Social Cognitive and Affective Neuroscience, 2020, 371–381

doi: 10.1093/scan/nsaa055
Advance Access Publication Date: 25 April 2020
Original manuscript

Shifting prosocial intuitions: neurocognitive evidence for a value-based account of group-based cooperation

Leor M. Hackel,[1] Julian A. Wills,[2] and Jay J. Van Bavel[2,]

[1]Department of Psychology, University of Southern California, Los Angeles, CA 90007, USA, and [2]Department of Psychology & Center for Neural Science, New York University, New York, NY 10003, USA

Correspondence should be addressed to Leor M. Hackel, Assistant Professor of Psychology, University of Southern California, 3620 South McClintock Ave, Los Angeles, CA 90089, USA. E-mail: lhackel@usc.edu

Abstract

Cooperation is necessary for solving numerous social issues, including climate change, effective governance and economic stability. Value-based decision models contend that prosocial tendencies and social context shape people's preferences for cooperative or selfish behavior. Using functional neuroimaging and computational modeling, we tested these predictions by comparing activity in brain regions previously linked to valuation and executive function during decision-making—the ventromedial prefrontal cortex (vmPFC) and dorsolateral prefrontal cortex (dlPFC), respectively. Participants played Public Goods Games with students from fictitious universities, where social norms were selfish or cooperative. Prosocial participants showed greater vmPFC activity when cooperating and dlPFC-vmPFC connectivity when acting selfishly, whereas selfish participants displayed the opposite pattern. Norm-sensitive participants showed greater dlPFC-vmPFC connectivity when defying group norms. Modeling expectations of cooperation was associated with activity near the right temporoparietal junction. Consistent with value-based models, this suggests that prosocial tendencies and contextual norms flexibly determine whether people prefer cooperation or defection.

Key words: cooperation; fMRI; preferences; norms; prosocial; social neuroscience

Introduction

When individuals prioritize themselves over their communities, the consequences can damage global economies, scientific institutions and the planet. Philosophers and scientists have debated the origins of human prosociality for centuries (Curry, 2016). The study of cooperation has evolved beyond philosophy and consumed the energy of scientists across numerous disciplines, from primatology to economics (Brosnan, 2018; Declerck and Boone, 2018). To better understand the nature of cooperation, we took an interdisciplinary approach that combined neuroeconomics, social and personality psychology, and cognitive neuroscience. We examined the neural systems that guide cooperation in groups, and how these systems are shaped by social norms and individual differences.

For many years, *prosocial restraint* models asserted that cooperation stems from deliberate restraint of selfish impulses (Stevens and Hauser, 2004; DeWall *et al.*, 2008; Kocher *et al.*, 2017). More recently, *prosocial intuition* models have argued that cooperation stems from intuition, whereas deliberation always maximizes self-interest (Rand *et al.*, 2012; Rand, 2016). Both models carve the mind into two core processes: intuition (i.e. rapid, automatic, reflexive mental processing) and deliberation (i.e. delayed, controlled, reflective processing). They differ on the role these processes play in promoting selfish *vs* collective interest. Despite extensive research on this issue, studies of the

Received: 14 November 2019; **Revised:** 6 February 2020; **Accepted:** 17 April 2020

mental processes underlying cooperation have yielded mixed evidence for existing theoretical models (Bouwmeester *et al.*, 2017; Rand, 2017). To help reconcile these issues, we considered an alternative approach to cooperation.

Value-based decision models argue that the cooperative decisions hinge on preferences that vary between individuals and situations. According to this approach, the neural systems involved in storing, representing and learning value play a role in all decisions, whether they are motor actions, food preferences or cooperative behavior (Chib *et al.*, 2009; Levy and Glimcher, 2012; Krajbich *et al.*, 2015b). By specifying the conditions under which cooperation requires effortful deliberation, value-based models explain when (and for whom) interventions to boost cooperation will be most effective. Furthermore, these models can reconcile opposing predictions from prosocial restraint and prosocial intuition models. For instance, consider two car drivers (one prosocial and the other selfish) who witness a collision and decide to pull over and help. Although the driver with prosocial preferences might find the decision to pull over easy, the selfish driver might find the decision hard. Understanding the value each individual places on human welfare may determine which processes produce cooperation.

Research on human cooperation often involves social dilemmas, such as the public goods game (PGG). In this economic game, players can make monetary contributions to their group that get multiplied and distributed equally (including to players who keep all their money for themselves). It is in the group's collective interest if all players contribute, but it is in each individual's self-interest to contribute nothing and reap the benefits of other's generosity (i.e. free-riding). Consistent with the prosocial intuition model, faster decisions have been associated with larger group contributions in a PGG, suggesting intuitive cooperation (Rand *et al.*, 2012). However, value-based models posit that longer decisions reflect decision-conflict during choices in which the agent is close to being indifferent between options (Evans *et al.*, 2015). In addition, prosocial individuals (i.e. those who generally help others) are faster to cooperate than free-ride (i.e. prioritize themselves over the group), whereas selfish individuals are faster to free-ride than cooperate (Hutcherson *et al.*, 2015; Krajbich *et al.*, 2015a). Thus, individual differences in prosociality shape which mental computations steer cooperation (Van Lange *et al.*, 1992, 2013).

The value-based approach also contends that situational factors shift these mental computations. Increasing the cost of cooperation renders selfish decisions faster and less conflicted (Krajbich *et al.*, 2015a). Therefore, descriptive norms (i.e. perceptions of typical social behavior) may also shape cooperation. Humans have strong needs for belonging and tendencies toward conformity (Asch, 1951; Cialdini *et al.*, 1990; Baumeister and Leary, 1995), and norms can increase prosocial behavior (Nook *et al.*, 2016) as well as the intrinsic value of socially preferred stimuli (Zaki *et al.*, 2011). Indeed, one's cooperation is related to the mean levels of cooperation of others around (Smith *et al.*, 2018). As a result, it may take more effort to defy group norms—cooperating when others are selfish or free-riding when others are generous. In sum, whether people are faster to cooperate or free-ride depends on their prosocial tendencies and context.

We examined whether neural regions involved in decision-making show a similar pattern. While early evidence from neuroimaging studies implicated the ventromedial prefrontal cortex (vmPFC) in affective evaluations (Bechara *et al.*, 1997; Satpute and Lieberman, 2006; Lebreton *et al.*, 2009), the vmPFC is now widely considered a hub for value-based decision-making. For instance, activity in vmPFC is proportional to the expected value that would be obtained by taking an action (Kable and Glimcher, 2007; Levy and Glimcher, 2012). In contrast, the dorsolateral prefrontal cortex (dlPFC) is associated with executive function and supports goal pursuit (Carter and Van Veen, 2007). Indeed, dlPFC damage impairs executive functions like working memory, reasoning and self-regulation (Zhu *et al.*, 2014). Functional connectivity between dlPFC and vmPFC is thought to reflect modulation of value during goal-directed decisions (Hare *et al.*, 2009). Moreover, the connectivity of these regions hinges on individual preferences (e.g. healthy eating among dieters; Hare *et al.*, 2009) and context (e.g. regulating cravings; Hutcherson *et al.*, 2012). In sum, vmPFC activation reflects expected value whereas dlPFC may index overcoming prepotent response tendencies (Barber and Carter, 2004), decision difficulty (Saraiva and Marshall, 2015) and goal-directed modulation of vmPFC's expected value signal (Rangel and Hare, 2010).

