Neuropsychologia 147 (2020) 107576

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

Me first: Neural representations of fairness during three-party interactions

Keith J. Yoder [a,*], Jean Decety [a,b]

[a] Department of Psychology, 5848 S. University Ave, University of Chicago, Chicago, IL, 60637, United States
[b] Department of Psychiatry and Behavioral Neuroscience, University of Chicago Medicine, Chicago, IL, 60637, United States

ARTICLE INFO

Keywords:
Fairness
Social decision-making
fMRI
EEG
Machine learning
Egocentric bias
Neuroeconomics
Morality

ABSTRACT

One hallmark of human morality is a deep sense of fairness. People are motivated by both self-interest and a concern for the welfare of others. However, it remains unclear whether these motivations rely on similar neural computations, and the extent to which such computations influence social decision-making when self-fairness and other-fairness motivations compete. In this study, two groups of participants engaged in the role of responder in a three-party Ultimatum Game while being scanned with functional MRI (N = 32) or while undergoing high-density electroencephalography (N = 40). In both studies, participants accepted more OtherFair offers when they themselves received fair offers. Though SelfFairness was reliably decoded from scalp voltages by 170 ms, and from hemodynamic responses in right insula and dorsolateral prefrontal cortex, there was no overlap between neural representations of fairness for self and for other. Distinct neural computations and mechanisms seem to be involved when making decisions about fairness in three-party contexts, which are anchored in an egocentric, self-serving bias.

1. Introduction

A motivation for fairness constitutes a universal cornerstone of the moral sense. Humans are deeply sensitive to issues of justice and fairness, both in their own lives and the lives of others (DesChamps et al., 2016). This sense of fairness emerges during early childhood (Cowell et al., 2019; Sloane et al., 2012; Ziv and Sommerville, 2017) and across cultures (Huppert et al., 2019), although some aspects show variability (Blake et al., 2015). However, successful social interactions often require individuals to balance their own self-interest with the wants and needs of others within specific ecological contexts.

The interplay between self-interest and other-regarding concerns plays a prominent role in behavioral economics. Arguably, the main function of morality is to regulate an individual's social interactions with others in the general direction of cooperation (Curry et al., 2019). Our moral intuitions about how to treat others and how they ought to treat us are produced by a number of evolved systems, each specialized for regulating different classes of social interactions (Cosmides et al., 2019). Though moral cognitions and behaviors are routine, they rely on both heuristics (i.e., cognitive short-cuts) and deliberate processing, involving neurocognitive computations for updating and maintaining value orientations and social expectations, representing the goals and beliefs of both self and others, and selecting adaptive responses while accounting for specific social norms (Buckholtz and Marois, 2012; Decety and Yoder, 2017; Ruff and Fehr, 2014; Stallen et al., 2018).

The standard microeconomic model of the profit maximizing firm assigns essentially no role for prosociality and social conscience. It assumes that people are mostly motivated by their material self-interest and strive to maximize their own payoffs (e.g. Bolton and Ockenfels, 2000; Kahneman et al., 1986). However, there is a vast literature in behavioral economics and psychology documenting that many people are strongly motivated by fairness and reciprocity and are willing to reward or punish other people at a considerable cost to themselves (Gintis et al., 2005). This does not mean that the notion of self-interest as being an important motivator of behavior should be abandoned altogether (van Dijk, 2013; van Dijk et al., 2004). Rather, the empirical literature has supplemented this motive with other motives that economists call social preferences (e.g., Bowles, 2016). Moreover, while research in behavioral economics has revealed prosocial tendencies in human interactions, some argue that people primarily pursue self-interest as long as they can maintain the appearance of being fair (Caviola and Faulmüller, 2014). Indeed, prosocial behavior declines when it is not observable by other affected parties or expectations to behave prosocially are reduced (Engelmann et al., 2013; Leimgruber et al., 2012; List, 2007; Overgaauw et al., 2012).

Evolutionary theory suggests that inequity aversion (i.e., the

* Corresponding author. Social Cognitive Neuroscience Laboratory, 5848 South University Avenue, Chicago, IL, 60637United States.
E-mail address: kjyoder@uchicago.edu (K.J. Yoder).

https://doi.org/10.1016/j.neuropsychologia.2020.107576
Received 20 March 2020; Received in revised form 28 July 2020; Accepted 29 July 2020
Available online 3 August 2020

negative reaction to inequitable rewards) is essential for establishing long-term cooperation among genetically heterogeneous individuals (Bowles and Gintis, 2013; Brosnan and Bshary, 2016). Consistent with this view, inequity aversion has only been observed in species which engage in repeated interactions with non-kin such as chimpanzees, macaques, ravens, and crows (Brosnan, 2013; Wascher and Bugnyar, 2013). Importantly, humans are ultra-social and care deeply about fairness, and appear to possess a genuine concern for the welfare of others which exists alongside their own self-interest (Baumard et al., 2013; Crocker et al., 2017; Henrich et al., 2010; Tomasello, 2014; Vermunt, 2014). In other words, prosocial and egoistic motivations both shape decision-making but can conflict when individuals' choices simultaneously affect themselves and others (Volz et al., 2017).

Economic games provide experimental control of objective measures of relative payoffs, and so have been previously effective at investigating fairness-related social decision-making. One popular game is the Ultimatum Game (UG) which involves two players, a Proposer and a Responder (Güth et al., 1982). The Proposer divides an endowment and offers a portion to the Responder. The Responder must decide to accept or reject the offer. If the offer is rejected, both players get nothing. If people were motivated solely by self-interest, then Proposers would offer very little money and Responders would accept any nonzero offer. However, this is rarely the case, with many Proposers choosing to offer close to 50%, and most Responders rejecting offers of less than one third of the initial amount (Gabay et al., 2014; Güth et al., 1982; Sanfey, 2007). Thus, individuals in both roles appear motivated to adhere to an "equal division" rule or fairness norm.

This appears to be a *social* norm, because unfair offers are accepted at a higher rate if they are randomly determined by a computer instead of by another person (Sanfey et al., 2003). Though there are some cultural variations, these norms appear to be universal across human societies (Cowell et al., 2016; Henrich et al., 2010). In fact, fairness preferences emerge even when individuals respond to hypothetical distributions (Camerer and Hogarth, 1999; Eriksson et al., 2017; Gillis and Hettler, 2007). Importantly, both in hypothetical contexts and when using real financial incentives, when individuals explain why they choose to reject an offer, they are most likely to select "be fair" rather than being morally right, earn money, or punish the proposer (Eriksson et al., 2017). Thus, the current study utilized hypothetical offers in order to focus on basic fairness preferences and their neural instantiations.

Past neuroeconomics work has characterized the neural mechanisms that guide Responders to accept or reject fair and unfair offers during the UG (Feng et al., 2015; Gabay et al., 2014). Converging evidence points towards a social decision-making network, which supports evaluations of fairness across social contexts (Decety and Yoder, 2017). In this network, the temporoparietal junction (TPJ) and medial prefrontal cortex play a key role in inferring the intentions of others, while the anterior insula (aINS) and dorsal anterior cingulate (dACC) conjointly work with amygdala, orbitofrontal cortex (OFC), and the striatum to facilitate updating and maintaining value representations which are integrated in vmPFC (Balleine and Killcross, 2006; Morrison and Salzman, 2010; Ruff and Fehr, 2014; Tremblay et al., 2017; Wassum and Izquierdo, 2015). Selecting appropriate responses in light of social norms is facilitated by dlPFC (Ruff et al., 2013).

Research has also used ultimatum-style games to investigate the interaction between self-interest and concern for others' welfare. One approach asks participants to play the role of Responder twice, once for themselves, and then again for another person. Another approach adds one or more players. In such three-party games, the endowment is divided between one Responder and one or more Observers (Dawes et al., 2012). While the Responder plays the same role as in the classical UG, Observers have no opportunity to influence the outcome and must passively observe the offer and decision. Two regions, dACC and aINS, extending into the ventrolateral aspect of the inferior frontal gyrus, have garnered special attention. These regions serve as core input and output nodes of the salience network, which coordinates responses to motivationally relevant stimuli (Harsay et al., 2012; Seeley et al., 2007; Shackman et al., 2011). In fMRI studies using sequential self-other decisions, hemodynamic activity in dACC and aINS is greater for unfair compared to fair offers when deciding for oneself, but only aINS responds when deciding for another person (e.g. Civai et al., 2012; Corradi-Dell'Acqua et al., 2013). Increased aINS response is observed even when distributions are randomly generated by a computer (Dawes et al., 2012), suggesting that aINS may encode deviations from expectations (e.g. equal division) to support a cognitive heuristic to reject inequality (Civai, 2013).

