the alone condition. Scores updated after each “go” trial as a function of participants’ response times in both conditions [previous trial’s points + (4s - current trial’s RT * 5)], such that shorter response times yielded a greater increase in points for that trial.

The self-reference localizer task consisted of 4 runs, each of which included 4 blocks (2 self, 2 other) of trait judgments. The self-judgment blocks began with the prompt, “Does the word apply to you?” (2s), followed by a series of 10 traits (3s each); participants responded “yes” or “no” for each trait with a button box. The other-judgment blocks began with the prompt, “Does the word apply to President Obama?” and followed the same structure. Blocks were separated by a 10s fixation period. Two runs followed an ABBA block-pattern, and the other two a BAAB block-pattern; the order of patterns was counterbalanced between participants. Trait words were randomly selected from 4 word-banks (taken from Saxe et al., 2006) such that the same word never appeared twice within participant.

After scanning, participants completed two behavioral tasks privately at a laptop. In order to minimize self-presentation concerns and public self-awareness, participants were also assured that all of their responses were completely confidential. We assessed competitor harm by asking participants to select one photo from a set of 6 for each of 4 targets: a male competitor, female competitor, male teammate, and female teammate. We told participants that we had permission to publish these images in the final publication of the study, as well as to present them at conferences and to put them on a website that was publically accessible. The same 4 individuals were presented to all participants, but the assignment of each face to be the “competitor” or “teammate” was counterbalanced across participants. The 6 images were

stills from the video feeds. In an independent pre-test (N=35), each image was ranked from very unflattering (1) to very flattering (6). For each photo, we averaged flatteringness responses across pre-test participants and then reverse-coded the resulting mean, such that higher values indicated more *un*flattering photos. In the current study, we defined competitor harm as the difference in average "unflatteringness" of the photos selected for supposed dissemination of competitors versus teammates (i.e., out-group minus in-group), such that a higher score reflects more harm to the out-group. Participants in the current study did not see the rankings produced by the pre-test, and images appeared in random order. Next, we administered a surprise recognition memory task for the morally-relevant stimuli participants had viewed in the scanner as part of the go/no-go task. Participants made an "old/new" judgment for 80 moral behavior sentences: 10 first-person from the group condition, 10 third-person/group condition, 10 first-person/alone condition, 10 third-person/alone condition, 20 first-person foils, and 20 third-person foils.

Finally, all participants were thoroughly debriefed: none of the participants had noticed that the first-person statements were true or that the third-person statements were false for them. Though 8 participants expressed some suspicion about the presence of their teammates, we observed no differences between these and the remaining participants on any of the outcome variables.

2.4 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 Tim Trio 3T System (Siemens Solutions, Erlangen, Germany) in the Athinoula A. Martinos Imagining Center at the McGovern Institute for Brain Research at MIT (TR = 2000 msec, TE = 30 msec, field of view = 196 mm, matrix size = 64 × 64). Near whole-brain coverage was achieved with 32 interleaved 3.6 mm near-axial slices.

2.5 fMRI preprocessing and data analysis.

SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) analyzed each participant's MRI data, which were motion corrected and then normalized 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. Normalization warp was produced by SPM combined segmentation and normalization and then applied to the anatomical image and the coregistered functionals. Analyses high-pass filtered and smoothed the data using a Gaussian filter (full width half maximum= 5mm).

Functional images were analyzed using both whole brain random effects analyses, and using group-level regions of interest. For whole brain analyses, we first built a modified general linear model of the experimental design, and used this model to analyze the BOLD response in each voxel. Both the main experiment and localizer models included covariates of interest (the experimental conditions) as well as nuisance covariates (i.e., a mean term). The main experiment used an event related design; each event consisted of the 2 TRs during which each stimulus was presented on the screen. The group and alone runs were modeled separately. The localizer used a block design; each block consisted of 15 TRs during which participants judged whether a series of adjectives applied to them (or President Obama). We modeled the conditions as a boxcar (matching the onset and duration of each event) convolved with a

standard hemodynamic response function (HRF). To identify voxels in which effects of condition were reliable across participants, BOLD signal differences between conditions (linear combinations of the beta parameters for condition covariates) were submitted to second level, random-effects analysis. All whole brain analyses used corrected p thresholds, at $p < 0.05$, based on Monte Carlo simulations of the false positive rate in these data (Nichols & Holmes, 2004; http://go.warwick.ac.uk/tenichols/snpm).