This suggests that individuals' prosociality and the social contexts in which they are embedded should determine the value of cooperation. According to the value-based approach, we should observe greater vmPFC activity during choices that align with one's prosocial tendency since value computations will be higher. Likewise, we should observe dlPFC-vmPFC connectivity during choices that conflict with this tendency, since the dlPFC will need to modulate value signals in the vmPFC to steer decision-making. In this manner, fMRI can provide insight into group-based cooperation.

Indeed, past research has found roles for vmPFC and dlPFC in cooperation consistent with a value-based approach (Pärnamets *et al.*, in press). For instance, people show greater activation in vmPFC when mutually cooperating in a Prisoner's Dilemma (Rilling *et al.*, 2002, 2007) or when inferring cooperative intentions in others (Cooper *et al.*, 2010). However, people show enhanced dlPFC activity when cooperating in a Prisoner's Dilemma with someone who typically defects (Suzuki *et al.*, 2011) or when trusting out-group members in a Trust Game (Hughes *et al.*, 2017)—situations that might require modulating prepotent value representations. Moreover, a recent study found that patients with dlPFC damage were less likely to cooperate in a Public Goods Game (Wills et al., 2018), again highlighting that dlPFC can contribute to cooperative choice. To clarify the contributions of these regions in group-based cooperation, we examined here how individual differences and social norms shape vmPFC and dlPFC activity during contributions to public goods.

Current research

We examined two variables that could influence the neural computations underlying cooperation: (1) prosocial tendencies and (2) descriptive norms. We measured brain activity while participants played one-shot PGGs ostensibly with other university students. Prosocial tendencies were measured as the proportion of cooperative PGG decisions made by each participant (Krajbich *et al.*, 2015a). To test that PGG decisions reflected broader individual differences in prosociality, we further measured giving in a dictator game (see Supplemental Section S7). We created prosocial and antisocial social norms by manipulating the feedback from the other university students in the PGG (see Figure 1). To verify these norms, we asked participants to estimate how often students cooperated at each university. These estimates indexed individual differences in norm detection.

We tested four hypotheses derived from the value-based approach: (VB_{H1}) prosocial tendencies will moderate the relative contribution of vmPFC and dlPFC, such that choices aligned with one's tendency (i.e. selfish participants free-riding or prosocial

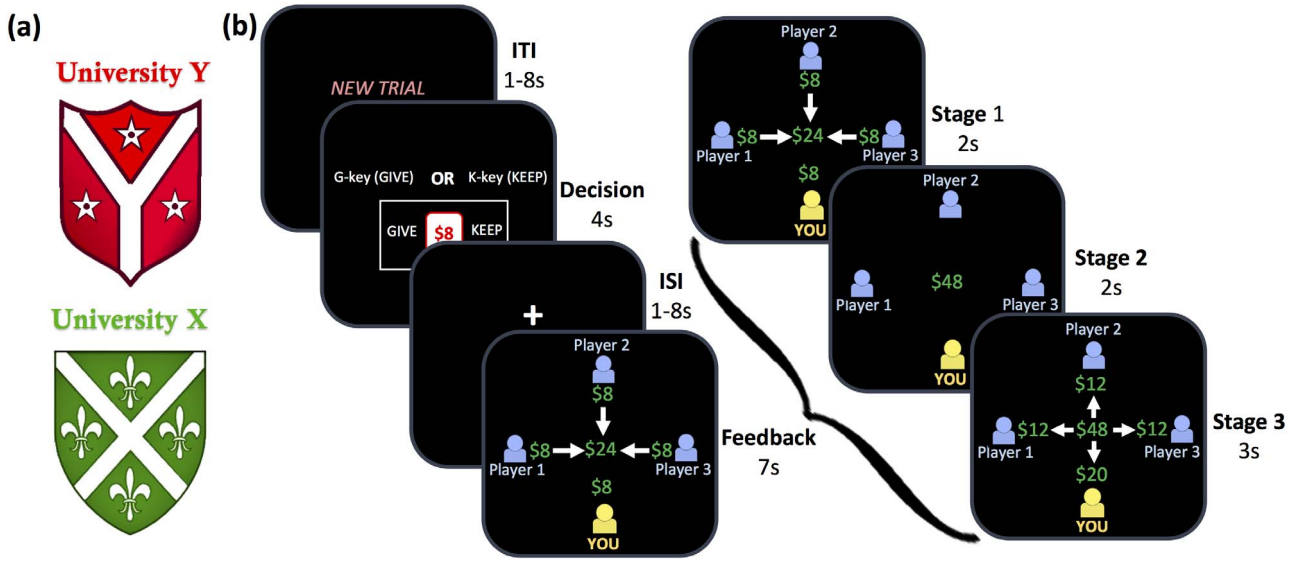

Fig. 1. Public goods game with descriptive norm manipulation. (a) At the beginning of each block, participants were instructed, 'You will now encounter students from University [X/Y]' with one of the corresponding emblems shown. The base rates were fixed for each university such that 30% of students cooperated in one university (antisocial) whereas 70% of students cooperated in the other (prosocial). Participants alternated between antisocial and prosocial schools (order counterbalanced) for a total of four blocks with 25 trials within each block. **(b)** Each trial consisted of a decision phase (4 s) and a feedback phase (7 s) that was broken down into three stages: (1) each student's contribution to the public pot, (2) the pot multiplying by two and (3) each student's earnings after the pot is evenly divided four ways. Students who gave to the pot were always pictured in blue whereas students who kept their money were displayed in yellow. ITI and ISI durations were jittered in order to dissociate neural activity between decision and outcome phases.

participants cooperating) will elicit greater vmPFC responses whereas (VB_{H2}) choices that conflict with one's tendency (e.g. selfish participants cooperating or prosocial participants free-riding) will elicit greater dlPFC-vmPFC connectivity. Social context should moderate activity in these regions, such that (VB_{H3}) deviating from group norms (i.e. cooperating with free-riders or free-riding against cooperators) will elicit greater dlPFC-vmPFC connectivity whereas (VB_{H4}) complying with group norms (e.g. cooperating with cooperators or free-riding against other free-riders) will elicit greater vmPFC activity.

Methods

Participants

Our target sample was 45–50 participants. Forty-seven university students (32 female) were recruited (via online posters) from the New York City area and paid $20 (or two research credit hours). Students ($M = 20.85$, $SD = 2.60$) spanned across 16 north-eastern colleges. All participants were right-handed, healthy, had normal or corrected-to-normal vision, and had no history of psychiatric diagnoses, neurological or metabolic illnesses. The study was approved by the review board of New York University. We report how we determined sample size, all data exclusions, all manipulations and all measures.

Five participants were excluded from analyses, leaving a sample of 42 participants. Participants were excluded if they met at least one of the criteria defined prior to data collection: (1) moving at least 3 mm across all scan sessions and (2) reporting suspicion (during debriefing) that their decisions could actually affect the payment of other participants. One participant was excluded due to criterion 1 (>5 mm motion in each scan session), and four participants were excluded due to criterion 2.

