Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Uncovering individual variations in bystander intervention of injustice through intrinsic brain connectivity patterns

Yancheng Tang[a,1], Yang Hu[b,1,*], Jie Zhuang[c], Chunliang Feng[d], Xiaolin Zhou[a,b,*]

[a] Key Laboratory of Brain-Machine Intelligence for Information Behavior (Ministry of Education and Shanghai), School of Business and Management, Shanghai International Studies University, Shanghai, China
[b] School of Psychology and Cognitive Science, Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, East China Normal University, Shanghai, China
[c] School of Psychology, Shanghai University of Sport, Shanghai, China
[d] School of Psychology, Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, Center for Studies of Psychological Application, Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, China

ARTICLE INFO

Keywords:
Third-party intervention
Resting-state fMRI (rs-fMRI)
Resting-state functional connectivity (rs-FC)
Inter-subject representational similarity analysis (IS-RSA)
Individual difference

ABSTRACT

When confronted with injustice, individuals often intervene as third parties to restore justice by either punishing the perpetrator or helping the victim, even at their own expense. However, little is known about how individual differences in third-party intervention propensity are related to inter-individual variability in intrinsic brain connectivity patterns and how these associations vary between help and punishment intervention. To address these questions, we employed a novel behavioral paradigm in combination with resting-state fMRI and inter-subject representational similarity analysis (IS-RSA). Participants acted as third-party bystanders and needed to decide whether to maintain the *status quo* or intervene by either helping the disadvantaged recipient (*Help* condition) or punishing the proposer (*Punish* condition) at a specific cost. Our analyses focused on three brain networks proposed in the third-party punishment (TPP) model: the salience (e.g., dorsal anterior cingulate cortex, dACC), central executive (e.g., dorsolateral prefrontal cortex, dlPFC), and default mode (e.g., dorsomedial prefrontal cortex, dmPFC; temporoparietal junction, TPJ) networks. IS-RSA showed that individual differences in resting-state functional connectivity (rs-FC) patterns within these networks were associated with the general third-party intervention propensity. Moreover, rs-FC patterns of the right dlPFC and right TPJ were more strongly associated with individual differences in the helping propensity rather than the punishment propensity, whereas the opposite pattern was observed for the dmPFC. *Post-hoc* predictive modeling confirmed the predictive power of rs-FC in these regions for intervention propensity across individuals. Collectively, these findings shed light on the shared and distinct roles of key regions in TPP brain networks at rest in accounting for individual variations in justice-restoring intervention behaviors.

1. Introduction

One of the striking features of human society is the inherent concern for justice (Sabbagh and Schmitt, 2016). In situations where justice has been violated, individuals often take action to restore justice by either punishing perpetrators or helping victims, even when they are not directly affected by the injustice, and/or when such acts come at a personal cost. These behaviors, known as third-party interventions, encompass both third-party punishment (TPP) and help (TPH), and are unlikely motivated by self-interest (Fehr and Fischbacher, 2004a, 2004b; Leliveld et al., 2012). As such, they are considered a hallmark of morality, playing a pivotal role in upholding and enhancing the social norm (Fehr and Fischbacher, 2004a; Skarlicki et al., 2015).

However, the extent to which third-party bystanders engage in these interventions can vary. On the one hand, studies utilizing incentivized economic games have yielded mixed findings regarding the preference of third-party bystanders for a specific type of intervention. Typically, these studies create an injustice scenario in which a decision-maker (the transgressor) divides an amount of money unequally between oneself and the other person, often in a way advantageous to oneself. The

* Corresponding authors.
E-mail addresses: yanghu@psy.ecnu.edu.cn (Y. Hu), xz104@psy.ecnu.edu.cn (X. Zhou).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2023.120468
Received 3 August 2023; Received in revised form 1 November 2023; Accepted 20 November 2023
Available online 1 December 2023

participant, acting as a third party and endowed with an extra amount of money (i.e., an endowment), has to decide whether to costly intervene with their own endowment by punishing the transgressor or compensating the victim, or keep the endowment. Several studies demonstrated a clear preference for helping the victim over punishing the transgressor (Dhaliwal et al., 2021; FeldmanHall et al., 2014; Raihani and Bshary, 2015; Van Doorn and Brouwers, 2017; van Doorn et al., 2018), even when punishing was more efficient in restoring justice than helping (van Doorn et al. (2018). However, there was also evidence indicating an opposite preference (McAuliffe and Dunham, 2021; Stallen et al., 2018). On the other hand, there is large heterogeneity among individuals in terms of their propensity (e.g., frequency, cost amount) for each specific type of intervention behavior (Fehr and Fischbacher, 2004b; Leliveld et al., 2012). A representative example from a seminal study on third-party punishment identified four different subgroups of participants who differed in either overall punishment amount or the pattern of punishment, depending on injustice scenarios (Fehr and Fischbacher, 2004b). Although previous research has explored personality traits (Hu et al., 2020, 2015; Leliveld et al., 2012; Lotz et al., 2011) and endogenous hormones (Wang et al., 2022a) as potential contributors, little is known about the general and differential neurobiological bases underlying individual variations in different forms of third-party intervention behaviors.

Studies leveraging functional Magnetic Resonance Imaging (fMRI) have explored the neural substrates of third-party intervention, with a primary focus on third-party punishment (Buckholtz et al., 2008, 2015; Ginther et al., 2016; Zhong et al., 2016). In particular, researchers employing the incentivized TPP task showed that affective-related regions, such as anterior insula (AI) and anterior cingulate cortex (ACC), are involved in encoding the severity of injustice (i.e., inequity level of the monetary distribution), while the valuation of punishment behaviors is associated with ventral medial prefrontal cortex (vmPFC), posterior cingulate cortex (PCC), and temporoparietal junction (TPJ). Moreover, studies using scenario-based tasks that manipulated the outcome and intention of injustice revealed a key role of dorsolateral prefrontal cortex (dlPFC) in integrating both types of information and guiding third-party punishment behaviors (Ginther et al., 2016). This dlPFC-punishment association was further demonstrated by a transcranial magnetic stimulation (TMS) study that provided causal evidence for the association (Buckholtz et al., 2015). Some of these regions, such as dlPFC and TPJ extending to inferior parietal cortex, were also reported to be engaged in distinguishing the two types of intervention or related decision-making processes (Civai et al., 2019; Hu et al., 2015).