Studies using electroencephalography (EEG) have also investigated event-related potentials (ERPs) that are responsive to fairness. Most of this work focuses on the medial frontal negativity (MFN), which is a relatively early negative-going potential over frontocentral sites. MFN shows more negative amplitudes in response to unfair compared to fair offers (e.g. Peterburs et al., 2017). The dACC is the purported neural generator of the MFN (Gehring and Willoughby, 2002), so it is not surprising that the few studies that have used three-party UG variants with EEG find self-specific MFN amplitude increases to unfair offers (Alexopoulos et al., 2013, 2012; Ma and Hu, 2015). Positive-going potentials after the MFN, such as the late positive potential (LPP) and the closely related P3/P300, are thought to be a general marker of attention allocation towards salient stimuli (Cacioppo et al., 1996). Previous work finds greater amplitudes for these ERPs in response to larger rewards (Peterburs et al., 2013; Yeung and Sanfey, 2004). For example, in gambling tasks, P3 amplitudes are greater when outcomes affect the self, as compared to another, whether they are positive or negative (Shen et al., 2013). However, LPP effects are sometimes absent in ultimatum-style games (Peterburs et al., 2017), though one study using a three-party variant found reduced amplitudes for mutually disadvantageous outcomes (Ma and Hu, 2015). Finally, the early posterior negativity (EPN) reflects early, nearly obligatory processing (Hajcak et al., 2010). The EPN is enhanced for visual stimuli with positive valence (Keil et al., 2002; Weinberg and Hajcak, 2010). Importantly, EPN amplitudes are greater when viewing or evaluating morally good, compared to morally bad social interactions, both in children and adults (Cowell and Decety, 2015; Yoder and Decety, 2014). Investigating early stages of information processing is crucial for clarifying how interactions between self-interest and concern for others unfold over time.

A few studies have attempted to distinguish between egoistic processing and other-focused processing during social decision-making. For instance, a positive peak emerges around 260 ms when individuals make decisions about sacrificial moral dilemmas, and this component appears to track the subjective unpleasantness of those decisions (Sarlo et al., 2012). Importantly, this P260 response is specifically correlated with egoistic empathic dispositions (i.e., personal distress), but not altruistic empathic dispositions (i.e., empathic concern), suggesting that self-relevance may play an important role at this early stage of information processing (Sarlo et al., 2014). In fact, posterior scalp voltages during the same time period distinguish between decisions on behalf of strangers compared to decisions on behalf of a close friend or oneself (Zhan et al., 2020). However, it is not currently known how self-interest and concern for others impact this stage of information processing when both motivations are directly relevant to a making a decision.

To date, neuroeconomics studies of fairness preferences in three-party contexts have relied on activation-based univariate analyses. Information-based techniques, such as multivariate pattern analysis (MVPA) provide an important complement to activation-based techniques for mapping out the neural computations of decision-making. One study demonstrated overlapping representations of unpleasantness for self and other in left aINS and dACC using MVPA across three modalities: electrical shocks, disgusting tastes, and unfair monetary offers (Corradi-Dell'Acqua et al., 2016). However, participants in that study responded to UG offers for self and other in different trials. Thus, an important but unanswered question is the extent to which neural representations of fairness for the self and fairness for others are

encoded by similar or distinct neural computations, particularly when these motivations can be in conflict. The concept of shared representations was originally proposed to account for common information representation (both at the computational and neural levels) between self and other for a variety of psychological functions (Decety and Sommerville, 2003). This framework has been successfully applied mostly to the domains of actions, intentions, and emotions (Cross and Obhi, 2016). Moreover, there is now strong evidence from neuroeconomics that the expected subjective value of myriad diverse options is internally transformed into a single common currency (Levy and Glimcher, 2015; Ruff and Fehr, 2014). However, the ultimate decision value of each choice incorporates contextual information, including relevant social norms. Given that many everyday decisions, especially moral decisions, involve a tension between self-interest and concern for others, this study was designed to determine the impact of these respective motivations at the behavioral and neural levels.

In order to investigate the extent to which similar or distinct neural computations guide decision-making for self and for others, participants in the current study were requested to play the role of Responder in a three-party variant of the Ultimatum Game while undergoing functional MRI (n = 32) or EEG measurements (n = 40). Importantly, this task independently manipulated the participants' own payoffs and the payoffs of a powerless neutral observer, allowing for self-interest and concern for others to potentially compete when influencing rejection decisions (Fig. 1A). Univariate techniques were used to characterize which neuro-hemodynamic and electrophysiological signals were associated with self-interest, the welfare of the observer, or both. A complementary machine-learning approach examined whether neural signals which reliably distinguish between fair and unfair offers for the self also encode fairness for others.

2. Materials and methods

2.1. Participants

32 healthy adults participated in Study 1 between May and September of 2016. Exclusion criteria included metal in the body, reported history of major psychiatric illness, current medication for mood or behavioral problems, or a head injury resulting in unconsciousness lasting more than 10 min. One participant was excluded because of excessive head movement (>3 mm translation) and four participants were excluded for excessive noise (see fMRI data preprocessing), leaving a final sample of 27 participants in Study 1 (15 female, 13 male, M_{age} = 26.7, range = 18–54). Concurrently, 40 healthy adults participated in Study 2 between May 2016 and February 2017. Metal in the body was permitted, but all other exclusion criteria were the same as Study 1. Three participants were excluded from EEG analysis for insufficient number of clean trials per condition. The final sample in Study 2 consisted of 36 participants (23 female, 13 male, M_{age} = 23.9, range = 18–44). Participants were compensated with $20 cash. All participants provided informed written consent. All procedures were approved by the Institutional Review Board at the University of Chicago.

2.2. Task description

Participants were asked to play the role of Responder in a three-party Ultimate Game. In each trial, an anonymous hypothetical Proposer distributed $12 (displayed as stacks of cartoon $1 bills) between the Proposer, a neutral observer, and the participant. The participant's allocation was shown at the bottom of the screen. The Proposer was identified by a blue box and appeared on the left in 50% of trials (Fig. 1A). Participants could accept the offer, meaning all three players would be allowed to keep the money from the hypothetical offer, or reject the offer, meaning no player would receive any money. Participants were instructed to indicate their response as soon as they had made their decision by pressing one of two buttons. The words "Accept" and "Reject" were shown at the center of the display to remind participants of the key-response mappings. As soon as a response was selected, this text was replaced with *"You get:"* and the dollar amount earned. If the participant did not respond within 3.5 s, the trial was counted as a miss and the task was continued. Trials were separated by a jittered fixation cross (M = 4 s, SD = 1.1 s). The response-key mappings were counter-balanced across participants.

Four types of offers independently manipulated fairness for the self ("Self") and fairness for the neutral observer ("Other"). "Fair" offers were at least one third of the initial endowment (>= $4). "Unfair" offers were those less than 15% (i.e. $1). Thus, the possible offers were: SelfFair-OtherFair (4:4:4), SelfFair-OtherUnfair (6:1:5), SelfUnfair-OtherFair (6:5:1), and SelfUnfair-OtherUnfair (10:1:1).

In Study 1, participants viewed 64 distributions in each of two scanning runs. Each run was further divided into two "chunks" which were separated by a blank fixation cross. Each chunk contained 32 distributions (eight of each type) presented in random order. Responses were provided using the first and second finger of the right hand on an MRI-compatible button box (hardware issues prevented the use a left-handed button-box, though all participants expressed comfort with the response box). In Study 2, the number of trials was doubled to meet EEG/ERPs signal/noise requirements, and participants responded by pressing either the 'A' or 'L' key on a computer keyboard.

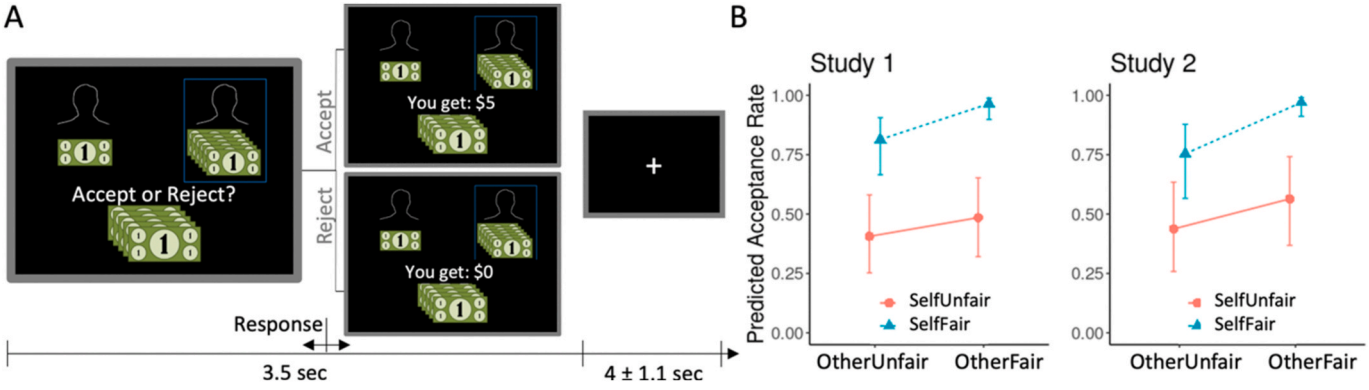

Fig. 1. Task Structure and Behavioral Responses. (A) Example SelfFair-OtherUnfair trial followed by fixation. The Proposer is marked with a blue box (50% probability of left side). Distributions remained on the screen for 3.5 s. The possible outcomes for deciding to accept (top) or reject (bottom) are shown. (B) Estimated marginal mean responses from the logistic hierarchical linear models. Errors bars represent 95% confidence intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2.3. Behavioral data analysis

Responses were analyzed separately for each study. In each, the proportion of "Accept" decisions for each distribution type was calculated for each participant. Unsurprisingly, nearly all SelfFair:OtherFair offers were accepted (M = 98.0% in Study 1 and M = 98.3% in Study 2). To account for this overdispersion, beta-binomial distributions were used to model acceptance rates. Specifically, proportions were regressed on SelfFairness (Unfair|Fair), OtherFairness (Unfair|Fair), and the Self * Other interaction with a random intercept for each subject in the 'glmmTMB' package (Brooks et al., 2017). Significant interactions were interrogated by applying a simple slopes approach for two level-1 predictors (Preacher et al., 2006).