To define regions of interest (i.e., mPFC), we conducted mixed effects analyses on the localizer experiment, using a threshold of $p < 0.005$ (voxel-wise), and a cluster threshold of $k > 368$, $p < .05$, corrected by SnPM. Coordinates of the peak voxel in the group ROI were identified, and all contiguous suprathreshold voxels within the cluster defined the region of interest (ROI). The response at each time point for each condition in the main experiment was calculated as the average BOLD response across all voxels in each ROI, for each participant. For the purposes of statistical analyses, we averaged 6–10 s after stimulus onset. This time accounted for hemodynamic lag. The data extracted from the ROIs were not filtered, other than averaging. All peak voxels are reported in MNI coordinates.

Group analyses treated the variability between participants as a random effect. We used a self > other contrast in the localizer data to identify a group-average mPFC ROI. We used a first-person/moral > third-person/moral contrast in the group and alone runs, respectively, in the mPFC ROI to test the prediction that the first > third-person difference in mPFC would be greater in the alone than the group run.

3. Results

Reduced self-referential processing and harm. The key prediction of this study was that, in a region of mPFC associated with self-referential processing, reduced spontaneous activation to self-relevant moral statements during group competition would predict willingness to harm an out-group member. We used the choice of relatively unflattering photos of out-group members (out-group minus in-group; higher score = more harm toward out-group) as a measure of harm. MPFC response was measured in a group ROI, based on the explicit self>other contrast in an independent localizer experiment (peak:-4,34,-4, k=377; see Table 1, Figure 2). No other supra-threshold clusters emerged for the self>other localizer contrast.

Overall, participants exhibited a tendency to harm competitors (*M*=3.17, *SD*=1.38) more than teammates (*M*=2.49, *SD*=.98; *t*(20)=1.88, *p*=.04, one-tailed). Importantly, when participants competed in a group, we observed a significant negative correlation between participants' mPFC response and their willingness to harm competitors, *r*(19)=-.44, *p*=.05. Specifically, individuals with reduced mPFC response to first-person (versus third-person) moral statements during intergroup competition selected less flattering photos of (i.e., inflicted more harm on) competitors relative to teammates. Critically, this relationship was specific to competing in a group: when participants competed alone, the mPFC response to self>other was unrelated to competitor harm, *r*(19)=.09, *p*=.70. These correlations are significantly different from one another, z=1.68, *p*=.05, one-tailed. The relationship between reduced mPFC activation and competitor harm was also specific to thinking about one's own *moral* behaviors; mPFC response to sentences describing self-relevant remote communication behaviors was not correlated with competitor harm in either condition: group, *r*(19)=-.07, *p*=.76; alone, *r*(19)=.06,

p=.80. Thus, willingness to harm an out-group member was specifically associated with reduced mPFC response to self-relevant moral items while competing in a group.

The above analyses show that some individuals exhibited reduced mPFC response to self-relevant moral items during intergroup competition, and that those individuals engaged in more harm towards competitors. Overall, however, we observed no difference in the mPFC response to self>other moral sentences when playing alone versus in a group: $F_{self/other}(1,20)= 0.88$, $p=.36$; $F_{alone/group}(1,20)= 0.003$, $p=.95$; $F_{interaction}(1,20)= 0.34$, $p=.56$, although the mean differences were in the expected direction (i.e., smaller difference in the mPFC response to self versus other in the group as compared to the alone condition): $M_{self/alone}=-.17$; $M_{other/alone}=-.20$; $M_{self/group}=-.19$; $M_{other/group}=-.19$. A whole brain analysis also failed to find any regions with significantly different responses to first-person versus third-person moral sentences in the alone relative to group condition (whole brain results are located in Table 2). These results are consistent with the complexity of intergroup behavior, demonstrating that for some but not all individuals, competing in a group is associated with reduced mPFC response to self-relevant moral stimuli, and this reduction is associated with a greater propensity to harm competitors.