Built upon these findings, recent studies have begun to address the issue of individual differences by appealing to resting-state fMRI (rs-fMRI; Biswal et al., 1995; Power et al., 2014). Compared to task-based fMRI studies, rs-fMRI studies are by definition independent of tasks, offering the advantages of minimal requirements on participants (Dubois and Adolphs, 2016) and larger sample size, and making them generally suitable for investigating individual differences. Importantly, rs-fMRI studies utilize spontaneous functional connectivity (rs-FC), which characterizes the temporal correlations of spontaneous low-frequency BOLD signals between brain regions. This neural index provides a robust and unique fingerprint for characterizing an individual's intrinsic brain functional architecture from a network perspective (Fox and Raichle, 2007; Greicius et al., 2003; Lee et al., 2013). Previous studies have combined connectome-based predictive modeling with sophisticated network measures (e.g., graph theory; Bullmore and Sporns, 2009; He and Evans, 2010) to reveal the potential of utilizing spontaneous brain activities in accounting for individual variations in a range of social behaviors (Bellucci et al., 2018; Feng et al., 2021, 2018; Li et al., 2022a, 2022b; Lu et al., 2019). However, the intrinsic neural network basis of third-party intervention behaviors remains poorly understood, and to our knowledge, only one study has examined individual differences in TPP propensity using rs-FC (Yang et al., 2021).

Here, we seek to examine the link between spontaneous brain connectivity and third-party intervention behaviors across individuals by taking a novel analytical approach, known as inter-subject representational similarity analysis (IS-RSA). IS-RSA allows us to test the intuition that individuals who exhibit similar behavioral responses or possess similar personality traits should also display similar neural signals (Finn et al., 2020). IS-RSA focuses on the relationships between individuals and on geometric properties in the high-dimensional space created using rs-FC patterns (i.e., where each single rs-FC serves as a dimension) or behavioral patterns (i.e., where each single behavioral measure serves as a dimension), rather than on individual differences in a single rs-FC or behavioral tendency. Consequently, IS-RSA offers the advantage of using second-order isomorphism to associate the high-dimensional space (i.e., representation geometry) of brain data with behavioral data across individuals (Kriegeskorte and Kievit, 2013). This characteristic sets it apart from other analytical approaches (e.g., regression, predictive modeling) that establish connections between behavioral responses and neural signals. Indeed, the IS-RSA approach has been employed to reveal how interindividual differences in neural patterns correspond to variations in a broad range of individual measures across various domains, including behavioral propensity (Hu et al., 2021; van Baar et al., 2019), personality traits (Finn et al., 2018; Wang et al., 2022b), and subjective experiences (Chen et al., 2020). Notably, IS-RSA was recently applied to rs-FC data to investigate the relationship between spontaneous neural patterns and social cognitive abilities across individuals (Iyer et al., 2023; Li et al., 2023; Wu et al., 2023).

In the present study, participants first underwent an rs-fMRI session and then completed a third-party intervention task. In this task, they were presented with a series of unequal monetary split, each made by a proposer, and decided whether to maintain the *status quo* or intervene either by helping the disadvantaged recipient in half of the trials (the *Help* condition) or punishing the proposer (the *Punish* condition) in the other half, at a specific cost. Compared to traditional tasks with both intervention options available at the same time, our design offers two advantages. First, it allows for independent measurement of the propensity for each specific type of intervention. Second, and more importantly, it enables the construction of a general intervention propensity by mapping individuals' help and punishment propensities in a 2-dimension (2D) space. This 2D index contains all information about the propensity for each type of intervention, with no need to calculating average propensities across the two conditions.

We aimed to achieve two goals using IS-RSA. The primary goal was to examine individual differences in the rs-FC pattern that accounts for third-party intervention behaviors across individuals. The secondary goal was to investigate how this association differs between punishment and help interventions. Our analyses took advantage of a theoretical TPP brain network model identified in the previous literature (Krueger and Hoffman, 2016) by focusing our IS-RSA on 18 regions of interests (ROIs; known as *nodes*) distributed across three core networks, namely the SN (i.e., bilateral AI, dACC, and amygdala), the DMN (i.e., bilateral vmPFC, dmPFC, PCC, and TPJ) and the CEN (i.e., bilateral dlPFC and PPC). For each node, we constructed a neural inter-subject representational dissimilarity matrix (IS-RDM) that captures the differences in node-specific rs-FC patterns between each pair of participants. These node-specific neural IS-RDMs were then correlated with different behavioral IS-RDMs that characterize inter-individual differences in either the general intervention propensity or the propensity specific to each type of intervention. By examining these correlations, we were able to identify nodes that commonly and differentially reflect individual variations in third-party intervention behaviors. To provide additional validation for these findings, we performed *post-hoc* predictive regression analyses to examine whether the rs-FC patterns of nodes identified through IR-RSA could effectively predict intervention behaviors across individuals.

Given that nodes in these networks (SN, DMN, CEN) have been found to be crucially engaged in predicting second- or third-party punishment behaviors across individuals (Li et al., 2022b; Yang et al., 2021), we hypothesized that inter-individual differences in rs-FC patterns of nodes distributed across three networks would account for individual variations in the general intervention propensity. Considering the potential differences between TPH and TPP, we further hypothesized that the role of nodes in these networks could diverge between the *Help* and *Punish* conditions.

2. Materials and methods

2.1. Participants

In total, 61 healthy undergraduates or graduates were recruited. Seven participants failed to complete the entire experiment due to reasons such as fatigue, claustrophobia, mismatched equipment, or other subjective reasons, resulting in a final sample of 54 participants (30 females; 22.2 ± 1.6 years old, ranging from 19 to 26, 2 left-handedness). None of them reported a history of neurological or psychiatric disorders. Written informed consents were collected from all participants. The study was conducted in accordance with the Declaration of Helsinki and were approved by the ethical committee of Shanghai University of Sport.