Procedure

Public goods game

Participants received instructions on how to play a PGG, including four practice trials and a thorough comprehension quiz. In each round, participants were given $8.00, which they could either keep for themselves or contribute to benefit the group. Players interacted in groups of four and contributions were doubled and split equally by all group members. On each trial, participants were given 4 s to make a decision and received no payment on trials where no response was recorded. After each choice, participants were shown feedback about the other players' decisions and resulting payouts. An intertrial interval (ITI) signaled the beginning of each round and a fixation cross interstimulus interval (ISI) separated the decision and feedback stages (Figure 1B). ITIs and ISIs were jittered 1–8 s using a Poisson distribution and randomized between participants. Participants were instructed they would never encounter the same player more than once such that each trial resembled a one-shot PGG.

Before beginning the game, we informed participants that the other players' decisions were based on previous responses from students attending two universities whose identities we had concealed (labeled University X and University Y; Figure 1A). Participants were further instructed that students from one university were more likely to give their money (prosocial school) whereas students from the other university were more likely to keep their money (antisocial school). We did not specify which university was prosocial and antisocial. Participants alternated between playing with students from the prosocial and antisocial schools across four blocks for a total of 100 trials. The identity of the universities and the order they were encountered were counterbalanced across participants. In reality, we adjusted the base rates such that players from each university gave 70% (prosocial) or 30% (antisocial) of the time. Of the 50 trials playing with

the prosocial university, for instance, participants encountered 105 'givers' (out of 150 students) with the following feedback distribution: 0 givers (2 trials), 1 giver (10 trials), 2 givers (20 trials) and 3 givers (18 trials).

Earnings across all trials were averaged and paid to participants after the study. Prosocial tendencies were computed for each participant by computing their mean level of cooperation (i.e. the proportion of trials they decided to give). Two survey measures were collected and analyzed for exploratory analyses reported in the supplement: emotional ratings for each PGG outcome (see Supplementary Sec. S1) and group identification with each university (Van Bavel and Cunningham, 2012) (see Supplementary Sec. S2). Additional survey measures were collected at the end of the study but are not included in the present analyses[1].

Dictator game

As an independent measure of prosociality, participants played two dictator games upon exiting the scanner: one with a student from the prosocial school and one with a student from the antisocial school. Participants could allocate anywhere from $0.00—$1.00 to each student using a slider bar. The amount given to both groups was averaged into an overall measure of prosociality. At the end of the study, participants were informed that no students were actually going to be paid and they were paid the full $2.00 they were originally endowed with (regardless of their actual allotment).

Explicit estimation

We next assessed the extent to which students' explicitly learned the social norms. Using a slider bar, participants estimated the percentage of students that cooperated from each school. We took the difference between these scores for each participant to compute a measure of norm detection (i.e. the extent to which students from the prosocial school cooperated more often than the antisocial school). Positive scores indicate students who (correctly) remembered the prosocial school giving more often whereas negative scores indicate students who (incorrectly) remembered the antisocial school giving more often.

Behavioral analyses

Analyses of cooperative behavior were conducted using generalized estimating equations (Liang and Zeger, 1986) (GEE) with an exchangeable correlation structure clustered on each participant[2]. These procedures account for repeated measures in a regression framework and were implemented using the geepack package (Halekoh *et al.*, 2006) in R (R Core Team, 2016). We used logistic regression with cooperation operationalized as a binary outcome for each trial (0 = Keep, 1 = Give). Parametric tests were only conducted on measures where we observed insufficient evidence for non-normality ($\alpha = 0.05$; Shapiro–Wilk test). Otherwise, bootstrap confidence intervals were computed with 10 000 simulations.

1 These include the Levenson Self-Report Psychopathy Scale (Levenson *et al.*, 1995), Groupiness Scale (Dunham & Van Bavel, unpublished), Social Value Orientation (Van Lange, 1999), and additional demographic information (e.g. political ideology, socioeconomic status, etc.)

2 Unlike traditional OLS regression, GEE parameter estimates are typically evaluated using a Wald χ^2 distribution with 1 degree of freedom.

fMRI data acquisition

Functional imaging was conducted using a Siemens (Erlangen, Germany) 3.0 Tesla Allegra head-only MRI scanner. Functional images were acquired using a customized multi-echo EPI sequence developed by the NYU Center for Brain Imaging to mitigate the effects of susceptibility artifacts in medial temporal and ventromedial regions (TR = 2000 ms; TE = 15 ms; Flip Angle = 82°; 34 3 mm slices with a 0.45 mm gap for whole-brain coverage, Matrix = 64 × 80; FOV = 192 × 240 mm; Acquisition voxel size = 3 × 3 × 3.45 mm). This sequence has been described in detail in prior work (Hackel *et al.*, 2015).

Each volume comprised 34 axial slices collected in an interleaved-ascending manner and parallel to the AC-PC line. Data were collected in four sessions, 225 volumes each (7 min and 30 sec). Six scans were acquired at the start of each run and dropped from analysis to allow magnet equilibration. During this time, participants were told which university they would be playing with. Finally, whole-brain high-resolution structural scans (T1-weighted, MPRAGE, 1 × 1 × 1 mm resolution) were acquired from all participants, coregistered with their mean EPI images and averaged together to permit anatomical localization of the functional activations at the group level.

fMRI data pre-processing

Image analysis was performed using SPM12. Images were corrected for slice-time acquisition and realigned to correct for participant motion, coregistered to structural images, transformed to conform to the default T1 Montreal Neurological Institute (MNI) brain interpolated to 3 × 3 × 3 mm, smoothed using a Gaussian kernel with a full-width-at-half-maximum of 6-mm, corrected for artifacts using the ArtRepair toolbox (Mazaika *et al.*, 2007) and detrended using the LMGS toolbox (Macey *et al.*, 2004). The blood-oxygenation-level-dependent (BOLD) signal was modeled using a canonical hemodynamic response function.

A general linear model (GLM) included (1) onset of give decisions, (2) onset of keep decisions, (3) onset of feedback after give decisions and (4) onset of feedback after keep decisions. Reaction times were entered for decision epoch durations, such that each trial was modeled as having a duration of the participant's reaction time (Grinband *et al.*, 2008). Further regressors of no interest included (5) onsets of choice epochs for missed trials (i.e. non-responses), (6) feedback epochs for missed trials, as well as six movement parameters from the realignment stage. A high-pass filter with cutoff period of 128 s was used. To analyze the impact of prosocial tendencies, first-level contrasts for Give > Keep were generated and entered into a second-level random effects analysis along with each participant's mean cooperation (i.e. proportion of give decisions) as an interaction term. To analyze the impact of descriptive norms, first-level contrasts for Give > Keep were generated separately for each of the four scanner sessions. Norm-level contrasts were computed for each participant by averaging session contrasts corresponding to each norm and then subtracting the antisocial contrast from the prosocial contrast. These contrasts were then entered into second-level random effects analyses. For a psychophysiological interaction (PPI) analysis of overall choice, vmPFC activity served as a physiological variable, choice type (free-riding *vs* giving) served as a psychological variable, and the interaction of these variables was examined; individual differences in prosociality served as a moderator in a second-level analysis. For a PPI analysis of norm congruence, vmPFC activity served as a physiological variable, choice type (congruent with norm *vs*

deviant from norm) served as a psychological variable, and the interaction of these variables was examined; individual differences in norm detection served as a moderator in a second-level analysis.