Behavioral self-reference effect. To explore the extent to which reduced mPFC activation in the group competition context genuinely indexed reduced processing of self-relevant information, we examined the relationship between the neural self reference-effect (i.e., mPFC activation to first-person versus third-person statements) and the behavioral self-reference effect (i.e., recognition memory for first-person versus third-person statement). Typically, individuals show better subsequent memory for items encoded with regard to the self, compared to other encoding conditions; this memory advantage is associated with mPFC

activation. Consistent with past research, we observed a positive correlation across individuals between the self-reference effect in memory and the self-reference effect in mPFC in the "alone" condition, $r(19)=.47$, $p=.03$. However, consistent with the hypothesis that competing in a group can reduce self-referential processing, we observed a marginally significant *negative* correlation between mPFC activity and subsequent memory in the group condition, $r(19)=-.39$, $p=.08$; these correlations are significantly different from one another, $z = 2.78$, $p = .005$. Thus, although activation in mPFC was indeed associated with subsequent memory for self-relevant items when competing alone, competing in a group inverted this relationship. Participants' d' scores were as follows: for items seen in the group condition, self d' = 2.42, other d' = 2.62; for items seen in the alone condition, self d' = 2.53, other d' = 2.38. We found no overall memory advantage for items encountered in the first-person versus third-person ($F_{self/other}(1,20)= 0.02$, $p=.90$), or while playing in a group versus alone ($F_{alone/group}(1,20)= 0.36$, $p=.57$), and no interaction ($F_{interaction}(1,20)= 1.19$, $p=.29$).

Alternative explanations. Analyses of participants' responses and RTs across the conditions revealed that participants were not faster or more accurate when competing alongside their teams relative to competing alone: out of 40 items in each condition, M_{group} misses = 1.76; M_{alone} misses = 2.19; $t(20) = -0.70$, $p = .49$; M_{group} RT = 1.19; M_{alone} RT = 1.13, $t(20) = 0.99$, $p = .33$. These results suggest that alternative explanations such as differential effort across conditions cannot account for our findings. Another possible explanation for why some participants might be more prone to exhibit reduced self > other mPFC in in the group and/or to harm members of the opposing team was the degree of identification with one's own team. Although participants valued, liked, and felt connected to their own team more than the

competing team (all *ts*(20)>3.1, *ps*>.01), individual differences in team identification did not predict mPFC activity or competitor harm (all *r*s<.3, *ns*).

4. Discussion

The current study examined whether acting as a member of a competitive group resulted in some participants losing touch with their moral selves, and in turn, whether those participants were more likely to harm competitors. Indeed, participants who exhibited reduced mPFC response to self-relevant moral items while performing a competitive task in a group were more willing to select unflattering photographs of competitors. Importantly, this relationship was specific to *moral* items; there was no relationship between competitor harm and mPFC response to communication items. When participants competed alone, we replicated previous findings showing that mPFC activation correlates positively with better memory for items encoded with reference to oneself vs. another person, confirming that activation in our mPFC ROI indexed self-referential processing in the context of our experiment. This relationship between self-related mPFC activation and self-relevant memory disappeared, however, when participants were performing the same task alongside their teammates.

These findings add to a growing literature on the cognitive and neural processes that facilitate harm in competitive contexts. Previous studies have focused on the role of empathic failures and Schadenfreude (Cikara, Bruneau, & Saxe, 2011). In interpersonal competition, reward-related brain regions (i.e., ventral striatum) respond when a competitor receives a painful electric shock (Singer et al., 2006), or when an envied target experiences misfortunes (Takahashi et al., 2009). Parallel effects occur at the group level: Red Sox and Yankees baseball

fans report feeling pleasure and show activity in the same reward-related brain regions when they watch the rival team fail to score; participants exhibiting greater reward-related activity also report being more likely to harm the rival team's fans (Cikara et al., 2011). Similarly, soccer fans exhibit reward-related activity when watching a rival team's fan receive a painful electric shock; again, participants who exhibit greater reward-related activity are less willing to relieve the rival's pain by receiving an electric shock themselves (Hein et al., 2010). Although research has begun to explore the neural substrates of real-life intergroup processing outside of a competitive context (Morrison, Decety, & Molenberghs, 2012; see Cikara & Van Bavel, 2014 for a review), no studies of which we are aware have examined the neural substrates of overt intergroup competition, in which participants actively compete. The current study adds to this literature by exploring reductions in spontaneous self-referential processing in the moral domain, perhaps due to reduced accessibility to one's own moral standards in competitive intergroup contexts: a complementary mechanism leading to harmful behavior against a competitive out-group member. The phenomenon observed here likely combines with these other factors to produce an overall shift towards greater harm in intergroup competition.