2.4. fMRI methods

2.4.1. fMRI data acquisition

Data for Study 1 were acquired with a 3.0 T Philips Achieva MRI scanner equipped with a 16-channel head coil. High-resolution T1-weighted anatomical scans were acquired using a 3D MP-RAGE sequence (repetition time = 8 ms; echo time = 7 ms; voxel size = 0.9 × 0.9 × 0.9 mm^3; matrix = 256 × 256). Functional images were acquired along the transverse plane oriented to the AC-PC line using a single-shot EPI sequence (voxel size = 3.5 × 3.5 × 3.5 mm^3; skip gap = 0.5 mm; flip angle = 77°; matrix size = 64 × 64; echo time = 26 ms; repetition time = 2 s, reconstructed 32 slices, interleaved acquisition). Each scanning run acquired 246 volumes.

2.4.2. fMRI data preprocessing

Preprocessing was carried out in SPM12. For activation-based univariate analysis, functional images were first realigned and high-pass filtered (cutoff = 128 s). Next, the mean realigned image was coregistered to the participant's structural scan. Structural scans were then segmented into five tissue types, and those were separately normalized to the SPM12 structural template. The warping parameters obtained from this process were then applied to the functional images before smoothing using a 7 mm full width half maximum (FWHM) Gaussian smoothing kernel. For information-based multivariate analysis, each chunk was individually realigned, high-pass filtered, and coregistered to the participant's structural scan. No smoothing steps were applied to these native-spaced chunks (four per participant). In each case the ArtRepair toolbox (Mazaika et al., 2009) was applied to identify individual volumes with high motion or noise. Four participants were excluded because they required more than 15% of their volumes within a run to be repaired.

2.4.3. Contrast analysis

For the univariate fMRI analysis, general linear models (GLMs) generated three specific contrasts aimed at identifying regions whose hemodynamic responses were associated with SelfFairness, OtherFairness, or the Self * Other interaction. At the first-level, stimulus onsets and durations were modeled with a boxcar function and convolved with the canonical hemodynamic response function (HRF) in SPM12. Trials were modeled beginning at the onset of the distributions slide and ended when the participant pressed a response key. Movement parameters were modeled as nuisance regressors. Using the ArtRepair toolbox (Mazaika et al., 2009), functional images with more than 0.5 mm/TR were interpolated and a separate regressor for every such image was added to the first-level design matrix in order to "deweight" these images.

Residuals were also calculated from the first-level GLMs. Participant-specific smoothness estimates were generated using the 3dFHWMx function in AFNI (Cox, 1996). Smoothness parameters were then averaged across participants and passed to 3dClustSim to generate non-parametric threshold maps. Cluster extent was chosen to produce a bi-directional alpha level of 0.05 with a height threshold of $p < .005$.

2.4.4. Spatial MVPA analysis

For the multivariate pattern analysis, the rest block was removed, and runs were split into two chunks (before the rest or after the rest), for a total of four chunks per participant. Separate GLMs were estimated for each chunk, as in the univariate analysis, except no smoothing was applied. A 3-voxel radius searchlight was then passed through beta maps for each condition to generate samples for classification (Kriegeskorte et al., 2006). Standardized beta weights from each sphere were passed to a linear support vector machine (SVM) with the default cost parameter of 1. SVMs were trained to classify SelfFairness, OtherFairness, or the four distinct stimulus classes (one-vs-one) using leave-one-chunk-out (i.e. 4-fold) cross-validation. Threshold-free cluster enhancement (TFCE) was used to assess statistical significant while controlling for multiple comparisons at $FWE_p < .05$ (Smith and Nichols, 2009).

2.5. EEG methods

2.5.1. EEG data acquisition

Stimuli were presented via a 23″ Samsung S23A700D monitor with 2 ms response rate with a native display resolution of 1920 × 1080 at 60 Hz refresh rate. Participants viewed the stimuli while seated in a chair approximately 80 cm from the monitor resulting in a visual angle of 16°. EEG data were collected using a BrainVision ActiChamp measurement system with 32-channels laid out according to the 10–20 system. Electrode impedances were kept within the manufacturer's recommendations (max 25 kΩ). Scalp voltages were digitized at 2000 Hz with reference to Cz.

2.5.2. EEG data preprocessing

EEG data were processed offline using BrainVision Analyzer (BVA; Version 2.1). First, Cz was recovered, and data were re-referenced to the average of all channels. The average reference was selected in order to maximize detection of EPN effects (Luck and Kappenman, 2012). Next, low-pass (0.1 Hz, 12 dB/octave), high-pass (30 Hz, 24 dB/octave), and notch (60 Hz) filtering was performed using a Butterworth zero phase IIR filter. Data were then downsampled to 256 Hz before using independent component analysis (ICA) to remove ocular artifacts. As implemented in BVA, restricted infomax ICA was performed to identify 31 ICs. Ostensible blink components were identified by correlation with algorithmically defined blinks in Fp1 or Fp2. Components were then visually inspected for time-course and scalp distribution, then blink components were removed and the remaining components were back-projected into sensor-space. Following ocular correction, semi-automatic procedures were used to identity periods of artifact within individual channels. In addition to automated procedures (max gradient = 100 μV/ms; max difference within 200 ms = 200 μV; low activity cutoff = 0.5 μV) data were visually inspected for artifacts. Any channel with more than 20% of its data marked as artifact was removed and not included in ERP analyses (M = 0.6, SD = 1.1, range = 0–5). Excluded channels were topographically interpolated prior to classification analysis.

2.5.3. ERP analysis

After data cleaning, segments were extracted beginning 200 ms before stimulus onset and continuing until 1000 ms after stimulus onset. Segments were time-locked to the offer slide. After segmentation, baseline correction was applied using the average of the 200 ms preceding stimulus onset. Baseline-corrected trials were averaged together to create individual ERPs. Three difference waves were created: SelfUnfair > SelfFair, OtherUnfair > OtherFair, and Self * Other. For each difference wave, mean amplitudes within specific time windows for the posterior EPN (125–255 ms) and frontal MFN (300–350 ms) and LPP (350–800 ms) were extracted based on previous work and visual inspection of the overall grand averages (Luck and Kappenman, 2012).

Based on previous literature and visual inspection of time-course voltages and scalp maps, two clusters were created: frontal (F3, Fz,

F4) and posterior (O1, Oz, O2). Mean amplitudes for EPN (125–225 ms), MFN (300–350 ms), and LPP (350–800 ms) were extracted from the three difference waves. Only trials to which participants provided a response were included in the analysis. Any participant with fewer than 16 artifact-free segments for any of the individual conditions was excluded from analysis (n = 3). For the remaining participants, the number of clean trials per condition ranged from 39 to 64 (M = 59.0, SD = 4.9). Prior to statistical analysis, averages within each time window in each condition were examined, and any individual whose mean amplitude was more than three standard deviations above or below the mean for that condition was excluded from analysis. This identified one participant. Each ERP was assessed via *t*-test, and the false discovery rate was applied across all tests to control for multiple comparisons. In order to identify the neural generators of the observed scalp voltages, low-resolution electromagnetic tomography analysis (LORETA), as implemented in BVA, was applied to significant ERP components (Pascual--Marqui et al., 1994). Estimated current source densities were rendered on an MNI-space template brain for visualization.

2.5.4. Temporal MVPA analysis

Baseline-corrected segments for each trial were exported separately for each distribution type for each participant. When the number of available clean segments was unbalanced between types, random subsampling was used to equalize segment counts. For each of the 358 timepoints, scalp voltages at each electrode were standardized, then used as samples to train a linear SVM to distinguish between SelfFairness, OtherFairness, or the four distinct distributions. As with the MRI MVPA analysis, the default cost parameter of 1 was used. Performance was evaluated with 5 repetitions of 5-fold cross-validation. Timepoint-by-timepoint accuracies were then assess at the second-level using threshold-free cluster enhancement, with extent (e = 2/3) and height (h = 2) parameters chosen based on previous simulations (Mensen and Khatami, 2013; Smith and Nichols, 2009). 5000 Monte Carlo simulations were used to identify timepoints with accuracy above chance at FWEp < .05. Beta weights for each electrode were extracted from the linear SVMs during periods of significant decoding accuracy during EPN or LPP. Mean importance values where then normalized and projected to the scalp. LORETA was used to generate source estimates for the normalized scalp importance projections.