Based on past work, our hypotheses targeted specific regions of interest (ROIs): the vmPFC and the dlPFC. To constrain our analysis to these regions, we used existing ROI masks for vmPFC and dlPFC based on prior work (Wills et al., 2018). These ROIs were constructed with the MarsBar toolbox by combining corresponding structures from the Harvard–Oxford Maximum Probability Atlases (Fischl et al., 2004) (see Supplementary Figure S6). The vmPFC ROI consisted of the frontal pole, frontal medial cortex, paracingulate gyrus, subcallosal cortex, constrained by rectangular prism X = [−14, 14], Y = [10, 80], Z = [−35, 0]. The dlPFC ROI consisted of the frontal pole, inferior frontal gyrus and middle frontal gyrus, constrained by bilateral rectangular prism X = [−60, −30 (L); 30, 60 (R)], Y = [20, 70], Z = [5, 55]. Small-volume corrections were conducted using these ROI masks. Unless otherwise noted, all analyses were corrected for multiple comparisons using a voxel-wise height threshold of $P < 0.001$ combined with an appropriate cluster extent to maintain a family-wise error (FWE) rate of $P < 0.05$, using Gaussian random field theory as implemented in SPM (Friston et al., 1994).

ROI effect size estimation

Given the difficulty of drawing conclusions from null results, we conducted the following procedure to estimate effect sizes for non-significant contrasts. Using Marsbar, we first extracted each subject's average betas (i.e. predicted amplitude) for each contrast within each ROI mask. We then computed the mean amplitude across 10 000 bootstrapped simulations and report the 95% confidence interval of the resulting distribution of means.

Computational model

We fit a computational model that estimated participants' trial-by-trial expectations about the number of expected givers on each round. This model assumed participants chose to give or keep based on (1) a constant term and (2) the number of expected givers on a given round:

$$pG_t = \frac{1}{1 + e^{(-c + -\beta \times G)}} \quad (1)$$

where pG_t is the probability of giving on trial t, c is a constant term, β is a weight on the expected number of givers for that trial and G is the expected number of givers. Separate estimates of G were held for each university. On each trial, G was updated for the relevant university using a delta-learning rule:

$$G_t = G_{t-1} + \alpha\delta \quad (2)$$

where α is a learning rate parameter and δ is a prediction error related to the expected number of givers:

$$\delta = Number\ of\ Givers\ - G \quad (3)$$

The number of givers was rescaled to range from −1 to 1, so that the constant term would be interpretable at the mean number of givers. Finally, given that the proportion of givers in each group did not change over time, participants could learn the contingencies relatively early on. We therefore allowed the learning rate to decay over time with a decay parameter d between 0 and 1 (Niv et al., 2012):

$$\alpha_t = \alpha_{t-1} \times d \quad (4)$$

This model was fit using maximum likelihood estimation. We applied parameters derived from the model to each participant's data to compute trial-by-trial estimates for the expected number of givers for each participant (see Supplementary Sec. S5 for more details on the modeling approach and Supplementary Table S2 for parameter estimates). This time series was entered as a parametric modulator of choice, using the same first-level fMRI GLM described in the primary analysis. This analysis was whole-brain corrected for multiple comparisons.

Data availability

All relevant and deidentified, pre-processed data and materials have been made publicly available at the following OSF link: (https://osf.io/hfngs/).

Results

Behavioral Results

Prosocial tendencies. On average, participants cooperated on 33.57% of the trials (SD = 28.40%). Nonetheless, there were large individual differences between participants, ranging from one participant who always cooperated to seven free-riders who chose to keep on all 100 trials (see Supplementary Figure S1). Reaction times did not vary between giving and keeping or between norm condition. However, RTs did depend on individual differences in prosocial tendencies, such that cooperation tended to be faster among prosocial participants (see Supplementary Sec. S3).

Descriptive norms. To ensure participants encoded the descriptive norm manipulation, we conducted a paired t-test to assess whether participants estimated different rates of cooperation between the schools (i.e. norm detection). Participants noticed that students from the prosocial school cooperated more often ($M = 54.71\%$, $SD = 23.37\%$) than students from the antisocial school ($M = 32.10\%$, $SD = 21.30\%$), suggesting they learned that group norms differed, $t(41) = 4.34$, $d = 0.67$, $P < .001$ (see Supplementary Figure S3).

In turn, descriptive norms impacted cooperation. Participants were more likely to cooperate with students from the prosocial school ($M = 40.42\%$, $SD = 49.08\%$) than the antisocial school ($M = 24.08\%$, $SD = 42.77\%$; Odds ratio = 2.15, Wald $\chi^2(1) = 26.4$, $P < 0.001$; see Supplementary Sec. S3 for reaction time analysis). Moreover, norm compliance (i.e. preferential cooperation with the prosocial *vs* antisocial school) was associated with norm detection ($M = 22.62\%$, $SD = 33.77\%$; $r(40) = 0.36$, bootstrapped 95% CI: [0.12, 0.55]): participants who correctly recalled greater cooperation from the prosocial school were more cooperative towards the prosocial school.

Neuroimaging results

Overall cooperation. We first tested whether cooperation and defection overall were associated with neural activation in

vmPFC and dlPFC. Such findings could be consistent with models asserting that intuition drives cooperation and deliberation promotes self-interest (prosocial intuition; Rand *et al.*, 2012, 2014) or that humans are impulsively selfish and require self-control to cooperate (prosocial restraint; Stevens and Hauser, 2004; Kocher *et al.*, 2017). We tested the first prosocial restraint hypothesis by examining whether the mean signal within vmPFC (defined through an anatomical region of interest; see Methods) was higher during free-riding as opposed to cooperation (i.e. Give > Keep). Therefore, seven participants with invariant decisions (i.e. cooperating or free-riding on every trial) were excluded from this analysis, leaving 35 participants. No significant vmPFC activation was observed in this contrast at thresholds with small volume corrections ($p_{voxel} < 0.001$; $p_{cluster} < 0.05$). We then examined whether dlPFC activity increased during cooperation *vs* free-riding (i.e. Keep > Give), but observed no significant difference.

We reversed these contrasts to test the two prosocial intuition hypotheses: vmPFC is more active when cooperating (i.e. Keep > Give) and dlPFC activity increases during free-riding (i.e. Give > Keep). However, we did not observe significant differences for these contrasts. Bootstrapped 95% confidence intervals of the signal change for cooperative (relative to selfish) choices indicated that mean vmPFC signal may increase up to 0.387% or decrease up to 0.474% whereas mean dlPFC signal may increase up to 0.224% or decrease up to 0.173%. Thus, we found insufficient evidence that these regions were differentially recruited for cooperation or free-riding. These results were not consistent with simplified prosocial restraint or prosocial intuition models of cooperation. Thus, we examined whether activation in these regions depends on cooperative tendencies and social norms.