Although many psychological and neural mechanisms promote interpersonal and intergroup harm in competitive contexts, some of the same mechanisms may also contribute to prosocial behavior and cooperation in the absence of competition. For example, the ventral striatum responds when individuals observe cooperation (Rilling et al., 2002) and fair resource distribution (Tricomi et al., 2010), as well as when individuals choose to act equitably (Zaki & Mitchell, 2012). Reduced self-referential processing may itself facilitate prosocial behavior, at least towards in-group members, in some contexts: group-oriented participants can be swayed

to donate more money than individual-oriented participants (Spivey & Prentice-Dunn, 1990). That the same mechanisms may promote both pro-social and anti-social behavior highlights the importance of investigating these phenomena in the context of different functional relational structures (e.g., cooperative, competitive, independent).

One limitation of the present study is that we did not assess whether participants deliberately intended to do harm when selecting relatively unflattering photographs of competitors. It is possible that their selections reflect a subconscious bias (e.g., a top-down distortion of out-group faces; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). Future studies will have to clarify which of these two, or if both of these motivations are related to reduced self-referential neural responses.

Although humans exhibit strong preferences for equity and moral prohibitions against harm in many contexts, people's priorities change when there is an "us" and a "them." Groups dynamically shape our perceptions, emotions, motivations, and behaviors. The current research furthers our understanding of the psychological processes within individuals that facilitate intergroup hostility in competitive contexts. Of course, it will be the task of future research to understand why certain individuals are more prone than others to "lose themselves" in intergroup competition. Nonetheless, these findings suggest a possible intervention—increasing self-referential processing—as one means of attenuating intergroup conflict. Understanding how individuals' self-representations can change in intergroup contexts will be necessary to complete the picture of the cognitive and neural mechanisms that support intergroup understanding and interaction.

Footnotes

[1]A separate sample (N=52) rated the communication (M=6.27, SD=.97) and moral items (M=6.26, SD=2.15) as equivalent in valence, $t(59)=0.14$, $p = 89$. Communication and moral items were also matched on characters per sentence ($M_{comm}=46.3$, $M_{moral}= 47.9$; $t(59)=0.88$, p=.38), words per sentence ($M_{comm}=8.5$, $M_{moral}= 8.9$; $t(59)=0.93$, $p=.36$), and Flesch Reading Ease ($M_{comm}=75.4$, $M_{mora}=71.5$; $t(59)=1.05$, $p = .30$).

Acknowledgments

The authors thank the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, including Dr. Christina Triantafyllou, Steven Shannon and Sheeba Arnold Anteraper. We also thank Hilary Richardson for assistance with scan acquisition and data analyses. The authors gratefully acknowledge support of this project by a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (NRSA grant # 1F32HD068086-01A1 awarded to MC), a grant from the Air Force Office of Scientific Research, managed through the Office of Naval Research (grant #N000140910845 awarded to RS), and a grant from the Packard Foundation (awarded to RS).

References

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*, 193–209.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science, 22,* 306-313.

Cikara, M., Bruneau, E. G., & Saxe, R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, *20*, 149-153.

Cikara, M., & Van Bavel, J. J. (in press). The neuroscience of intergroup relations: An integrative review. Perspectives on Psychological Science.

Cohen, T.R., Montoya, R.M., & Insko, C.A. (2006). Group morality and intergroup relations: Cross-cultural and experimental evidence. *Personality and Social Psychology Bulletin*, *32*, 1559–1572.

Darley, J. M. & Latané, B. (1968). "Bystander intervention in emergencies: Diffusion of responsibility".*Journal of Personality and Social Psychology* 8: 377–383

Diener, E. (1979). Deindividuation, self-awareness and disinhibition. *Journal of Personality and Social Psychology, 37*, 116-71.

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness.* New York: Academic Press.

Ellemers, N. (2012). The Group Self. *Science, 336*, 848-852.

Festinger, L., Pepitone, A., & Newcomb, T. (1952). Some consequences of deindividuation in a group. *Journal of Abnormal and Social Psychology, 47,* 382-389.

Gaertner, L., & Schopler, J. (1998). Perceived in-group entitativity and intergroup bias: An

interconnection of self and others. *European Journal of Social Psychology*, *28*, 963-980.

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior and Organization.* Special Issue on "Deception, Incentives and Behavior." Forthcoming.

Hamilton, D. L., Sherman, S. J., & Lickel, B. (1998). Perceiving Social Groups: The Importance of Entitativity. *Intergroup cognition and intergroup behavior*.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to in-group and out-group members' suffering predict individual differences in costly helping. *Neuron, 68*, 149-160.

Hogg, M. A. (1992). *The social psychology of group cohesiveness: From attraction to social identity.* London: Harvester Wheatsheaf.