2.2. Image acquisition

Images were acquired with a Siemens Prisma 3-Tesla scanner at the Shanghai University of Sport, China. The rs-fMRI images were collected using an echo-planar imaging (EPI) sequence, consisting of 244 volumes (repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle = 80°, slice number = 62, slice thickness = 2 mm, field of view (FOV) = 212 × 212 mm^2, voxel size = 2 × 2 × 2 mm^3). Moreover, the high-resolution T1-weighted structural images were collected through a magnetization-prepared rapid gradient-echo (MPRAGE) sequence (TR = 2300 ms, TE = 2.98 ms, flip angle = 9°, slice number = 176, slice thickness = 1 mm, FOV = 248 × 256 mm^2, voxel size = 1 × 1 × 1 mm^3).

2.3. Procedures

Participants underwent an rs-fMRI scan and completed a third-party intervention task in the scanner. During the 8-min rs-fMRI scan, participants were required to keep their eyes open, fixate on a cross, let their mind wander, and avoid falling asleep (Speer et al., 2022; Yang et al., 2021).

The third-party intervention task was adapted from the classical TPP game. On each trial, participants were presented with an unequal monetary split made by a proposer and decided whether to maintain the *status quo* or change it by implementing an intervention proposal to help the disadvantaged recipient (the *Help* condition) or to punish the proposer (the *Punish* condition) at a specific cost (see **Supplementary Fig. S1**). In each condition, we manipulated two orthogonalized independent variables, the injustice level between the proposer and the recipient (i.e., ranging from 10 to 120 points in increments of 10, resulting in 12 levels), as well as the intervention cost incurred by the participant (i.e., ranging from 8 to 38 out of 260 points, in increments of 5, resulting in 7 levels) in a parametric manner, resulting in a total of 84 unique justice-restoring intervention proposals (i.e., 84 trials; see **Supplementary Methods** for details). To measure the intervention propensity, we calculated the proportion of trials in which participants chose to change across the 84 trials in the *Help* and *Punish* conditions, respectively. Analysis of covariance only revealed a significant main effect of intervention type on intervention propensity ($F_{(1,\,51)} = 24.48$, $p < 0.001$, $\eta^2 = 0.32$, see **Supplementary Fig. S2**), with neither gender nor age showing a significant effect (both $ps > 0.1$). These measures were used to construct the general intervention propensity for later analyses (see below for details). Visual stimuli were presented using *Psychtoolbox* (http://psychtoolbox.org/; Brainard and Vision, 1997; Pelli and Vision, 1997), and they were back-projected on a screen outside the scanner using a mirror system attached to the head coil.

2.4. Image preprocessing

The rs-fMRI data preprocessing was performed using the Configurable Pipeline for the Analysis of Connectomes (C-PAC, https://fcp-indi.github.com; Craddock et al., 2013), on a cloud-based platform (http://www.humanbrain.cn, Beijing Intelligent Brain Cloud, Inc). In particular, the first 10 volumes of each participant's rs-fMRI data were discarded. Next, motion correction was applied to correct for head movement between volumes. Then, the framewise displacement was calculated based on the rigid body image realignment parameters (Power et al., 2012, 2014). No participant was excluded due to excessive head motion (>20 % time points with FD >0.5 mm or mean FD > 0.5 mm, Snyder et al., 2021; Wu et al., 2016). Note that the pipeline did not entail the removal or interpolation of frames with excessive motion. Additionally, skull stripping was carried out to remove non-brain tissues. These skull-stripped images were registered to anatomical space using linear transformation, followed by a white-matter boundary-based transformation and then the prior white-matter tissue segmentation from FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL). These images were normalized to the standard Montreal Neurological Institute (MNI) space. Motion artifacts were removed using ICA-AROMA with partial component regression (Pruim et al., 2015). Finally, a series of nuisance variables, including global mean signals, white matter signals, cerebrospinal fluid signals, and 24-parameter head-motion parameters, were regressed out from the time course of each voxel. The pre-processed rs-fMRI time courses are available via OSF (https://osf.io/s8dvg/).

2.5. Construction of node-specific resting-state functional connectivity (rs-FC) pattern

All follow-up analyses were performed using R 4.2.3 (R Core Team, 2014). The behavioral data and codes for replicating these analyses are available via OSF (https://osf.io/s8dvg/). We focused our analyses on 18 brain nodes in the three core neural networks proposed by the TPP brain model (Krueger and Hoffman, 2016). Note that all nodes (expect the right dmPFC) have multiple sites with different coordinates based on a previous study (**Supplementary Table S1;** for visualization, see **Supplementary Table S2 and Fig. S3**, Li et al., 2022b). Hence, for each node, we built a series of spheres with a 5-mm radius centering at these coordinates. The node-specific time courses during rs-fMRI scan were first averaged over all voxels within each sphere (i.e., applicable for right dmPFC only) and then over all spheres (i.e., applicable for all other nodes except the right dmPFC). We performed a *Pearson* correlation between each pair of node-specific time courses, resulting in a 17 × 18 matrix (see **Supplementary Fig. S4** for the mean connectivity matrix). We referred to each column in this matrix as the node-specific rs-FC pattern (i.e., a 17 × 1 vector; Fig 1). These procedures were repeated for all participants.