3. Results

3.1. Self-interest is a better predictor of behavior

In Study 1, SelfFair offers were more likely to be accepted than SelfUnfair offers (OR = 6.39, 95% CI [2.67, 15.35], p < .001), and the influence of SelfFairness was significantly greater than the influence of OtherFairness (Z = 2.57, p = .005). There was also a significant Self * Other interaction (OR = 4.52, 95% CI [1.18, 17.28], p = .028). Comparison of the simple slopes (Preacher et al., 2006) indicated that while participants were more likely to accept OtherFair distributions than OtherUnfair distributions (p < .001), this effect was more pronounced for distributions which were also SelfFair (β = 3.36) than SelfUnfair distributions (β = 1.85).

Behavioral responses in Study 2 followed a similar pattern (Fig. 1). There was a significant main effect of SelfFairness (OR = 3.94, 95% CI [1.73, 8.99], p = .001), as well as a significant Self * Other interaction (OR = 6.27, 95% CI [1.80, 21.81], p = .004). Simple slopes analysis again indicated that OtherFair distributions were more likely to be accepted, but the effect of OtherFairness was stronger for SelfFair distributions (β = 3.21, p = .001) than SelfUnfair distributions (β = 1.37, p < .001). In Study 2, the main effect of SelfFairness was only marginally greater than the effect of OtherFairness (Z = 1.50, p = .067). Model parameters are shown in Table 1.

3.2. Self-interest and third-party fairness elicit activations in different regions

Though no regions showed greater response to SelfFair distributions, multiple regions showed significantly greater response to SelfUnfair distributions (Fig. 2, Table S1), including bilateral midcingulate cortex (MCC), precentral and postcentral gyri, dorsomedial prefrontal cortex (dmPFC), and bilateral aINS.

For the OtherUnfair > OtherFair contrast, significant voxels were identified in right precentral and postcentral gyri which did not overlap with the clusters defined by the SelfUnfair > SelfFair contrast. Several regions showed greater response to OtherFair distributions, including precuneus, left inferior parietal cortex, and right inferior cortex. A partially overlapping cluster in precuneus was identified in the Self * Other interaction contrast. This contrast also revealed clusters in bilateral supramarginal gyrus.

Two additional contrasts were generated to isolate regions sensitive to self-interest when other-fairness was preserved or vice versa (Table S2). Specifically, SelfUnfair-OtherFair distributions compared to SelfFair-OtherFair distributions revealed two large clusters in bilateral occipital cortices. In contrast, SelfFair-OtherUnfair distributions were associated with decreased signal in right posterior inferior temporal cortex, compared to SelfFair-OtherFair distributions. No significant activations overlapped between the contrasts.

3.3. Self-interest predicts MFN, EPN, and LPP amplitudes

SelfFairness was associated with multiple ERPs (Fig. 3). SelfUnfair distributions elicited a larger MFN (t(35) = −3.05, FDRq = 0.032, *d* = 0.51). Conversely, SelfFair distributions were associated with more positive EPN (t(35) = 6.33, FDRq < 0.001, *d* = 1.06) and LPP (t(35) = −3.46, FDRq = 0.014, *d* = 0.58). OtherFair distributions elicited a more positive EPN than OtherUnfair distributions (Fig. 4; t(35) = 3.88, FDRq = 0.017, *d* = 0.65). There were no significant mean amplitude differences in the Self * Other difference wave. LORETA estimations identified sources throughout the frontal midline during the EPN, MFN, and LPP windows for SelfFairness and during MFN and LPP for SelfUnfairness. Sources were also identified in posterior temporal lobe and lateral occipital lobes, posterior midline cortices, and right insula. These sources appeared during EPN and expanded through the MFN and LPP time windows. During the EPN, OtherFairness and OtherUnfairness demonstrated putative generators in vmPFC, laterial occipital cortex and posterior temporal cortex. Multi-slice LORETA results are shown in Supplementary Figures S1-S4. In addition, LORETA sources were identified throughout vmPFC in each time window and dACC/SMA in middle

Table 1
Model summary for behavioral responses. A) Estimates from the multi-level logistic beta-binomial regressions of acceptance decisions on distribution type. B) Random effect variance estimate. CI: 95% confidence interval.

		Study 1			Study 2		
A	Fixed effect	Odds Ratio (CI)	Z	p	Odds Ratio (CI)	Z	p
	SelfFairness	6.39 (2.66, 15.35)	4.15	0.0000	3.94 (1.73, 8.99)	3.26	0.0011
	OtherFairness	1.38 (0.63, 2.99)	0.81	0.4184	1.66 (0.77, 3.57)	1.31	0.1916
	Self * Other	4.52 (1.18, 17.28)	2.20	0.0277	6.27 (1.8, 21.81)	2.88	0.0039
B	Random Effect	Variance	SD		Variance	SD	
	Participant	1.42	1.19		3.15	1.77	

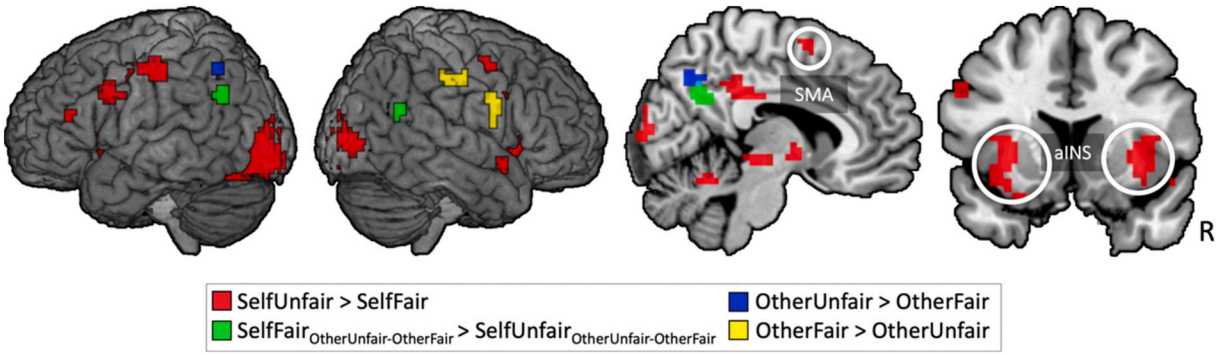

Fig. 2. Whole-brain effects of fairness evaluations. All clusters significant at FWE-p<.05 (height p < .005).

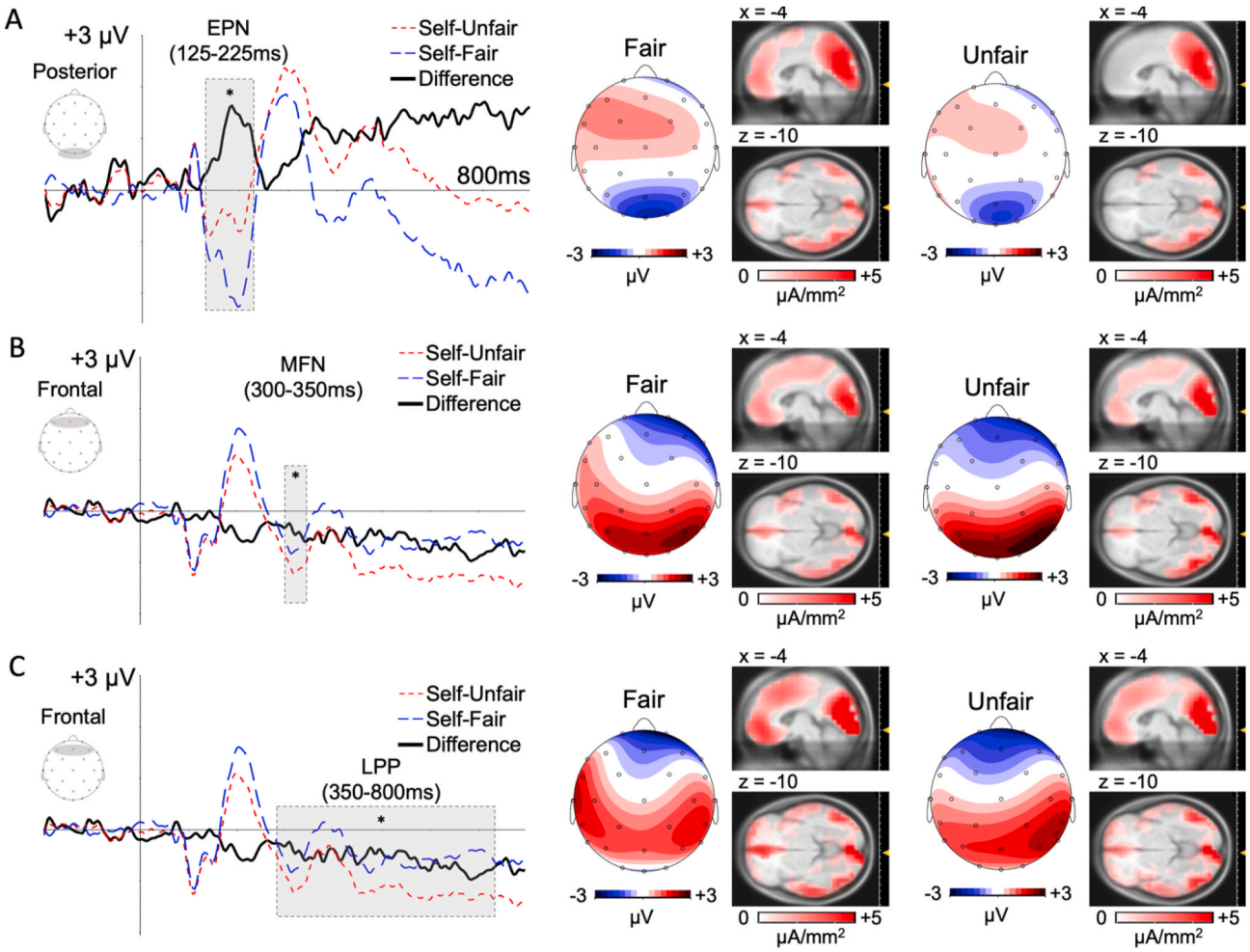

Fig. 3. ERP Effects of SelfFairness. Traces for SelfFair (blue), SelfUnfair (red), and the Unfair-Fair difference wave (black) at posterior (A) and frontal (B and C) sites. Scalp projections and rendered LORETA source density estimates are shown for (MNI z = −10 and x = 4). *$p < .05$ (FDR-corrected). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and later time windows.