Hogg, M. A. (1993). Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology, 4*, 85-111.

Jaffe, Y., Shapir, N., & Yinon, Y. (1981). Aggression and its escalation. *Journal of Cross-Cultural Psychology, 12,* 21– 36.

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.

Jenkins, A. C. & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience, 6(3),* 211-218.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*, 785–794.

Kelman, H. C. (1973). Violence Without Moral Restraint: Reflections on the Dehumanization of Victims and Victimizers. *Journal of Social Issues, 29*, 25-61.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex*, *14*(6), 647–654.

Mazar, N., Amir, O. & Ariely, D. (2008). The dishonesty of honest people. *Journal of Marketing Research,* 45, 633-644.

Meier, B. P., & Hinsz, V. B. (2004). A comparison of human aggression committed by groups and individuals: An interindividual-intergroup discontinuity. *Journal of Experimental Social Psychology*, *40*, 551–559.

Milgram, S. (1965). "Some Conditions of Obedience and Disobedience to Authority". *Human Relations,* 18, 57–76.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience, 24*, 4912-4917.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655-663.

Moran, J.M., Wyland, C.L., Macrae, C.N., Heatherton, T.F., and Kelley, W.M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience,* 18, 1586-1594.

Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-reference evaluation. *Social Neuroscience*, *4*(3), 197–211.

Morrison, S., Decety, J., & Molenberghs, P. (2012). The neuroscience of group membership. *Neuropsychologia*, *50*(8), 2114-2120.

Mullen, B., Brown, R., & Smith, C. (1992). In-group bias as a function of salience, relevance, and status: an integration*. European Journal of Social Psychology*, *22*, 103–122.

Nichols, T. & Holmes, A. (2004). Nonparametric permutation tests for functional neuroimaging. *Human Brain Function, 2*, 887–910.

Pinter, B., & Wildschut, T. (2012). Self-interest masquerading as ingroup beneficence: An altruistic rationalization explanation of the interindividual– intergroup discontinuity effect. *Small Group Research, 43*, 105–123.

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin, 123*, 238-259.

Prentice-Dunn, S., & Rogers, R. W. (1989). Deindividuation and the self-regulation of behavior. In P. B. Paulus (Ed.), *The psychology of group influence* (2nd ed., pp. 86-109). Hillsdale, NJ: Erlbaum.

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 6, pp. 161-198). Chichester, England: Wiley.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kitts, C.D., 2002. A neural basis for social cooperation. *Neuron 35*, 395 – 405.

Saxe, R., Moran, J.M., Scholz, J.K., and Gabrieli, J.D.E. (2006). Overlapping and non-overlapping brain regions for theory of mind and self-reflection in individual subjects. *Social Cognitive and Affective Neuroscience,* 1, 229-234.

Schopler, J., Insko, C. A., Drigotas, S. M., Wieselquist, J., Pemberton, M. B., & Cox, C. (1995). The role of identifiability in the reduction of interindividual–intergroup discontinuity. Journal of Experimental Social Psychology, 31(6), 553–574. http:// dx.doi.org/10.1006/jesp. 1995.1025.

Shu, L., Mazar, N., Gino, F., Ariely, D., Bazerman, M. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109(38)*: 15197-15200.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*, 466–469.

Spivey, C. B., & Prentice-Dunn, S. (1990). Assessing the directionality of deindividuated behavior: Effects of deindividuation, modeling, and private self-consciousness on aggressive and prosocial responses. *Basic and Applied Social Psychology, 11*, 387–403.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. Psychological Bulletin, 121, 371–394.

Tajfel, H. (1982) *Social identity and intergroup relations*, Cambridge, England: Cambridge University Press

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: Neural correlates of envy and Schadenfreude. *Science, 323,* 937-939.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010) Neural evidence for inequality-averse social preferences. *Nature, 463*, 1089–1091.

Van Hiel, A., Hautman, L., Cornelis, I., De Clercq, B. (2007). Football hooliganism: comparing self-awareness and social identity theory explanations. *Journal of Community & Applied Social Psychology 17*(3): 169-186

Zaki, J., & Mitchel, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), 19761–19766.

Zhong, C., Bohns, V. K., & Gino, F. (2010). A good lamp is the best police: Darkness increases self-interested behavior and dishonesty. *Psychological Science*, 21(3), 311-314.

Zimbardo, P. G. (1995). The psychology of evil: A situationist perspective on recruiting good people to engage in anti-social acts. *Research in Social Psychology, 11*, 125-133.