2.6. Inter-subject representational similarity analysis (IS-RSA)

We employed IS-RSA to scrutinize the relationship between individual variability of rs-FC patterns and third-party intervention behaviors (Fig. 1; Chen et al., 2020; van Baar et al., 2019). First, we investigated the intrinsic neural substrates associated with individual differences in the general third-party intervention propensity. To this end, we constructed a neural representational dissimilarity matrix (RDM) for each node by calculating the Euclidean distance of the node-specific rs-FC pattern between each pair of participants. Next, we defined the general intervention propensity for each participant within a two-dimension (2D) space of intervention propensities in the *Help* and *Punish* conditions, instead of averaging across the two conditions. Specifically, we built a behavioral RDM (the *General* RDM), which captured individual differences in the general intervention propensity through

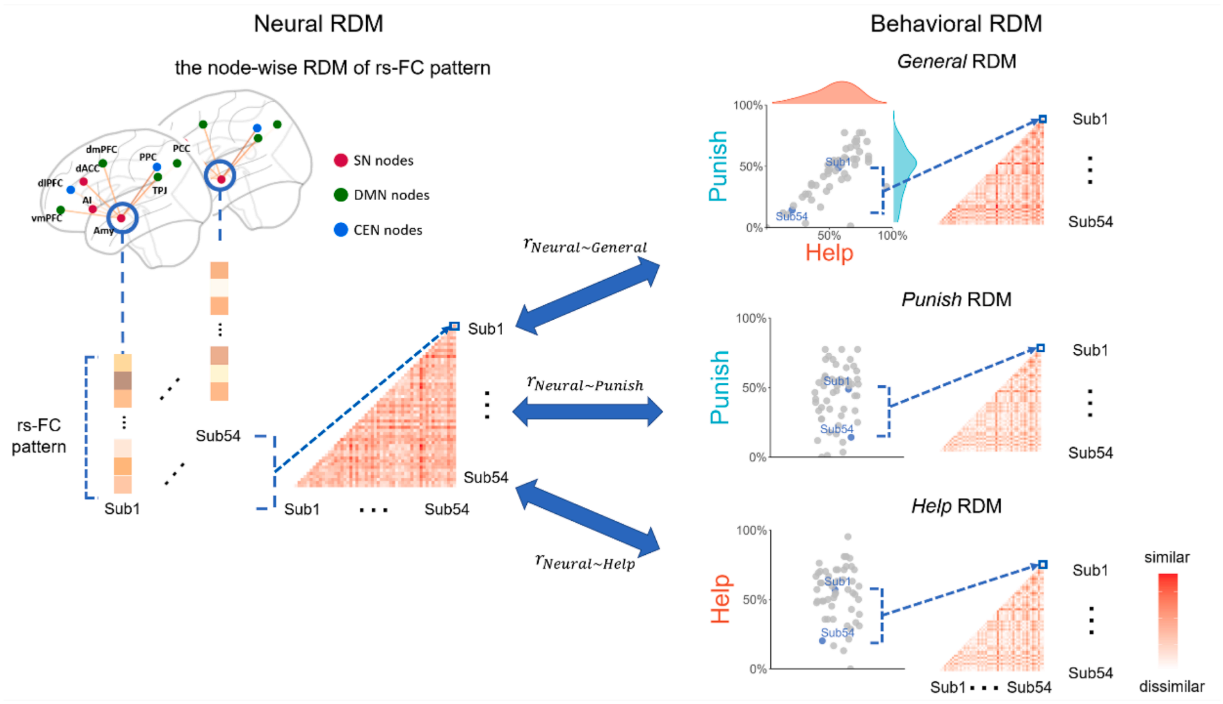


Fig. 1. Workflow of IS-RSA. Our analyses focused on the 18 brain nodes in the third-party punishment (TPP) brain model. For each node, we calculated subject-wise rs-FC patterns using the *Pearson* correlations between this node and all other nodes. Then, we built the neural RDM by calculating the Euclidean distance between all pairs of participants. Each grid in the neural RDM reflects the inter-subject dissimilarity of rs-FC patterns between a specific pair of participants. We built three behavioral RDMs for different goals. In particular, we defined the general intervention propensity of each participant on a two-dimension space of intervention propensity in the *Help* and *Punish* conditions. Then we built the *General* RDM using the pairwise Euclidean distance between individuals on this 2D propensity space to characterize individual differences in the general intervention propensity. The *Punish* and *Help* RDMs were built in a similar vein, except that the inter-subject dissimilarity was simply characterized as the 1D vector. We calculated *Spearman*-rank correlation coefficients between each of these behavioral RDMs and the neural RDM. Abbreviations: SN, salience network; DMN, default-mode network; CEN, central executive network; Amy, amygdala; AI, anterior insula; dACC, dorsal anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; PCC, posterior cingulate cortex; TPJ, temporoparietal junction; dlPFC, dorsolateral prefrontal cortex; PPC, posterior parietal cortex; and L, left; R, right.

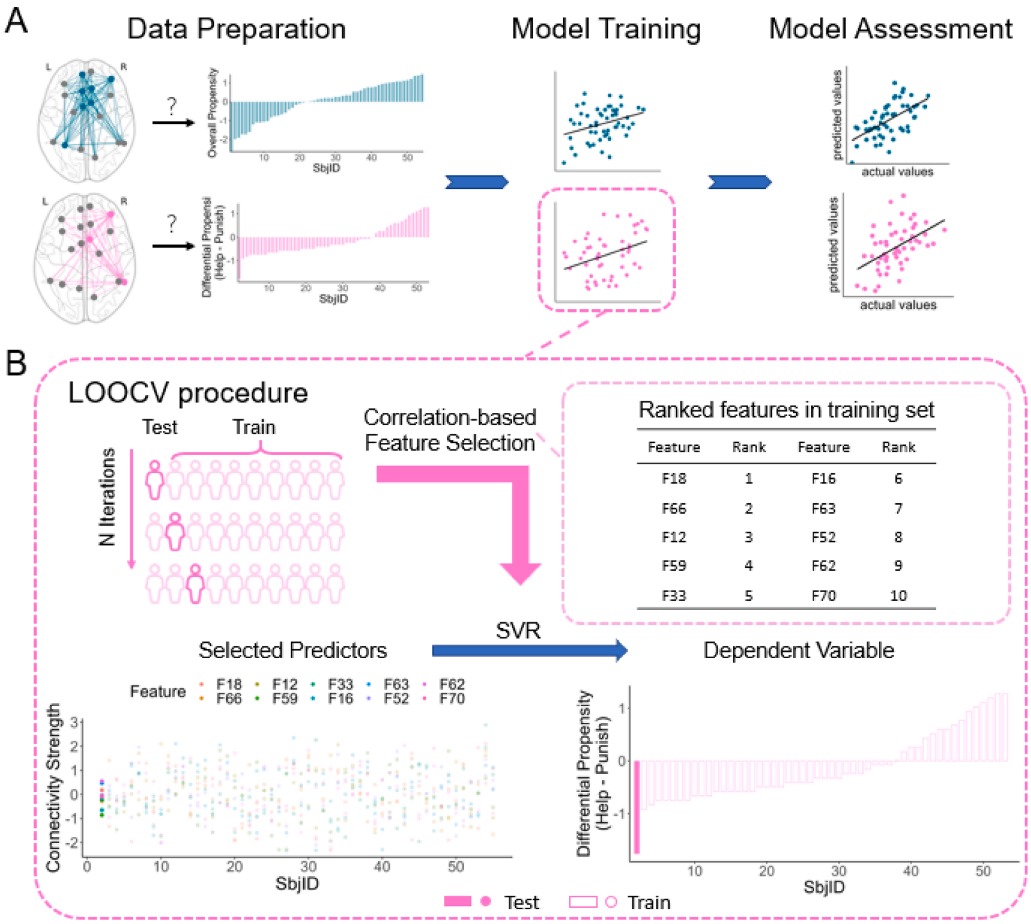
pairwise Euclidean distance between individuals (van Baar et al., 2019). To identify brain regions where rs-FC patterns are associated with third-party intervention, we calculated *Spearman* rank correlations between neural RDMs of each node and the *General* RDM. The statistic significances were obtained via permutation test. Specifically, we randomly shuffled the values of each RDM and re-computed the correlation, which was repeated for 5000 times to generate a null distribution of correlation. The permuted *p*-value was then computed based on the null distribution using a one-tailed test (Chen et al., 2020; Nili et al., 2014).