Prosocial tendencies. According to the value-based approaches, vmPFC should be more responsive during choices aligned with one's prosocial tendencies: selfish participants should display greater vmPFC responses when free-riding and prosocial participants should display greater vmPFC responses when cooperating (see VB_{H1}). This is because vmPFC has been found to represent the difference in value between chosen *vs* unchosen options (Boorman *et al.*, 2009; Nicolle *et al.*, 2012; Hackel *et al.*, 2015). As a result, vmPFC activity should be higher for a prosocial participant when giving (i.e. when relative value is higher) than when keeping (i.e. when relative value is lower). For a free-riding participant, the reverse should be true.

To test this hypothesis, we entered each participant's mean cooperation into a second-level GLM as an estimate of prosociality, and then tested the interaction between prosociality and decision type (i.e. cooperate *vs* free-ride)[3]. We observed a cluster of vmPFC for the interaction of Prosociality x Decision, $t(33) = 4.99$, $k = 53$, $p_{cluster} < 0.001$ (Figure 2A and B; see Supplementary Table S1)[4]. Consistent with value-based approach, decisions that aligned with one's prosociality elicited greater vmPFC activity: free-riders had more vmPFC activity when free-riding, whereas cooperators had greater vmPFC activity when cooperating.

We examined whether prosociality moderated dlPFC activity during choices that *conflicted* with prosociality (i.e. selfish participants cooperating or cooperative participants free-riding). For this prosociality × decision interaction, we observed two clusters within right dlPFC (both with equal cluster sizes, $k = 13$, $p_{cluster} = 0.050$). More prosocial participants showed greater right dlPFC activity when free-riding whereas participants with selfish tendencies showed greater activity when cooperating (see Figure 2C). Moreover, we obtained the same findings when using a self-report measure of cooperative preferences that was not derived from behavior and was not specific to either norm context. Specifically, after the main task, participants reported how positively and negatively they felt upon giving and keeping; when using relative positive affect during giving *vs* keeping as an alternative measure of individual cooperative preference, we observed the same patterns of activation in vmPFC and dlPFC described above (see Supplemental Figure S5, and Supplementary Sec. S1).

Some value-based models also assert that dlPFC and vmPFC become temporally correlated when making more difficult, goal-directed choices (see VB_{H2}). To test this prediction, we examined whether functional connectivity between dlPFC-vmPFC was moderated by prosociality. We conducted a psychophysiological interaction (PPI) analysis with the right dlPFC cluster for non-dominant choices (Figure 2C) and the vmPFC cluster for dominant choices as a seed region (Figure 2A). We examined connectivity as a function of cooperative *vs* free-riding decisions, with prosociality serving as an individual difference. Cooperative participants showed greater right dlPFC-vmPFC connectivity when free-riding, whereas selfish participants showed greater connectivity when cooperating, $r(33) = -0.23$, bootstrapped 95% CI: [−0.56, −0.04], $t_R = -2.00$ (Figure 2D)[5]. In other words, participants showed greater dlPFC-vmPFC connectivity when making choices that ran counter to their prosociality, consistent with the value-based approach.

Social norms. If norms amplify the value of conformity (Nook and Zaki, 2015), then a value-based approach would predict greater dlPFC-vmPFC connectivity when choosing to deviate from norms (i.e. cooperating in antisocial groups or free-riding in prosocial groups; see VB_{H3}). We tested this prediction by conducting another PPI analysis with the dominant choice vmPFC cluster (Figure 2A) as the seed region and the non-dominant choice dlPFC cluster (Figure 2C). We contrasted deviant *vs* compliant decisions (i.e. cooperating with the prosocial group and free-riding with the antisocial group *vs* free-riding with the prosocial group and cooperating with the antisocial group)[6]. We did not observe any significant connectivity changes across decision type: bootstrapped confidence intervals for deviant (*vs* compliant) choices indicated mean dlPFC connectivity may increase up to 0.172% or decrease up to 0.189%.

Our measure of norm detection indicated that not all participants learned the group norms. Since this norm-detection measure was positively associated with behavioral norm compliance, it seemed possible that norms modulated dlPFC-vmPFC connectivity in a manner dependent on how participants perceived the norms. To test this possibility, we conducted an exploratory analysis similar to our individual difference analysis above: we entered individual differences in norm detection as a predictor in the second-level GLM. We observed a positive

3 See Krajbich et al., (2015) for or a similar analytic strategy with behavioral data.

4 The *t* statistic reported for this contrast (as well as all subsequent contrasts) refers to the peak voxel within the cluster.

5 We use the t_R notation to indicate *t* statistics calculated using the robust linear regression procedure with an M estimator available in the MASS package (Venables and Ripley, 2002) for R (R Core Team, 2016).

6 One additional participant was excluded from this analysis because they never cooperated when playing with the antisocial students. This left thirty four valid participants in all fMRI analyses involving descriptive norms.