3.4. Self-interest can be decoded from multiple nodes of the social decision-making network

High classification accuracy for SelfFairness was observed throughout the visual cortex and in left precentral gyrus, rdlPFC, and raINS (Fig. 5A and B, Table S3). The searchlight analysis did not identify any regions capable of decoding OtherFairness, or the four distinct classes.

3.5. Self-interest can be decoded from scalp voltages

SelfFairness were reliably decoded during the EPN time window (176–207) and during the LPP time window (Fig. 5C). Between 500 and 1200 ms, 85% of samples (597 ms) showed above chance accuracy. The SVM distinguished between OtherFair and OtherUnfair distributions briefly during the LPP time window (three samples between 570 and

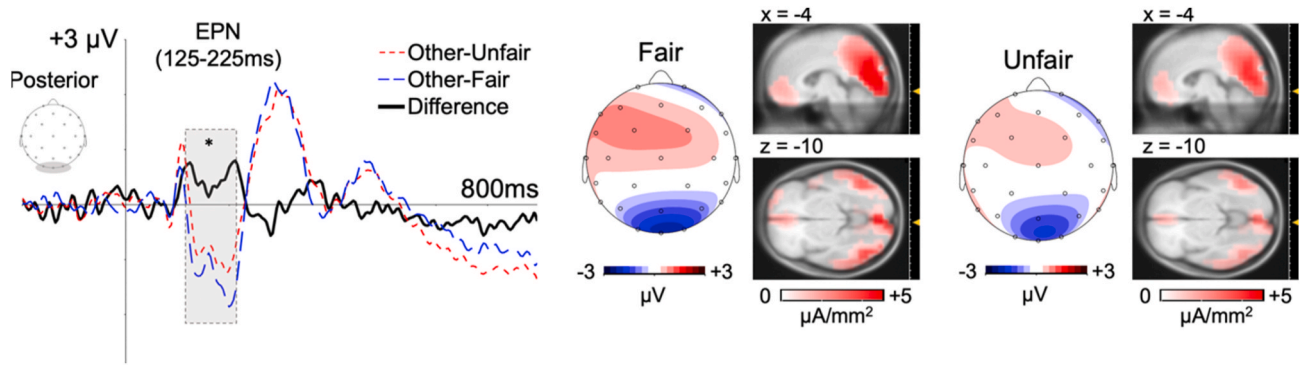

Fig. 4. ERP Effects of Other Fairness. Traces and OtherFair (blue), OtherUnfair (red), and the Unfair-Fair difference wave (black) at posterior sites. Scalp projections and rendered LORETA source density estimates are shown for (MNI z = −10 and x = 4). No significant other-directed fairness effects were observed at the frontal cluster. *$p < .05$ (FDR-corrected). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

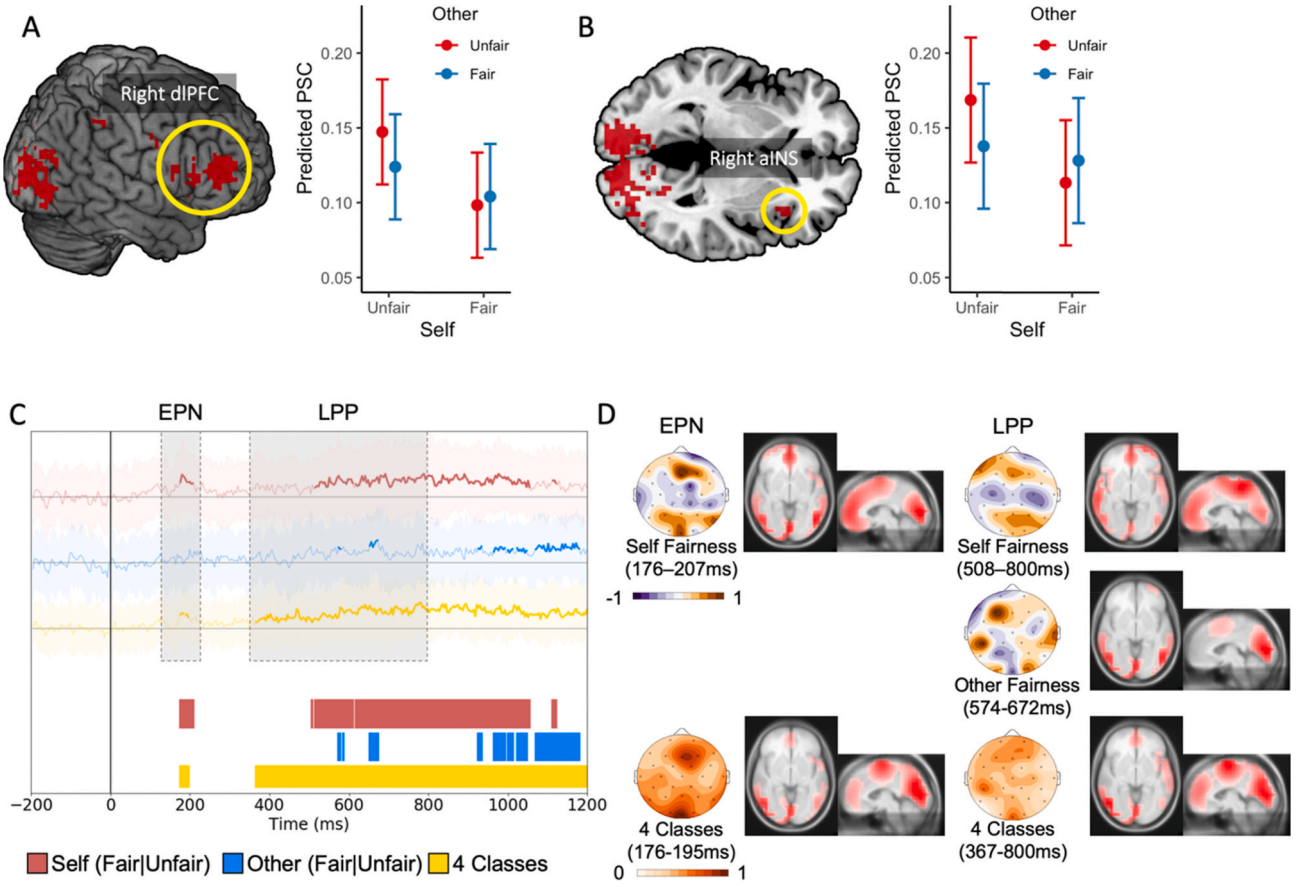

Fig. 5. Multivariate effects of fairness. Regions with above chance classification accuracy for SelfUnfair vs SelfFair are shown in red (TFCE-p < .05). Estimated marginal percent signal change are shown for dlPFC (A) and aINS (B). (C) Mean classifier accuracy for Self (red), Other (blue), and 4 Classes (yellow) support vector machine classifiers over time. Chance accuracy for each classifier is shown as a gray line (Self: 50%, Other: 50%: 4 Classes: 25%). Shaded bands represent classifier standard deviation. Thick line and rasters at bottom mark samples where accuracy was above chance at TFCE-p < .05. (D) Scalp maps of feature importance with earliest and latest significant sample. Importance is calculated as mean standardized beta weights from all SVMs with above chance accuracy during EPN or LPP time window. Importance for 4 Classes reflects mean absolute beta weights across the six one-vs-one linear SVMs. Rendered LORETA source estimates are shown beside scalp plots (MNI coordinates z = −10 and x = 4). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

590 ms, then 652–671 ms). Above chance decoding was observed for 199 ms of the final 275 ms of the trial. For the 4-distribution classification analysis, decoding accuracy was above chance during the EPN time window (176–195 ms), and then for all timepoints after 367 ms. LORETA estimates for classifiers trained to categorize SelfFairness or the 4-way classifier were associated with widespread frontal midline sources, including vmPFC, dmPFC, SMA, and ACC. All classifiers were associated with sources throughout visual cortex and middle cingulate cortex, extending into dorsal posterior mPFC. Similarly, dmPFC and right aINS were identified for the SelfFairness classifier and 4-way classifier, but not OtherFairness (Fig. 5D).