We also examined the node in which rs-FC patterns were selectively associated with inter-subject variations in the help and punishment propensities. We constructed the *Help* RDM and the *Punish* RDM by calculating the absolute difference in the intervention propensity between each pair of participants in the *Help* and *Punish* conditions respectively. We then computed *Spearman* correlations (*Spearman's* ρ) between the neural RDMs of each node and two behavioral RDMs separately (i.e., ρ_{Help}, ρ_{Punish}) and found out nodes showing significant neural-behavioral correlations in each condition. Then, we computed the differential correlation values (i.e., $\Delta\rho = \rho_{\text{Help}}$ - ρ_{Punish}) for these significant nodes. In particular, a positive $\Delta\rho$ indicates that the rs-FC patterns of a node were selectively associated with the helping propensity across individuals, while a negative $\Delta\rho$ indicates the opposite (i.e., a selective association with the punishment propensity). To obtain the statistical significance, we performed a permutation test in which we randomly shuffled the condition labels (*Help* and *Punish*) and recompute the *Spearman* correlation for 5000 times, resulting in a null distribution of $\Delta\rho$. Note that all these permuted *p*-values were corrected for multiple comparisons across nodes using the false-discovery rate (FDR) correction method.

2.7. *Post-hoc predictive modeling*

To validate the results of IS-RSA, we performed *post-hoc* predictive modeling analyses to examine whether rs-FC in the nodes identified by IS-RSA could predict general and differential intervention propensities, respectively. We built two predictive regression models (Fig. 2). The *Overall* model aimed to examine whether rs-FC of the nodes associated with individual differences in the general intervention propensity, as identified by IS-RSA, could predict the average intervention propensity over the two conditions (mean ± SD: 49.5 % ± 18.7 %). The *Differential* model aimed to examine whether the rs-FC of the nodes related with individual differences in the help or punishment propensity, as identified by IS-RSA, could predict the differential intervention propensity between the *Help* and *Punish* condition (Δpropensity = propensity$_{\text{Help}}$ – propensity$_{\text{Punish}}$: mean ± SD: 9.3 % ± 14.1 %). Note that two participants whose differential intervention propensities fell outside three SDs of the mean (Δpropensity: 61.9 % and 65.5 %) were excluded from the analysis of the *Differential* model.

Following the established procedure in previous studies (Shen et al., 2017; Song et al., 2021; Speer et al., 2022), we performed prediction analyses using support vector regression (SVR) with the R packages *kernlab* (version 0.9–32; Karatzoglou et al., 2004), *caret* (version 6.0–94; Kuhn, 2008) and *mlr3* (version 0.16.0; Lang et al., 2019). The data were normalized prior to analysis. Leave-one-out cross validation (LOOCV) was utilized to assess the performance of the prediction model. In LOOCV, one participant is held out as the test set while the remaining *N* - 1 participants are used as the training set, where *N* refers to the total



Feature	Rank	Feature	Rank
F18	1	F16	6
F66	2	F63	7
F12	3	F52	8
F59	4	F62	9
F33	5	F70	10

Fig. 2. Workflow of predictive modeling analysis. (A) We separately trained two predictive models (i.e., the *Overall* model and the *Differential* model) to examine whether the rs-FC of the regions identified in IS-RSA (i.e., solid dots in the glass brain) carried sufficient information to predict the intervention propensity across individuals. (B) The leave-one-out cross validation (LOOCV) procedure in support vector regression (SVR). In LOOCV, one participant is treated as test set while the remaining *N* - 1 participants are used as the training set, where *N* refers to the number of participants. This procedure was repeated *N* times to ensure each participant could be used once as testing set. Within each iteration, we initially selected the 10 features using Pearson-correlation-based feature selection (as listed in the table "Ranked features in training set"). Using these features, we trained the SVR model in the training set (predictors: blank circles in the scatter plot; dependent variables: blank bars in the histogram) to predict the intervention propensity of the test set (predictors: filled circles in the scatter plot; dependent variables: the filled bar in the histogram). Finally, we calculated the mean square error (MSE) and correlation coefficient between the actual and the predicted intervention propensity to assess the model performance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number of participants. This procedure was repeated *N* times to ensure each participant served as the test set once, maximizing the information obtained from limited data. Within each iteration, we initially selected the top 10 features (i.e., rs-FC between nodes) from the training set based on strongest *Pearson* correlation coefficients between rs-FCs and intervention propensities (see **Supplementary Table S3** for all correlations; see **Supplementary Table S4** for the prediction model performance using various feature selection criteria). Subsequently, these selected features were used to train the prediction model and test in test set. To determine the optimal hyperparameters in SVR, a nested cross validation procedure was implemented within the training set (see **Supplementary Fig. S5** and **Supplementary Methods** for details about nested cross-validation framework and SVR model). Finally, we calculated the mean square error (MSE) and correlation coefficient between the actual and predicted intervention propensities across all participants. To obtain the statistical significance, we conducted a permutation test with 5000 iterations. Since our primary goal was to explore the relationship between the rs-FC pattern and the third-party intervention propensity pattern rather than to build a prediction model of the third-party intervention behavior, we did not rigorously follow the full workflow of the connectome-based predictive modeling, which could be addressed in future studies that could include an independent sample for out-of-sample prediction.