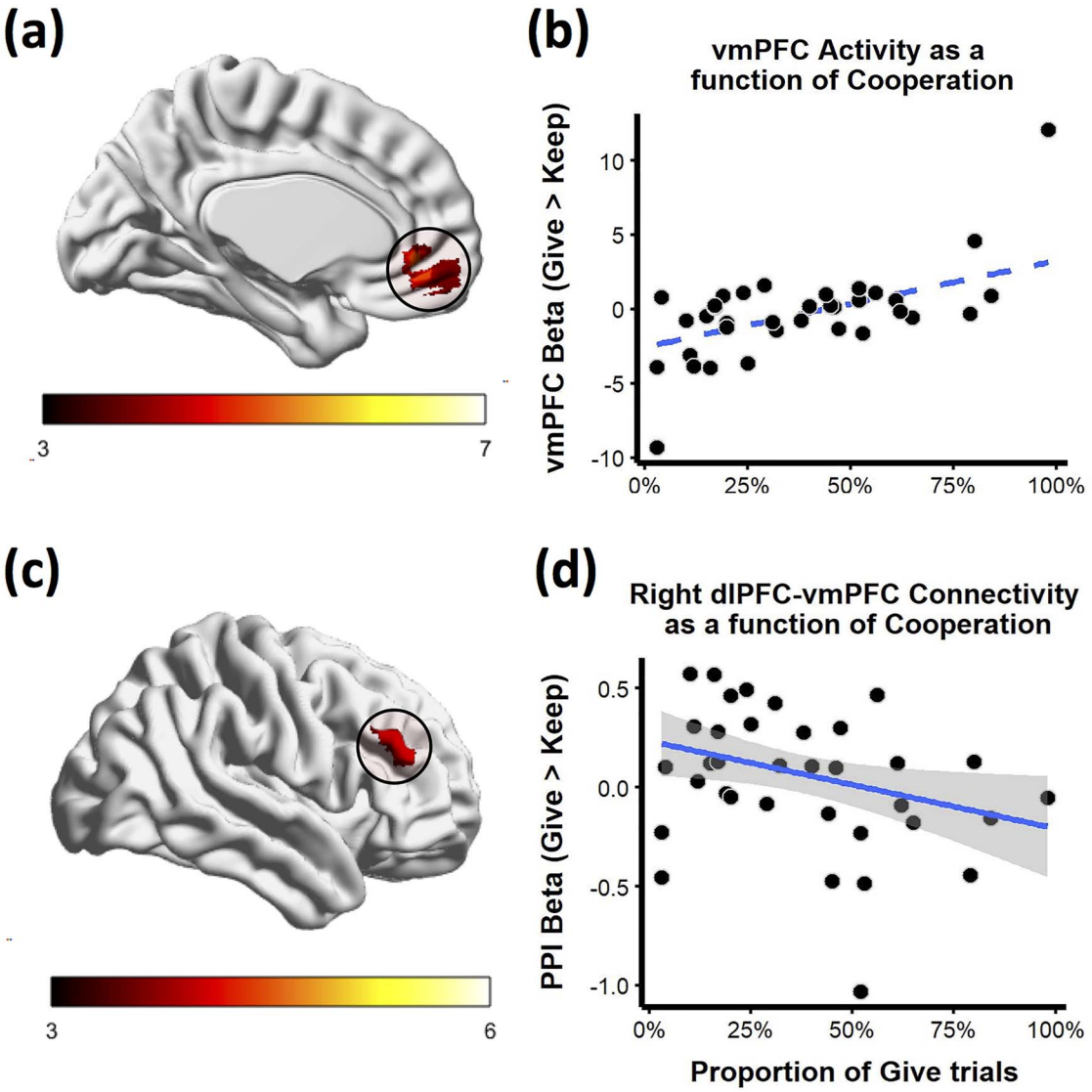

Fig. 2. vmPFC activity and dlPFC-vmPFC connectivity is moderated by prosocial tendencies. Average cooperation moderates **(a)** BOLD response in vmPFC and **(c)** right dlPFC activity during Give (*vs* Keep) decisions. Color indicates magnitude of *t* statistic. As an alternate visualization, **(b)** vmPFC cluster betas (*y*-axis) for each participant ($n = 35$) are plotted against the proportion of cooperative trials (*x*-axis). **(d)** Right dlPFC-vmPFC PPI cluster betas (*y*-axis) are plotted against the proportion of cooperative trials (*x*-axis). Robust linear regression fits are displayed with blue lines and surrounding 95% confidence interval band.

correlation between norm detection and dlPFC-vmPFC connectivity: participants who correctly perceived the norms elicited greater right dlPFC-vmPFC connectivity when deviating from the group, $r(33) = 0.50$, bootstrapped 95% CI: [0.22, 0.69], $t_R = 2.90$ (Figure 3). This suggests that vmPFC and dlPFC are more functionally correlated when (a) participants who correctly perceived norms deviated from actual group norms and (b) participants who incorrectly perceived norms complied with group norms, thus deviating from their *perceived* norms. That is, participants who incorrectly perceived norms showed greater vmPFC-dlPFC connectivity when deviating from their *subjectively* perceived norms, much as participants who correctly perceived norms showed greater vmPFC-dlPFC connectivity when deviating from the *objective* norms.

We examined whether this connectivity predicted norm compliance (i.e. preferential cooperation with the prosocial *vs* antisocial school). We observed an interaction whereby participants with greater right dlPFC-vmPFC connectivity during deviant decisions engaged in greater norm-compliant cooperation, OR = 1.97, 95% CI = [1.26, 3.10], Wald $\chi^2(1) = 8.72$, $P = 0.003$. This suggests that vmPFC and dlPFC are more functionally correlated when (a) norm-compliant participants deviate from norms and (b) non-compliant participants comply with norms; the latter finding may reflect the fact that non-compliance was correlated with incorrect perceptions of norms, as reported above. We did not observe a relationship between group norms and vmPFC activity (see Supplementary Sec. S6).

Evolving expectations. Adapting to norms requires people to update and deploy expectations about how others will cooperate. We therefore tested whether neural activity during choice reflected expectations about others' cooperation. We fit participant choices to a computational model of learning and choice, which assumed that people updated an estimate of average cooperation for each group following feedback on each trial. Estimates were updated with a prediction error, or the discrepancy between the actual and expected number of cooperators on each trial. This model allowed us to estimate, in a trial-by-trial manner, the number of givers participants expected while making

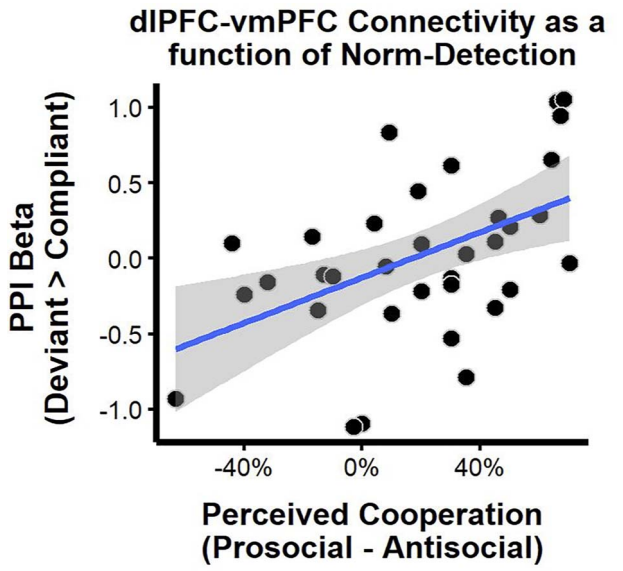

Fig. 3. Norm detection moderates dlPFC-vmPFC connectivity when deviating from the norm. Norm detection moderates right dlPFC-vmPFC connectivity for deviant decisions (e.g. free-riding in the prosocial norm or cooperating in the antisocial norm). Participants ($n = 34$) who perceived relatively more cooperation from the prosocial school *vs* antisocial school (*x*-axis) elicited heightened right dlPFC-vmPFC connectivity when deviating from (*vs* complying with) the group norm (*y*-axis). To mitigate influential points, a robust linear regression fit is displayed with blue line and surrounding 95% confidence interval band.

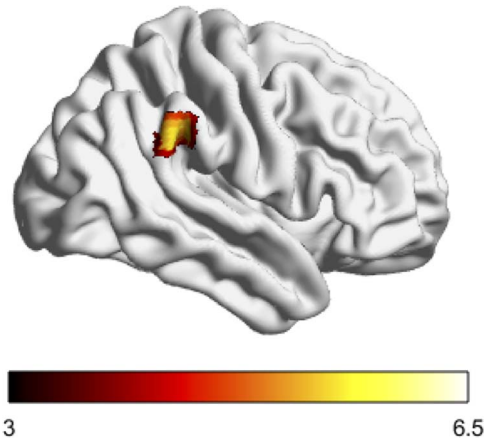

Fig. 4. rTPJ activity reflects the expected number of givers during decision-making. The number of givers anticipated during choice on a given trial—as estimated through a computational model of learning and choice—correlates with activation in rTPJ. Color indicates magnitude of *t* statistic.

decisions and analyze whether any neural regions tracked this quantity during choice.

Increased expectation of givers during choice was associated with activation near the right temporoparietal junction (Figure 4). Given this region's purported role in theory of mind tasks (Frith and Frith, 2006), this may reflect increased social cognitive processing during choices where others were expected to cooperate. If so, internalizing social norms may play a key role in future cooperation—a form of social prediction. It also suggests participants dynamically tracked expectations related to norms during decision-making.