4. Discussion

Human societies are becoming increasingly interconnected, interdependent, and cosmopolitan, with individual decisions often impacting both the decision maker and multiple others, making people more accustomed to the idea that strangers can be trusted or more morally responsible (Buchan et al., 2009). Both evolutionary theory and economics predict that individuals who constitute large social groups are motivated by a conjunction of selfish and selfless interests, and both motives can compete. The current study manipulated fairness for the self and other in a three-party economic game, and used multiple methodologies to map out spatiotemporal neural dynamics sensitive to the payoffs of oneself or a powerless anonymous observer. Together, the findings point towards a primacy of self-interest over third-party concerns, not only in terms of decision-making, but in its implemention at the neural representation level. Across the fMRI and the EEG studies, participants were more sensitive to the plight of the neutral observer when they themselves received at least one third of the initial endowment. Results from the neuroimaging studies (fMRI and EEG) provide consistent results. Neuro-hemodynamic responses and MFN amplitudes showed sensitivity to fairness for the participant, and importantly, those effects did not overlap OtherFairness responses. Finally, despite reliable decoding of fairness for the self as early as 170 ms in Study 2, and in several regions of the social decision-making network in Study 1, there was no evidence for shared fairness representations for self and other.

Results from previous fMRI studies comparing fairness for the self to third-party fairness using sequential decisions point towards a dissociation between regions encoding self-interest, such as dACC and vmPFC, and those encoding deviations from the social norm of equity, in particular aINS (Civai et al., 2012; Corradi-Dell'Acqua et al., 2013; Dawes et al., 2012). Here, all decisions affected the participant as well as the neutral observer. The SelfUnfair > SelfFair contrast revealed activation in dmPFC, primarily in supplementary motor area (Fig. 2). No such clusters were detected in the OtherUnfair > OtherFair contrast, nor in the Self * Other interaction. Greater hemodynamic response for unfairness towards the self, but not the neutral observer was detected in the aINS. Importantly, our task structure was designed for self-interest and concern for others to directly conflict, since any money allocated to the neutral observer could not be allocated to the participant. One study indicates that cooperative and competitive contexts impact the extent to which neural representations of self and other overlap (Wittmann et al., 2016), and these inherent tradeoffs may alter participants' framing of their relationship with the observer. Earlier interpretations of the functional significance of aINS extending into IFG highlighted its role in rapidly updating representations of relevant stimuli (Tops and Boksem, 2011). Thus, the results presented here suggest that when self-interest and third-party fairness concerns are manipulated independently, one's own payoffs become more salient, and the impact of third-party fairness on aINS response is diminished. It will be important for future studies to utilize different tasks or vary the social identity and group membership of the players to characterize other contextual factors which might cause representations of fairness for self and other to recruit similar regions.

Interestingly, participants in the EEG study showed a similar saliency effect, limited to SelfFairness. Frontal MFN amplitudes were more negative for unfair compared to fair offers directed at the self (Fig. 3), which replicates previous ERP studies of three-party UG responding (e.g., Alexopoulos et al., 2013, 2012). Self-interest also appears to drive attention allocation and deliberation, since LPP amplitudes were greater for SelfFair distributions. Such an interpretation is also consistent with increased right dlPFC response, identified in the SelfUnfair > SelfFair contrast (Fig. 2). LORETA source estimations identify possible neural generators, and are thus not well-suited for making strong claims about exact cortical locations (Polich, 2007). Nevertheless, when combined with the fMRI results from Study 1, they can provide converging evidence for which spatially segregated neural systems are involved in a given situation. LORETA identified robust sources along the frontal midline, include vmPFC, dACC, and SMA, as well as right aINS. However, vmPFC sources were more robust for SelfFair trials during EPN and much weaker in SelfUnfair, consistent with a rapid recruitment of prefrontal structures to facilitate valuation. Moreover, dACC/SMA did not appear until the MFN and LPP time windows for SelfUnfair trials, suggesting that these core nodes of the salience network may not be as important until these middle and later stages of processing.

Notably, mean amplitudes during the EPN time window were greater for fair offers, both for the self and the neutral observer (Figs. 3A and 4). Such amplitudes have previously been implicated in emotional processing, especially for stimuli with a positive valence (Keil et al., 2002; Weinberg and Hajcak, 2010). This result is thus consistent with conceptions of fairness as inherently rewarding (Decety and Yoder, 2017; Tabibnia et al., 2008). Moreover, increased EPN amplitudes have previously been observed for third-party moral evaluations of praiseworthy actions compared to blameworthy actions (Yoder and Decety, 2014), suggesting that social fairness norms influence information processing within 200 ms. Further support for early value-based processing comes from robust vmPFC sources identified by the LORETA analysis during the EPN time window (Fig. 3).

The multivariate techniques employed in this study provide further support for a "me first" mode of processing. MVPA searchlights identified right aINS and dlPFC as regions capable of decoding fairness, but only for the self (Fig. 5). This contrasts with recent work which identified overlapping self and other fairness representations in a sequential UG (Corradi-Dell'Acqua et al., 2016). Though interpreting null effects requires caution, these results are consistent with the conclusion that when payoffs to another involve reducing one's own payoffs, fairness for self and other are representationally distinct. Similarly, the temporal searchlight identified several timepoints with significant decoding during the EPN time window for SelfFairness, but not OtherFairness. During the LPP time window, SelfFairness was decoded in over 20 times as many samples as OtherFairness (Fig. 5). The two methods converge on the conclusion that variation in one's own payoffs is robustly represented in early sensory processing areas and in regions and timepoints known to play important roles in social decision-making. The LORETA results based on the importance maps should be interpreted with caution, given that the spatial distribution of classifier beta weights is not necessarily limited by the same physical constraints as scalp EEG. Thus, while the absence of robust vmPFC sources during LPP specifically for OtherFairness suggests that value-based deliberation may be unnecessary for third-party fairness decisions, much more work would be required to justify this claim.

An alternative, though not mutually exclusive interpretation comes from models of third-party punishment (Krueger and Hoffman, 2016). In this framework, aINS is proposed to generate an aversive experience related to norm violations, while dlPFC selects appropriate punishment. In the social decision-making network, dlPFC supports goal-directed response selection (Buckholtz and Marois, 2012; Mitchell et al., 2005; Tabibnia et al., 2008; Van Bavel et al., 2015). Moreover, right dlPFC is responsive to violations of social norms (Güroğlu et al., 2010) and causally involved in adhering to social fairness norms (Ruff et al., 2013). Thus, these results could indicate that when others' payoffs can conflict with one's own payoffs, people reliably prioritize self-interest, and the scope of punishment decisions (i.e. rejecting unfair offers) is restricted to how situations impact oneself.

One potential limitation of empirical investigations using functional neuroimaging is the necessity for repeated trials. Future work using one-shot or in-person naturalistic contexts will be required to determine whether results obtained here generalize to other sorts of everyday decisions in which individuals must balance their own self-interest against fairness concerns. Moreover, the more an individual is affiliated or familiar with another, such as a family member, close friend, or in-group member, the more that other would be expected to be perceived as an extension of the self, which leads to larger overlap in neural

representations (Cheng et al., 2010; Decety and Grèzes, 2006; Wittmann et al., 2016). Future work comparing strangers and close others could empirically address this effect. Another potential limitation is that, while statistically significant, the classification analyses obtained relatively moderate accuracy, suggesting that measures with greater sensitivity may be required to better disentangle neural representations of fairness.

5. Conclusion

Fairness is foundational to morality, and a cornerstone in the social values that facilitate human cooperation in large-scale societies. Yet many social interactions involve decisions where concern for the welfare of others can conflict with self-interest. Based on the studies presented here, it appears that self-focused fairness is encoded more rapidly and occupies more cortical real estate than other-focused fairness. This supports the view put forward by Camerer and Thaler (1995) that self-interested behavior is alive and well, even in ultimatum games, and that other-interested behavior is not ready to be buried either. It will be important for future work to examine whether this effect extends to in-group conspecifics and clarify the extent to which social value orientations and fairness judgments are further influenced by competitive or cooperative contexts.

CRediT authorship contribution statement

Keith J. Yoder: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Jean Decety:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interests

The authors declare no competing interests.

Acknowledgments

This work was supported by National Institutes of Health [R01MH109329].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107576.

References

Alexopoulos, J., Pfabigan, D.M., Göschl, F., Bauer, H., Fischmeister, F.P.S., 2013. Agency matters! Social preferences in the three-person ultimatum game. Front. Hum. Neurosci. 7, 1–10. https://doi.org/10.3389/fnhum.2013.00312.

Alexopoulos, J., Pfabigan, D.M., Lamm, C., Bauer, H., Fischmeister, F.P.S., 2012. Do we care about the powerless third? An ERP study of the three-person ultimatum game. Front. Hum. Neurosci. 6, 1–9. https://doi.org/10.3389/fnhum.2012.00059.

Balleine, B.W., Killcross, S., 2006. Parallel incentive processing: an integrated view of amygdala function. Trends Neurosci. 29, 272–279. https://doi.org/10.1016/j.tins.2006.03.002.

Baumard, N., André, J.-B., Sperber, D., 2013. A mutualistic approach to morality: the evolution of fairness by partner choice. Behav. Brain Sci. 36, 59–78. https://doi.org/10.1017/S0140525X11002202.