3. Results

3.1. IS-RSA results

3.1.1. Overall correlation between individual differences in the rs-FC pattern and the general intervention propensity

We found that inter-subject variations in the general intervention propensity were associated with the rs-FC patterns of bilateral dACC (left: $\rho = 0.069$, $p = 0.011$; right: $\rho = 0.056$, $p = 0.042$), bilateral dmPFC (left: $\rho = 0.097$, $p < 0.001$; right: $\rho = 0.073$, $p = 0.011$), left vmPFC ($\rho = 0.081$, $p < 0.001$), right dlPFC ($\rho = 0.11$, $p < 0.001$), and left PPC ($\rho = 0.079$, $p = 0.006$), covering all three subordinate networks proposed by the TPP theoretical brain model (Fig. 3; also see **Supplementary Fig. S6** for the inter-subject variations in the rs-FC pattern of all nodes reflecting individual differences in the general intervention propensity).

3.1.2. Differential correlations between individual differences in the rs-FC patterns and the propensities for help and punishment

We found that individual differences of rs-FC in the right TPJ ($\rho_{Help}= 0.09$, $\rho_{Punish} = -0.01$, $\Delta\rho = 0.10$, $p = 0.003$) and right dlPFC ($\rho_{Help}=$

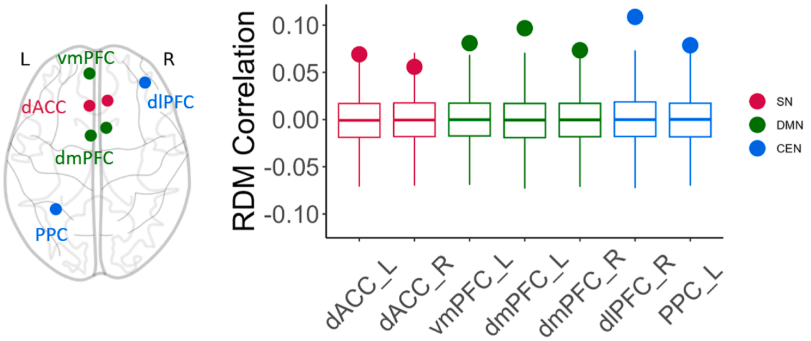


Fig. 3. Overall correlation between individual differences in rs-FC patterns and the general intervention propensity. Left: Nodes whose rs-FC pattern reflected the general intervention propensity. Right: The null distribution of correlation coefficients (boxplots) in permutation test and the true values (dots). Abbreviations: SN, salience network; DMN, default-mode network; CEN, central executive network; dACC, dorsal anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; PPC, posterior parietal cortex; L, left; R, right.

0.13, $\rho_{Punish} = 0.06$, $\Delta\rho = 0.07$, $p = 0.041$) selectively reflected heterogeneity in the helping propensity, whereas the opposite neural-behavioral association was observed in the right dmPFC ($\rho_{Help}= 0.04$, $\rho_{Punish} = 0.11$, $\Delta\rho = -0.07$, $p = 0.023$; Fig. 4; also see **Supplementary Fig. S7** for inter-subject variations in the rs-FC pattern of all nodes reflecting individual differences in the help and punishment propensity, respectively).

3.2. *Post-hoc predictive modeling*

Using SVR, we showed that the *Overall* model consisting of 11 rs-FCs between nodes reflecting individual differences in the general intervention propensity identified by IS-RSA were able to predict the general intervention propensity across participants ($MSE = 51.78$, $p = 0.048$; $r = 0.26$, $p = 0.024$; Fig. 5A; see Fig. 5B for contributing predictors). Likewise, as shown in our *Differential* model, the rs-FC between nodes selectively reflecting individual differences in either the help or punishment propensity identified by IS-RSA also contributed to predicting differential intervention propensities across participants ($MSE = 46.71$, $p = 0.046$; $r = 0.31$, $p = 0.007$, Fig. 6; see Fig. 6B for contributing predictors). Taken together, these findings validated the IS-RSA results by showing that regions exhibiting inter-subject variations in rs-FC, which captures the variations in intervention propensity, contain sufficient information to predict the intervention propensity across individuals.

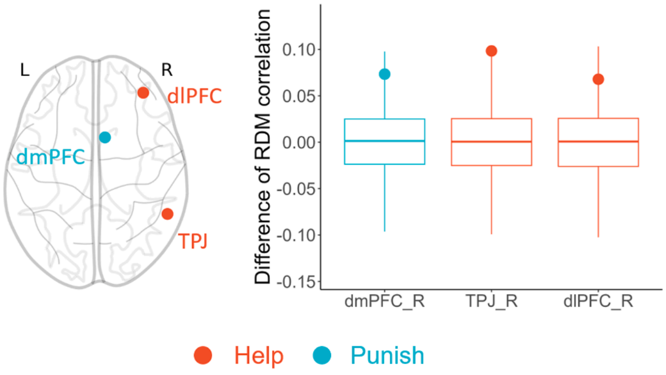


Fig. 4. Differential correlations between individual differences in the rs-FC pattern and the propensities for help and punishment. Left: Nodes whose rs-FC pattern reflected the intervention propensity difference between the *Help* and *Punish* conditions. Right: Null distributions of correlation coefficient differences (boxplots) in permutation test and the true values (dots). Abbreviations: dmPFC, dorsomedial prefrontal cortex; TPJ, temporoparietal junction; dlPFC, dorsolateral prefrontal cortex; L, left; R, right.