Discussion

Value-based models contend that psychological processes supporting cooperation may hinge on idiosyncratic preferences as well as contextual factors that shift these preferences (Krajbich *et al.*, 2015a; Pärnamets *et al.*, in press). Consistent with this approach, we find that vmPFC and dlPFC activity during cooperation depends upon individual differences in prosociality and sensitivity to social norms. Specifically, prosocial participants elicited greater vmPFC activity when cooperating as well as heightened right dlPFC activity and dlPFC-vmPFC connectivity when free-riding. Self-serving participants showed the reverse pattern. Thus, neither the vmPFC nor the dlPFC exhibited a consistent role in cooperation, but instead showed greater activation when people acted consistently or inconsistently with their prosocial tendencies, respectively. These findings are consistent with the idea that vmPFC activity reflects a decision value that can be modulated by dlPFC (Rangel and Hare, 2010).

In contrast, prominent models rooted in dual-process frameworks argue that cooperation is either reflexive (*prosocial intuition* models) or primarily stems from deliberate self-control (*prosocial restraint* models). Past neuroimaging work rooted in dual-process tradition associated vmPFC with intuitive processing and dlPFC with deliberative processing. Thus, the *prosocial restraint model* implied free-riding would be associated with vmPFC activity and cooperation would be associated with dlPFC activity. In contrast, the *prosocial intuition model* implied cooperation would be associated with greater vmPFC activity and free-riding would be associated with greater dlPFC activity. When adopting this interpretation of vmPFC and dlPFC, our data still suggest that prosocial participants are intuitive cooperators, which conflicts with prosocial restraint models, and that selfish participants are deliberative cooperators, which conflicts with prosocial intuition models. Thus, our findings support a value-based interpretation of vmPFC and dlPFC activity or require a more flexible dual-process model that can account for the moderating roles of prosociality and norms.

The current research further indicates that norms shape cooperation. Participants who were most attentive to norms aligned their behavior with norms and showed greater right dlPFC-vmPFC connectivity when deviating from norms, whereas the least attentive participants showed the reverse pattern. Curiously, we found no clear evidence that decisions to conform were more valued than decisions to deviate. This conflicts with work suggesting social norms boost the value of norm compliance (Nook and Zaki, 2015). Instead, our findings suggest that norm compliance can also stem from increased functional connectivity between vmPFC and dlPFC.

Our findings raise questions about how people model the dynamic shifts in norms that vary over time and between groups. We fit a computational model of learning to understand how people represent the cooperation expected from each group. Expectation of givers during decision-making was associated with activation in a region near the right temporoparietal junction—a region related to theory of mind (Frith and Frith, 2006). This finding comports well with recent work (Park *et al.*, 2019) and suggests that social cognitive systems may interface with the construction of value to guide decisions.

The present research highlights two important factors that modulate the neuro-cognitive processes guiding cooperation: prosociality and social norms. Rather than neural activity in vmPFC *always* reflecting either prosocial or selfish behavior, the value of cooperation may be lower among selfish individuals or in selfish environments. The notion that cooperation necessarily

stems from reflexive intuitions ignores the possibility that some individuals value the outcomes of others. This idea coheres with evidence that selfish participants more often cooperate by mistake, whereas prosocial participants accidentally prioritize their self-interest (Hutcherson *et al.*, 2015). Thus, intuition may trigger cooperative mistakes from selfish participants but deter prosocial participants from second guessing their cooperation.

Conclusion

In conclusion, more flexible models are needed to specify when, and for whom, cooperation is intuitive or deliberative. This study highlights the advantage of social neuroscience methods for disambiguating the decision-making processes that guide prosocial behavior. The present findings are largely inconsistent with models that assume one mental process always supports cooperation but are consistent with a value-based approach to understanding cooperation. Although our focus on dlPFC and vmPFC helped constrain our hypotheses, cooperation is complex and draws on a widely distributed network of neural systems implicated in prosocial behavior (Lamm *et al.*, 2011; Fareri *et al.*, 2015; FeldmanHall *et al.*, 2015; Hutcherson *et al.*, 2015; Fermin *et al.*, 2016). For instance, other reward-related regions, such as caudate, are involved in prosocial choices (Lemmers-Jansen *et al.*, 2018) and may show differential patterns of activation across individuals with strong *vs* weak prosocial preferences. Similarly, as evidence accumulates that moral decisions rely on more dynamic, distributed and multi-faceted neurocognitive processing (Van Bavel *et al.*, 2015), more flexible models of value-based decision-making appear to offer a fruitful account for prosociality.

Supplementary data

Supplementary data are available at *SOCAFN* online.

Acknowledgements

We thank E. Owens for assistance with data collection, and members of the New York University Social Perception and Evaluation Laboratory for comments on the manuscript.

Funding

This work was funded by the New York University Center for Brain Imaging and the US National Science Foundation grant #1349089 awarded to J.V.B., grant #1606959 awarded to L.H., and graduate research fellowship awarded to J.W.

Conflicts of interest

There are no conflicts of interest to declare.

References

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In: *Groups, leadership, and men*, 222–36). Oxford, UK: Carnegie Press.

Barber, A.D., Carter, C.S. (2004). Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. *Cerebral Cortex*, **15**(7), 899–912.

Baumeister, R.F., Leary, M.R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, **117**(3), 497.

Bechara, A., Damasio, H., Tranel, D., Damasio, A.R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, **275**(5304), 1293–5.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., Rushworth, M.F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, **62**(5), 733–43.

Bouwmeester, S., Verkoeijen, P.P., Aczel, B., *et al.* (2017). Registered replication report: Rand, Greene, and nowak (2012). *Perspectives on Psychological Science*, **12**(3), 527–42.

Brosnan, S.F. (2018). Insights into human cooperation from comparative economics. *Nature Human Behaviour*, **2**(7), 432.

Carter, C.S., Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience*, **7**(4), 367–79.

Chib, V.S., Rangel, A., Shimojo, S., O'Doherty, J.P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, **29**(39), 12315–20.

Cialdini, R.B., Reno, R.R., Kallgren, C.A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, **58**(6), 1015.

Cooper, J.C., Kreps, T.A., Wiebe, T., Pirkl, T., Knutson, B. (2010). When giving is good: ventromedial prefrontal cortex activation for others' intentions. *Neuron*, **67**(3), 511–21.

Curry, O. S. (2016). Morality as cooperation: a problem-centered approach. In T. K. Shackelford & R. D. Hansen (Eds.). In: *The evolution of morality*, Springer International Publishing, Switzerland, pp. 27–51).

Declerck, C.H., Boone, C. (2018). The neuroeconomics of cooperation. *Nature Human Behaviour*, **2**(7), 438.

DeWall, C.N., Baumeister, R.F., Gailliot, M.T., Maner, J.K. (2008). Depletion makes the heart grow less helpful: helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin*, **34**(12), 1653–62.

Evans, A.M., Dillon, K.D., Rand, D.G. (2015). Fast but not intuitive, slow but not reflective: decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General*, **144**(5), 951.

Fareri, D.S., Chang, L.J., Delgado, M.R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience*, **35**(21), 8170–80.