Table 1. *Whole-brain Analyses in Localizer Experiment*

Regions	*x*	*y*	*Z*	Pseudo t statistic	Cluster Size (Voxels)
Self > Other					
Medial prefrontal cortex/pregenual ACC	-4	34	-4	4.81	377
	4	40	-4	4.56	
	-2	36	4	4.15	
Other > Self					
Posterior cingulate/precuneus	6	-52	22	7.45	1144
	4	-54	26	7.12	
	0	-58	30	6.36	
R Temporal pole	60	-4	-18	7.1	573
	60	0	-24	6.2	
L Temporo-parietal junction	-42	-54	20	6.16	823
	-42	-64	26	5.68	
	-52	-66	28	5.29	
L Temporal pole	-54	-6	-18	6.1	548

Note. Peak voxel and cluster size (1 voxel = $3mm^3$); additional local maxima within each cluster are included beneath the peak description. Cluster-wise significance threshold, $p < .05$, corrected. Coordinates refer to the Montreal Neurological Institute stereotaxic space.

Table 2. *Whole-brain Analyses Collapsing Across Group/Alone Conditions in Main Experiment*

Regions	*x*	*y*	*Z*	Pseudo t statistic	Cluster Size (Voxels)
Communication > Moral					
(collapsing across self/other)					
Cerebellum	16	-52	-20	9.43	5342
	4	-62	-12	8.33	
	14	10	4	7.15	
L insula	-36	16	-2		1770
				9.28	
	-30	16	6	8.95	
	-48	6	28	7.66	
L Inferior parietal cortex	-46	-38	44		2767
				9.17	
	-26	-64	42	8.27	
	-28	-66	40	8.24	
L Inferior temporal cortex	-56	-54	-14	9.17	1655
	-42	-76	-10	6.03	
	-42	-80	-18	5.84	
R Inferior parietal cortex	36	-46	40	8.29	1905
	34	-60	42	8.17	
	32	-64	38	7.12	
R dlPFC	42	46	-2	7.38	819
	42	44	14	7.09	
	40	42	16	6.74	
R Insula	34	28	2	6.35	2445
	52	12	40	6.13	
	54	12	24	6.65	
Dorsal ACC/SMA	6	18	46	6.63	983
	4	32	42	6.6	
	-4	8	50	6.43	
L vlPFC	-36	52	-14	4.98	1057
	-26	46	-18	3.18	
	-44	42	4	6.09	
Moral > Communication					
(collapsing across self/other)					
R Superior parietal cortex, Precentral/postcentral gyrus, Superior temporal gyrus	42	-16	18	12.37	30156
	-14	-58	12	12.02	
	48	-12	-12	9.96	
mPFC	-8	48	10	6.61	3056
	32	32	-16	6.39	
	4	48	-12	6.31	

Self > Other (collapsing across moral/communication)					
Left dlPFC	-40	52	10	5.88	591
	-32	48	20	3.95	
	-44	48	-4	3.87	
L Inferior parietal cortex, L TPJ	-42	-56	52	5.49	757
	-38	-58	50	5.46	
	-36	-60	52	5.39	
L Middle temporal gyrus	-62	-46	-10	5.46	456
	-56	-34	-8	5	
	-56	-24	-18	4.32	
L Middle frontal gyrus	-48	12	42	4.59	795
	-42	10	36	4.51	
	-36	6	34	4.38	
Other > Self (collapsing across moral/communication)					
R Thalamus/Periaqueductal gray	10	-28	-8	5.94	628
	18	-38	8	5.53	
	10	-30	18	5.13	
R Posterior insula	42	-12	20	4.6	470
	50	0	26	5.32	
	50	-6	18	4.97	
R Amygdala	22	-4	-12	4.8	409
	36	-4	-34	5.26	
	-2	-4	-20	4.43	
L Hippocampus	-28	-44	16	4.4	485
	-32	-42	0	4.77	
	-30	-28	-14	4.72	

Note. Peak voxel and cluster size (1 voxel = $3mm^3$); additional local maxima within each cluster are included beneath the peak description. Cluster-wise significance threshold, $p < .05$, corrected. Coordinates refer to the Montreal Neurological Institute stereotaxic space. Accompanying images appear in supplementary materials.