4. Discussion

When witnessing a situation involving a transgression of justice, different bystanders may exhibit different responses in their effort to restore justice. While some individuals may turn a blind eye to the transgression, others are inclined to take action through various modes of intervention. Here, we adopted a novel behavioral paradigm together with rs-fMRI to explore the intrinsic neuro-network substrates underlying such heterogeneity in the third-party intervention behaviors across individuals. In particular, we applied IS-RSA to examine the association between individual variations in the general and specific intervention propensities (i.e., help or punishment) and the corresponding rs-FC patterns within nodes (regions) distributed across networks, as proposed by the TPP brain network model, i.e., the SN, DMN and CEN. This approach allowed us to identify the distinct roles of nodes and their within- or between-network connections in idiosyncratic associations with the third-party intervention behaviors.

In line with our hypothesis, we found that rs-FC patterns in regions covering all the three networks were associated with individual differences in the overall third-party intervention propensity. Specifically, rs-FC patterns in two nodes in the SN, namely the right AI and dACC, were found to be involved in capturing variations in the overall intervention propensity across third parties. Previous task-based fMRI studies have established that both regions serve as critical hubs in responding to norm violations, such as unfair situations encountered in Ultimatum Game (Feng et al., 2015; Sanfey et al., 2003) or TPP Game (Zhong et al., 2016). We also observed that rs-FC patterns in nodes within the DMN, including TPJ, dmPFC and vmPFC, significantly contributed to individual differences in the third-party intervention propensity. These regions have been implicated in processes concerning social cognition, especially mentalizing or theory-of-mind (ToM; Frith and Frith, 2006; Schurz et al., 2014). In the literature of third-party intervention, these regions are considered to integrate the affective processes targeting at the victim and the intention of the perpetrator into a blame signal (Buckholtz and Marois, 2012; Krueger and Hoffman, 2016; Treadway et al., 2014). Moreover, rs-FC patterns in dlPFC and PPC, key nodes in the CEN, also accounted for individual variations in the third-party intervention propensity. Two lines of evidence have suggested that the CEN is engaged in both types of intervention. On the one hand, both regions are engaged in judging the blameworthiness of the morally wrongful actions (Crockett et al., 2017; Cushman et al., 2012), and dlPFC is also shown to play a crucial role in transforming the blame signals into the decision to punish (Buckholtz et al., 2015). On the other hand, both regions have been shown to be essential for the computation processes that weigh self-risk and other-need during helping-related decisions (Hu et al., 2018). Taken together, our results are consistent with regions identified in previous task-based fMRI studies, and further demonstrate the involvement of specific nodes within the TPP brain

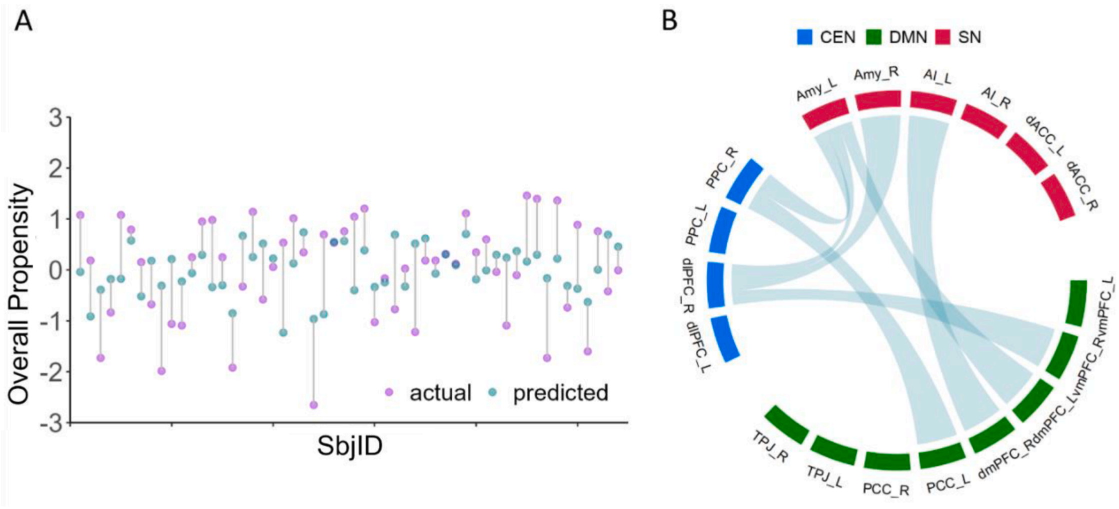


Fig. 5. Summary of results of the *Overall* predictive model. (A) The performance of the *Overall* predictive model. (B) Chord plot shows the most influential features (i.e., rs-FC predictors) in predicting the overall third-party intervention propensity. For each iteration of LOOCV, we recorded top 10 features that contributed the most to the predictive model. Then we identified features that appeared in more than 90 % of all iterations and considered them as the most influential features. Abbreviations about brain regions and networks are the same as Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