Blake, P.R., McAuliffe, K., Corbit, J., Callaghan, T.C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K.L., Ross, E., Vongsachang, H., Wrangham, R., Warneken, F., 2015. The ontogeny of fairness in seven societies. Nature 528, 258–261. https://doi.org/10.1038/nature15703.

Bolton, G.E., Ockenfels, A., 2000. ERC : a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90, 166–193.

Bowles, S., 2016. The Moral Economy. Yale University Press, New Haven, CT.

Bowles, S., Gintis, H., 2013. A Cooperative Species: Human Reciprocity and its Evolution. Princeton University Press.

Brooks, M.E.J.K.K., van Benthem, K., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J 9, 378–400.

Brosnan, S.F., 2013. Justice- and fairness-related behaviors in nonhuman primates. Proc. Natl. Acad. Sci. Unit. States Am. 110, 10416–10423. https://doi.org/10.1073/pnas.1301194110.

Brosnan, S.F., Bshary, R., 2016. On potential links between inequity aversion and the structure of interactions for the evolution of cooperation. Behaviour 153, 1267–1292. https://doi.org/10.1163/1568539X-00003355.

Buchan, N.R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., Foddy, M., 2009. Globalization and human cooperation. Proc. Natl. Acad. Sci. Unit. States Am. 106, 4138–4142. https://doi.org/10.1073/pnas.0809522106.

Buckholtz, J.W., Marois, R., 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat. Neurosci. 15, 655–661. https://doi.org/10.1038/nn.3087.

Cacioppo, J.T., Crites, S.L., Gardner, W.L., 1996. Attitudes to the right: evaluative processing is associated with lateralized late positive event-related brain potentials. Pers. Soc. Psychol. Bull. 22, 1205–1219. https://doi.org/10.1177/01461672962212002.

Camerer, C., Thaler, R.H., 1995. Anomalies: ultimatums, dictators and manners. J. Econ. Perspect. 9, 209–219. https://doi.org/10.1257/jep.9.2.209.

Camerer, C.F., Hogarth, R.M., 1999. The effects of financial incentives in experiments: a review and capital-labor-production framework. J. Risk Uncertain. 19, 7–42. https://doi.org/10.1007/978-94-017-1406-8_2.

Caviola, L., Faulmüller, N., 2014. Moral hypocrite in economic games - how prosocial behavior is shaped by social expectations. Front. Psychol. 5, 897. https://doi.org/10.3389/fpsyg.2014.00897.

Cheng, Y., Chen, C., Lin, C.-P., Chou, K.-H., Decety, J., 2010. Love hurts: an fMRI study. Neuroimage 51, 923–929. https://doi.org/10.1016/j.neuroimage.2010.02.047.

Civai, C., 2013. Rejecting unfairness: emotion-driven reaction or cognitive heuristic? Front. Hum. Neurosci. 7, 1–3. https://doi.org/10.3389/fnhum.2013.00126.

Civai, C., Crescentini, C., Rustichini, A., Rumiati, R.I., 2012. Equality versus self-interest in the brain: differential roles of anterior insula and medial prefrontal cortex. Neuroimage 62, 102–112. https://doi.org/10.1016/j.neuroimage.2012.04.037.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R.I., Fink, G.R., 2013. Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc. Cognit. Affect Neurosci. 8, 424–431. https://doi.org/10.1093/scan/nss014.

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., Singer, T., 2016. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. Nat. Commun. 7 https://doi.org/10.1038/ncomms10904.

Cosmides, L., Guzmán, R.A., Tooby, J., 2019. The evolution of moral cognition. In: Zimmerman, A., Jones, K., Timmons, M. (Eds.), The Routledge Handbook of Moral Epistemology. Routledge, New York, NY, pp. 174–228.

Cowell, J.M., Decety, J., 2015. The neuroscience of implicit moral evaluation and its relation to generosity in early childhood. Curr. Biol. 25, 93–97. https://doi.org/10.1016/j.cub.2014.11.002.

Cowell, J.M., Lee, K., Malcolm-Smith, S., Selcuk, B., Zhou, X., Decety, J., 2016. The development of generosity and moral cognition across five cultures. Dev. Sci. 20, e12403 https://doi.org/10.1111/desc.12403.

Cowell, J.M., Sommerville, J.A., Decety, J., 2019. That's not fair: children's neural computations of fairness and their impact on resource allocation behaviors and judgments. Dev. Psychol. 55, 2299–2310. https://doi.org/10.1037/dev0000813.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173. https://doi.org/10.1006/cbmr.1996.0014.

Crocker, J., Canevello, A., Brown, A.A., 2017. Social motivation: costs and benefits of selfishness and otherishness. Annu. Rev. Psychol. 68, 299–325. https://doi.org/10.1146/annurev-psych-010416-044145.

Cross, E.S., Obhi, S.S., 2016. Shared Representations: Sensorimotor Foundations of Social Life. Cambridge University Press, Cambridge, UK.

Curry, O.S., Mullins, D.A., Whitehouse, H., 2019. Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. Curr. Anthropol. 60, 47–69. https://doi.org/10.1086/701478.

Dawes, C.T., Loewen, P.J., Schreiber, D., Simmons, A.N., Flagan, T., McElreath, R., Bokemper, S.E., Fowler, J.H., Paulus, M.P., 2012. Neural basis of egalitarian behavior. Proc. Natl. Acad. Sci. Unit. States Am. 109, 6479–6483. https://doi.org/10.1073/pnas.1118653109.

Decety, J., Grèzes, J., 2006. The power of simulation: imagining one's own and other's behavior. Brain Res. 1079, 4–14. https://doi.org/10.1016/j.brainres.2005.12.115.

Decety, J., Sommerville, J.A., 2003. Shared representations between self and other: a social cognitive neuroscience view. Trends Cognit. Sci. 7, 527–533. https://doi.org/10.1016/j.tics.2003.10.004.

Decety, J., Yoder, K.J., 2017. The emerging social neuroscience of justice motivation. Trends Cognit. Sci. 21, 6–14. https://doi.org/10.1016/j.tics.2016.10.008.

DesChamps, T.D., Eason, A.E., Sommerville, J.A., 2016. Infants associate praise and admonishment with fair and unfair individuals. Infancy 21, 478–504. https://doi.org/10.1111/infa.12117.

Engelmann, J.M., Over, H., Herrmann, E., Tomasello, M., 2013. Young children care more about their reputation with ingroup members and potential reciprocators. Dev. Sci. 16, 952–958. https://doi.org/10.1111/desc.12086.

Eriksson, K., Strimling, P., Andersson, P.A., Lindholm, T., 2017. Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. J. Exp. Soc. Psychol. 69, 59–64. https://doi.org/10.1016/j.jesp.2016.09.004.