FeldmanHall, O., Dalgleish, T., Evans, D., Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage*, **105**, 347–56.

Fermin, A.S., Sakagami, M., Kiyonari, T., Li, Y., Matsumoto, Y., Yamagishi, T. (2016). Representation of economic preferences in the structure and function of the amygdala and prefrontal cortex. *Scientific Reports*, **6**, 20982.

Fischl, B., Van Der Kouwe, A., Destrieux, C., *et al.* (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, **14**(1), 11–22.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J., Frith, C.D., Frackowiak, R.S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, **2**(4), 189–210.

Frith, C.D., Frith, U. (2006). The neural basis of mentalizing. *Neuron*, **50**(4), 531–4.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *Neuroimage*, **43**(3), 509–20.

Hackel, L.M., Doll, B.B., Amodio, D.M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, **18**(9), 1233.

Halekoh, U., Højsgaard, S., Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, **15**(2), 1–11.

Hare, T.A., Camerer, C.F., Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, **324**(5927), 646–8.

Hughes, B.L., Ambady, N., Zaki, J. (2017). Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. *Social Cognitive and Affective Neuroscience*, **12**(3), 372–81.

Hutcherson, C.A., Plassmann, H., Gross, J.J., Rangel, A. (2012). Cognitive regulation during decision making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. *Journal of Neuroscience*, **32**(39), 13543–54.

Hutcherson, C.A., Bushong, B., Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, **87**(2), 451–62.

Kable, J.W., Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, **10**(12), 1625.

Kocher, M.G., Martinsson, P., Myrseth, K.O.R., Wollbrant, C.E. (2017). Strong, bold, and kind: self-control and cooperation in social dilemmas. *Experimental Economics*, **20**(1), 44–69.

Krajbich, I., Bartling, B., Hare, T., Fehr, E. (2015a). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, **6**, 7455.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., Fehr, E. (2015b). A common mechanism underlying food choice and social decisions. *PLoS Computational Biology*, **11**(10), e1004371.

Lamm, C., Decety, J., Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, **54**(3), 2492–502.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron*, **64**(3), 431–9.

Lemmers-Jansen, I.L., Krabbendam, L., Amodio, D.M., Van Doesum, N.J., Veltman, D.J., Van Lange, P.A. (2018). Giving others the option of choice: an fMRI study on low-cost cooperation. *Neuropsychologia*, **109**, 1–9.

Levenson, M.R., Kiehl, K.A., Fitzpatrick, C.M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, **68**(1), 151.

Levy, D.J., Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, **22**(6), 1027–38.

Liang, K.-Y., Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Macey, P.M., Macey, K.E., Kumar, R., Harper, R.M. (2004). A method for removal of global effects from fMRI time series. *Neuroimage*, **22**(1), 360–6.

Mazaika, P., Whitfield-Gabrieli, S., Reiss, A., Glover, G. (2007). Artifact repair for fMRI data from high motion clinical subjects. Presented at the Organization for Human Brain Mapping Annual Conference, Chicago, IL., **2007**.

Nicolle, A., Klein-Flügge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., Behrens, T.E. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, **75**(6), 1114–21.

Niv, Y., Edlund, J.A., Dayan, P., O'Doherty, J.P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, **32**(2), 551–62.

Nook, E.C., Zaki, J. (2015). Social norms shift behavioral and neural responses to foods. *Journal of Cognitive Neuroscience*, **27**(7), 1412–26.

Nook, E.C., Ong, D.C., Morelli, S.A., Mitchell, J.P., Zaki, J. (2016). Prosocial conformity: Prosocial norms generalize across behavior and empathy. *Personality and Social Psychology Bulletin*, **42**(8), 1045–62.

Park, S.A., Sestito, M., Boorman, E.D., Dreher, J.-C. (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications*, **10**(1), 1–12.

Pärnamets, P., Shuster, A., Reinero, D.A., Van Bavel, J.J. (in press). A value-based framework for understanding cooperation. *Current Directions in Psychological Science*. https://journals.sagepub.com/doi/10.1177/0963721420906200.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rand, D.G. (2016). Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, **27**(9), 1192–206.

Rand, D.G. (2017). Reflections on the time-pressure cooperation registered replication report. *Perspectives on Psychological Science*, **12**(3), 543–7.

Rand, D.G., Greene, J.D., Nowak, M.A. (2012). Spontaneous giving and calculated greed. *Nature*, **489**(7416), 427.

Rand, D.G., Peysakhovich, A., Kraft-Todd, G.T., Newman, G.E., Wurzbacher, O., Nowak, M.A., Greene, J.D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, **5**, 3677.

Rangel, A., Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, **20**(2), 262–70.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D. (2002). A neural basis for social cooperation. *Neuron*, **35**(2), 395–405.

Rilling, J.K., Glenn, A.L., Jairam, M.R., Pagnoni, G., Goldsmith, D.R., Elfenbein, H.A., Lilienfeld, S.O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry*, **61**(11), 1260–71.

Saraiva, A.C., Marshall, L. (2015). Dorsolateral–ventromedial prefrontal cortex interactions during value-guided choice: a function of context or difficulty? *Journal of Neuroscience*, **35**(13), 5087–8.

Satpute, A.B., Lieberman, M.D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research*, **1079**(1), 86–97.

Smith, K.M., Larroucau, T., Mabulla, I.A., Apicella, C.L. (2018). Hunter-gatherers maintain assortativity in cooperation despite high levels of residential change and mixing. *Current Biology*, **28**(19), 3152–7.

Stevens, J.R., Hauser, M.D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, **8**(2), 60–5.

Suzuki, S., Niki, K., Fujisaki, S., Akiyama, E. (2011). Neural basis of conditional cooperation. *Social Cognitive and Affective Neuroscience*, **6**(3), 338–47.

Van Bavel, J.J., Cunningham, W.A. (2012). A social identity approach to person memory: group membership, collective

identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin*, **38**(12), 1566–78.

Van Bavel, J.J., FeldmanHall, O., Mende-Siedlecki, P. (2015). The neuroscience of moral cognition: from dual processes to dynamic systems. *Current Opinion in Psychology*, **6**, 167–72.

Van Lange, P.A.M. (1999). The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. *Journal of Personality and Social Psychology*, **77**(2), 337.

Van Lange, P.A.M., Liebrand, W.B., Messick, D.M., Wilke, H.A. (1992). *Social Dilemmas: The State of the Art. Social Dilemmas*, Theoretical Issues and Research Findings, London: Pergamon Press, pp. 3–28.

Van Lange, P.A.M., Joireman, J., Parks, C.D., Van Dijk, E. (2013). The psychology of social dilemmas: a review. *Organizational Behavior and Human Decision Processes*, **120**(2), 125–41.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Wills, J., FeldmanHall, O., NYU PROSPEC Collaboration, *et al.* (2018). Dissociable contributions of the prefrontal cortex in group-based cooperation. *Social Cognitive and Affective Neuroscience*, **13**(4), 349–56.

Zaki, J., Schirmer, J., Mitchell, J.P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, **22**(7), 894–900.

Zhu, L., Jenkins, A.C., Set, E., *et al.* (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, **17**(10), 1319.