Figure Legends

Fig. 1. A) A schematic overview of the procedure. B) Examples of stimuli from the main experiment. C) Example of the picture-rating task for one target (independent pre-test rankings appear beneath each photo: 1 = most flattering to 6 = least flattering; participants in the current study did not see the rankings).

Fig. 2. *Left panel*: Group mPFC ROI, based on self>other contrast in the localizer experiment (peak: -4, 34, -4, k = 377). *Top right panel*: correlation between spontaneous mPFC response to self>other moral behavior and willingness to harm competitors when competing alone, $r(19) = .09$, p = ns. *Bottom panel*: significant negative correlation between spontaneous mPFC response to self>other moral behavior and willingness to harm competitors when competing with the group, $r(19) = -.44$, $p < .05$.

Figure 1.

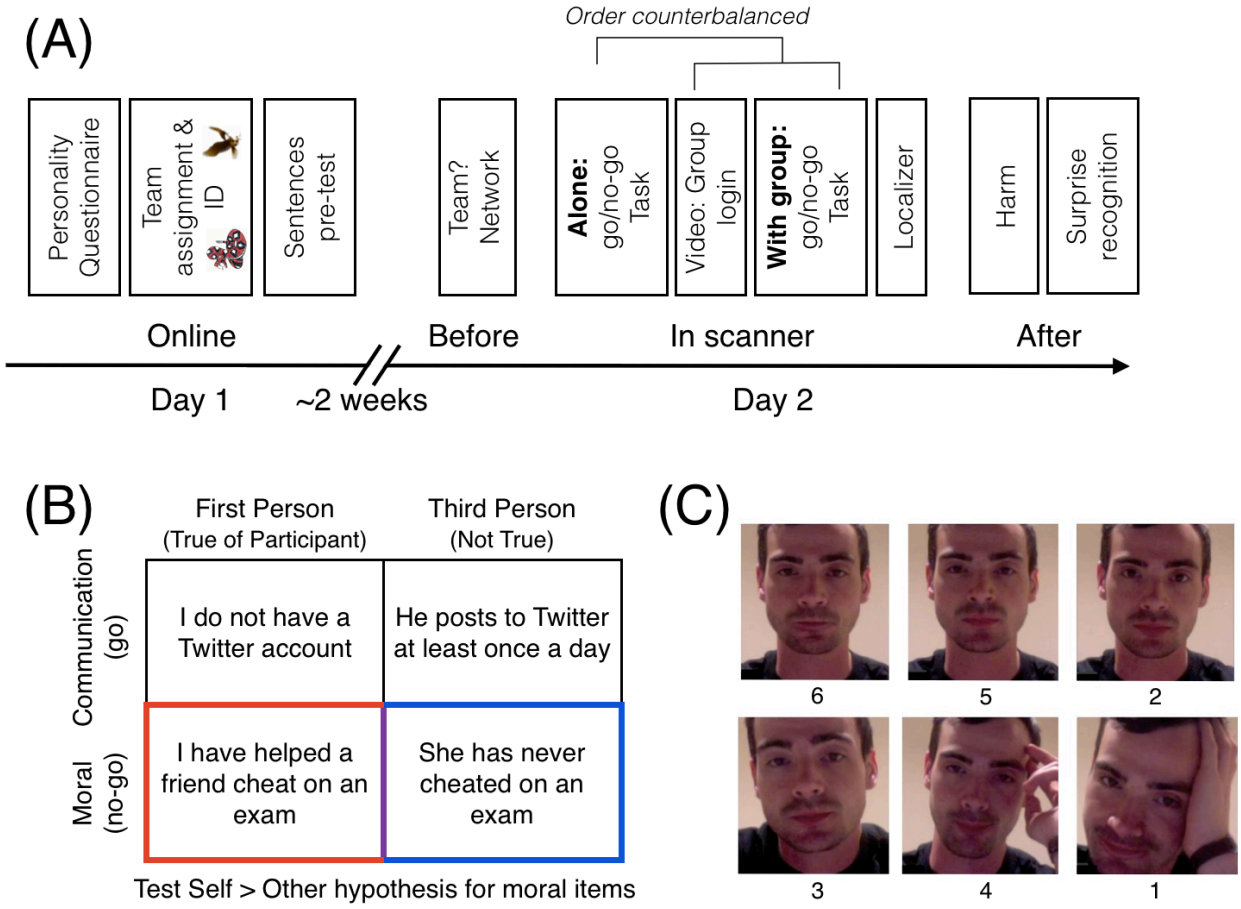
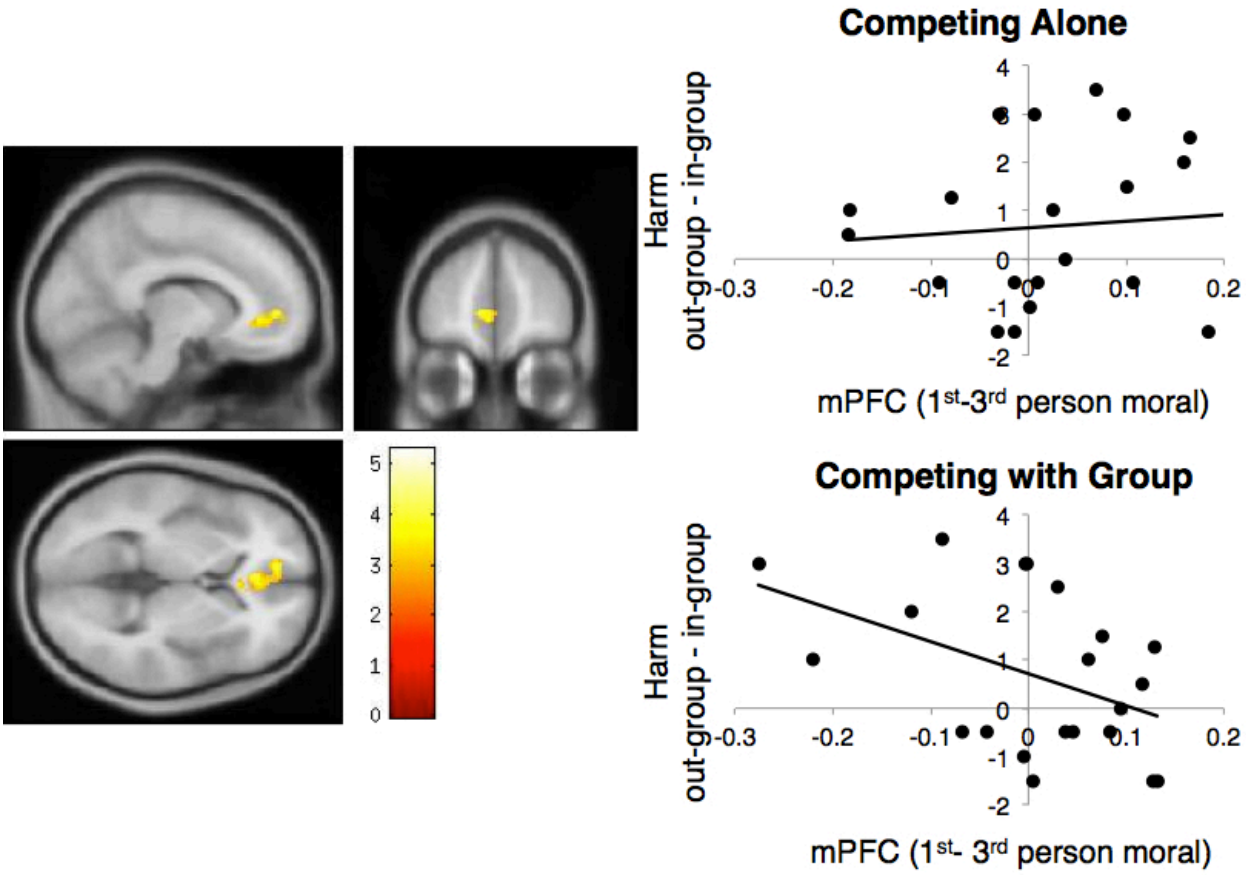

(A)
Order counterbalanced
Personality Questionnaire
Team assignment & ID
Sentences pre-test
Team? Network
Alone: go/no-go Task
Video: Group login
With group: go/no-go Task
Localizer
Harm
Surprise recognition
Online
Before
In scanner
After
Day 1
~2 weeks
Day 2
(B)
First Person (True of Participant)
Third Person (Not True)
Communication (go)
Moral (no-go)
I do not have a Twitter account
He posts to Twitter at least once a day
I have helped a friend cheat on an exam
She has never cheated on an exam
Test Self > Other hypothesis for moral items
(C)
6
5
2
3
4
1

Figure 2


Competing Alone
Harm
out-group - in-group
mPFC (1st-3rd person moral)
Competing with Group
Harm
out-group - in-group
mPFC (1st- 3rd person moral)
5
4
3
2
1
0