Feng, C., Luo, Y.J., Krueger, F., 2015. Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. Hum. Brain Mapp. 36, 591–602. https://doi.org/10.1002/hbm.22649.
Gabay, A.S., Radua, J., Kempton, M.J., Mehta, M.A., 2014. The Ultimatum Game and the brain: a meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. 47, 549–558. https://doi.org/10.1016/j.neubiorev.2014.10.014.
Gehring, W.J., Willoughby, A.R., 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. Science 84 295, 2279–2282. https://doi.org/10.1126/science.1066893.
Gillis, M.T., Hettler, P.L., 2007. Hypothetical and real incentives in the ultimatum game and Andreoni's public goods aame: an experimental study. E. Econ. J. 33, 491–510. https://doi.org/10.1057/eej.2007.37.
Gintis, H., Bowles, S., Boyd, R., Fehr, E., 2005. Moral Sentiments and Material Interests: the Foundations of Cooperation in Economic Life. Cambridge University Press.
Güroğlu, B., van den Bos, W., Rombouts, S.A.R.B., Crone, E.A., 2010. Unfair? It depends: neural correlates of fairness in social context. Soc. Cognit. Affect Neurosci. 5, 414–423. https://doi.org/10.1093/scan/nsq013.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388. https://doi.org/10.1016/0167-2681(82)90011-7.
Hajcak, G., MacNamara, A., Olvet, D.M., 2010. Event-related potentials, emotion, and emotion regulation: an integrative review. Dev. Neuropsychol. 35, 129–155. https://doi.org/10.1080/87565640903526504.
Harsay, H.A., Spaan, M., Wijnen, J.G., Ridderinkhof, K.R., 2012. Error awareness and salience processing in the oddball task: shared neural mechanisms. Front. Hum. Neurosci. 6, 246. https://doi.org/10.3389/fnhum.2012.00246.
Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2010. Markets, religion, community size, and the evolution of fairness and punishment. Science 84 327, 1480–1484.
Huppert, E., Cowell, J.M., Cheng, Y., Contreras-Ibáñez, C., Gomez-Sicard, N., Gonzalez-Gadea, M.L., Huepe, D., Ibanez, A., Lee, K., Mahasneh, R., Malcolm-Smith, S., Salas, N., Selcuk, B., Tungodden, B., Wong, A., Zhou, X., Decety, J., 2019. The development of children's preferences for equality and equity across 13 individualistic and collectivist cultures. Dev. Sci. 22, e12729 https://doi.org/10.1111/desc.12729.
Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness and assumptions of economics. J. Bus. 59, 285–300.
Keil, A., Bradley, M.M., Hauk, O., Rockstroh, B., Elbert, T., Lang, P.J., 2002. Large-scale neural correlates of affective picture processing. Psychophysiology 39, 641–649.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U.S.A. 103, 3863–3868. https://doi.org/10.1073/pnas.0600244103.
Krueger, F., Hoffman, M., 2016. The emerging neuroscience of third-party punishment. Trends Neurosci. 39, 499–501. https://doi.org/10.1016/j.tins.2016.06.004.
Leimgruber, K.L., Shaw, A., Santos, L.R., Olson, K.R., 2012. Young children are more generous when others are aware of their actions. PloS One 7, e48292. https://doi.org/10.1371/journal.pone.0048292.
Levy, D.J., Glimcher, P.W., 2015. Common Value Representation - a Neuroeconomic Perspective, Handbook of Value: Perspectives from Economics, Neuroscience, Philosophy, Psychology and Sociology. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198716600.001.0001.
List, J.A., 2007. On the interpretation of giving in dictator games. J. Polit. Econ. 115, 482–493. https://doi.org/10.1086/519249.
Luck, S.J., Kappenman, E.S., 2012. The Oxford Handbook of Event-Related Potential Components. Oxford University Press, New York, NY.
Ma, Q., Hu, Y., 2015. Beauty matters: social preferences in a three-person ultimatum game. PloS One 10, 1–17. https://doi.org/10.1371/journal.pone.0125806.
Mazaika, P.K., Hoeft, F., Glover, G.H., Reiss, A.L., 2009. Methods and software for fMRI analysis of clinical subjects. Neuroimage 47, S58.
Mensen, A., Khatami, R., 2013. Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. Neuroimage 67, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027.
Mitchell, J.P., Banaji, M.R., Macrae, C.N., 2005. General and specific contributions of the medial prefrontal cortex to knowledge about mental states. Neuroimage 28, 757–762. https://doi.org/10.1016/j.neuroimage.2005.03.011.
Morrison, S.E., Salzman, C.D., 2010. Re-valuing the amygdala. Curr. Opin. Neurobiol. 20, 221–230. https://doi.org/10.1016/j.conb.2010.02.007.
Overgaauw, S., Güroğlu, B., Crone, E.A., 2012. Fairness considerations when I know more than you do: developmental comparisons. Front. Psychol. 3, 242. https://doi.org/10.3389/fpsyg.2012.00424.
Pascual-Marqui, R.D., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int. J. Psychophysiol. 18, 49–65. https://doi.org/10.1016/0167-8760(84)90014-X.
Peterburs, J., Suchan, B., Bellebaum, C., 2013. You do the math: coding of bets and outcomes in a gambling task in the feedback-related negativity and P300 in healthy adults. PloS One 8, 1–7. https://doi.org/10.1371/journal.pone.0081262.
Peterburs, J., Voegler, R., Liepelt, R., Schulze, A., Wilhelm, S., Ocklenburg, S., Straube, T., 2017. Processing of fair and unfair offers in the ultimatum game under social observation. Sci. Rep. 7, 1–12. https://doi.org/10.1038/srep44062.
Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019.
Preacher, K.J., Curran, P.J., Bauer, D.J., 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. J. Educ. Behav. Stat. 31, 437–448. https://doi.org/10.3102/10769986031004437.
Ruff, C.C., Fehr, E., 2014. The neurobiology of rewards and values in social decision making. Nat. Rev. Neurosci. 15, 549–562. https://doi.org/10.1038/nrn3776.
Ruff, C.C., Ugazio, G., Fehr, E., 2013. Changing social norm compliance with noninvasive brain stimulation. Science 84 342, 482–484. https://doi.org/10.1126/science.1241399.
Sanfey, A.G., 2007. Social decision-making: insights from game theory and neuroscience. Science 84 318, 598–602. https://doi.org/10.1126/science.1142996.
Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 84 300, 1755–1758. https://doi.org/10.1126/science.1082976.
Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., Palomba, D., 2012. Temporal dynamics of cognitive-emotional interplay in moral decision-making. J. Cognit. Neurosci. 24, 1018–1029. https://doi.org/10.1162/jocn_a_00146.
Sarlo, M., Lotto, L., Rumiati, R., Palomba, D., 2014. If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. Physiol. Behav. 130, 127–134. https://doi.org/10.1016/j.physbeh.2014.04.002.
Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D., 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. J. Neurosci. 27, 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007.
Shackman, A.J., Salomons, T.V., Slagter, H.A., Fox, A.S., Winter, J.J., Davidson, R.J., 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nat. Rev. Neurosci. 12, 154–167. https://doi.org/10.1038/nrn2994.
Shen, Q., Jin, J., Ma, Q., 2013. The sweet side of inequality: how advantageous status modulates empathic response to others' gains and losses. Behav. Brain Res. 256, 609–617. https://doi.org/10.1016/j.bbr.2013.08.043.
Sloane, S., Baillargeon, R., Premack, D., 2012. Do infants have a sense of fairness? Psychol. Sci. 23, 196–204. https://doi.org/10.1177/0956797611422072.
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061.
Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C.K.W., Sanfey, A.G., 2018. Neurobiological mechanisms of responding to injustice. J. Neurosci. 38, 2944–2954. https://doi.org/10.1523/JNEUROSCI.1242-17.2018.
Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness. Psychol. Sci. 19, 339–347. https://doi.org/10.1111/j.1467-9280.2008.02091.x.
Tomasello, M., 2014. The ultra-social animal. Eur. J. Soc. Psychol. 44, 187–194. https://doi.org/10.1002/ejsp.2015.
Tops, M., Boksem, M.A.S., 2011. A potential role of the inferior frontal gyrus and anterior insula in cognitive control, brain rhythms, and event-related potentials. Front. Psychol. 2, 1–14. https://doi.org/10.3389/fpsyg.2011.00330.
Tremblay, S., Sharika, K.M., Platt, M.L., 2017. Social decision-making and the brain: a comparative perspective. Trends Cognit. Sci. 21, 265–276. https://doi.org/10.1016/j.tics.2017.01.007.
Van Bavel, J.J., FeldmanHall, O., Mende-Siedlecki, P., 2015. The neuroscience of moral cognition: from dual processes to dynamic systems. Curr. Opin. Psychol. 6, 167–172. https://doi.org/10.1016/j.copsyc.2015.08.009.
van Dijk, E., 2013. The economics of prosocial behavior. In: Schroeder, D.A., Graziano, W.G. (Eds.), The Oxford Handbook of Prosocial Behavior. Oxford University Press, New York, NY, pp. 86–99.
van Dijk, E., De Cremer, D., Handgraaf, M.J.J., 2004. Social value orientations and the strategic use of fairness in ultimatum bargaining. J. Exp. Soc. Psychol. 40, 697–707. https://doi.org/10.1016/j.jesp.2004.03.002.
Vermunt, R., 2014. The Good, the Bad, and the Just. Ashgate Publishing Company, Burlington, VA.
Volz, L.J., Welborn, B.L., Gobel, M.S., Gazzaniga, M.S., Grafton, S.T., 2017. Harm to self outweighs benefit to others in moral decision making. Proc. Natl. Acad. Sci. U.S.A. 114, 7963–7968. https://doi.org/10.1073/pnas.1706693114.
Wascher, C.A.F., Bugnyar, T., 2013. Behavioral responses to inequity in reward distribution and working effort in crows and ravens. PloS One 8. https://doi.org/10.1371/journal.pone.0056885.
Wassum, K.M., Izquierdo, A., 2015. The basolateral amygdala in reward learning and addiction. Neurosci. Biobehav. Rev. 57, 271–283. https://doi.org/10.1016/j.neubiorev.2015.08.017.
Weinberg, A., Hajcak, G., 2010. Beyond good and evil: the time-course of neural activity elicited by specific picture content. Emotion 10, 767–782. https://doi.org/10.1037/a0020242.
Wittmann, M.K., Kolling, N., Faber, N.S., Scholl, J., Nelissen, N., Rushworth, M.F.S., Wittmann, M.K., Kolling, N., Faber, N.S., Scholl, J., Nelissen, N., 2016. Self-other mergence in the frontal cortex during cooperation and competition. Neuron 91, 482–493. https://doi.org/10.1016/j.neuron.2016.06.022.
Yeung, N., Sanfey, A.G., 2004. Independent coding of reward magnitude and valence in the human brain. J. Neurosci. 24, 6258–6264. https://doi.org/10.1523/jneurosci.4537-03.2004.
Yoder, K.J., Decety, J., 2014. Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. Neuropsychologia 60, 39–45. https://doi.org/10.1016/j.neuropsychologia.2014.05.022.
Zhan, Y., Xiao, X., Tan, Q., Li, J., Fan, W., Chen, J., Zhong, Y., 2020. Neural correlations of the influence of self-relevance on moral decision-making involving a trade-off between harm and reward. Psychophysiology 1–15. https://doi.org/10.1111/psyp.13590.
Ziv, T., Sommerville, J.A., 2017. Developmental differences in infants' fairness expectations from 6 to 15 months of age. Child Dev. 88, 1930–1951. https://doi.org/10.1111/cdev.12674.