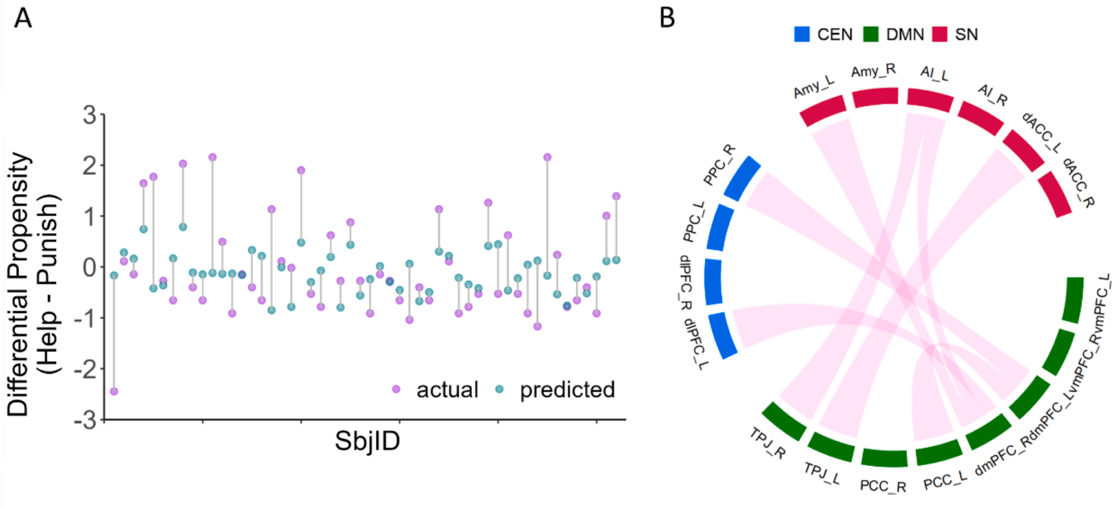


Fig. 6. Summary of results of the *Differential* predictive model. (A) The performance of the *Differential* predictive model. (B) Chord plot shows the most influential features (i.e., rs-FC predictors) in predicting the differential third-party intervention propensity. For each iteration of LOOCV, we recorded top-10 features that contributed the most to the predictive model. Then we identified features that appeared in more than 90 % of the iterations and considered them as the most influential features. Abbreviations about brain regions and networks are the same as Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

networks in characterizing the heterogeneity of third-party intervention behaviors.

Importantly, from the network perspective, our results are consistent with recent studies combining rs-fMRI and the connectome-based predictive modeling approach. Using comprehensive graph-based network analyses, Li et al. (2022b) found that the topological organization of nodes in all three networks were predictive of punishment rates across individuals. In this study, the graph-based predictive model was trained based on the Ultimatum Game that essentially assessed the second-party punishment, and was tested using an independent sample in which the TPP behaviors were measured. The consistency across networks indicated a shared neural network underlying both types of punishment. A subsequent study focusing on TPP also revealed the importance of the rs-FC patterns of CEN and other networks (such as SN) in predicting TPP across individuals (Yang et al., 2021). Our IS-RSA results further suggest that the spontaneous patterns within the TPP brain network could explain individual differences in a more generalized intervention preference, which was characterized by the Euclidean differences in the intervention propensity for help and punishment in a 2-dimensional space between pairs of participants.

We further examined how rs-FC patterns differentially reflect individual variations in the helping and punishment behaviors, which have been largely neglected in previous research. In particular, we identified a stronger association between rs-FC patterns of the right dlPFC and individual differences in TPH compared to TPP. While dlPFC was identified in both punishment and helping-related decisions in studies as mentioned earlier (Hu et al., 2018; Krueger and Hoffman, 2016), a recent study using functional near-infrared spectroscopy (fNIRS) concurred with our finding and showed an increased signal in dlPFC in TPH choices compared to TPP choices, suggesting that engaging in intervention through helping may involve higher cognitive and affective demands on evaluative processing (Xie et al., 2022). Our study expands on these findings by showing that the intrinsic FC patterns of dlPFC may serve as a neural fingerprint for characterizing individual variations in TPH.

Intriguingly, we observed dissociative patterns concerning different nodes within DMN in reflecting TPH and TPP across individuals. Specifically, the rs-FC pattern of the right TPJ was more associated with individual-level helping (vs. punishment) propensity, whereas the opposite pattern was found in dmPFC. Both regions have been implicated in representing intentions during social interaction (Hampton et al., 2008; Hill et al., 2017), and are thus considered key nodes in the mentalizing network (Frith and Frith, 2006; Schurz et al., 2014). However, the right TPJ is additionally related to prosociality (such as generosity, advantageous inequality aversion) and context-dependent morality. Studies have shown that the gray matter volume in the right TPJ is positively correlated with individuals' altruistic propensities indexed by advantageous inequality aversion. Moreover, this region plays a critical role in computations regarding others' interests during decision-making (Hutcherson et al., 2015; Nicolle et al., 2012; Strombach et al., 2015). In the third-party context, compensators (i.e., individuals who prefer helping over punishment), compared to punishers (i.e., individuals who prefer punishment over helping) have been shown to be more engaged in the right TPJ when deciding whether to intervene in unfair situations (Civai et al., 2019). Overall, our findings support the distinct roles of the right TPJ and dmPFC in third-party intervention, and, more broadly, social cognition, by highlighting their contributions to capturing heterogeneity in TPH and TPP respectively.

To further explore whether those nodes identified by IS-RSA would predict the intervention propensity across individuals, we conducted post-hoc predictive regression analyses. These analyses revealed that the intrinsic connectivity between nodes within or between these networks, knowns as features, indeed contributes to predicting individual-level the overall intervention propensity or the differential intervention propensity. These findings reveal a direct connection between intrinsic neural patterns and third-party intervention propensity across individuals, which, to some degree, corroborate and complement the IS-RSA results identifying the relationship between the inter-individual similarity in intrinsic neural patterns and their behavioral propensities.

Despite the strengths of our approach, there are several limitations that should be noted. First, while the present task offers obvious advantages in clearly measuring the propensity for each type of intervention, it suffers from the disadvantages of lower ecological validity compared to real-life transgression scenarios in which bystanders typically have both intervention options available at the same time (Dhaliwal et al., 2021; Hu et al., 2020, 2015; Leliveld et al., 2012). This limitation to some extent reduces the generalizability of the present findings. Second, the intensity of injustice, elicited via monetary unequal splits between two strangers, can be subject to debate, particularly when the intention of the transgressors is not considered. Last but not the least, given the relatively modest effect size observed in the present IS-RSA results, future studies may consider recruiting larger samples to assess the robustness and replicability of the current findings.

In conclusion, the present study offers novel evidence regarding the neurobiological substrates underpinning the large heterogeneity in third-party interventions. By adopting the cutting-edge IR-RSA approach from an intrinsic brain network perspective, we have gained insights into the common and distinct roles of brain networks (and key nodes) at rest in accounting for individual variations in justice-restoring behaviors. More broadly, our study showcases the potential of integrating multivariate analyses with task-free neural data in elucidating the underlying mechanisms concerning why people vary in their social behaviors.

5. Data and code availability statement

The pre-processed rs-fMRI time courses, behavioral performance, and scripts visualizing main results are available via OSF (https://osf.io/s8dvg/).

CRediT authorship contribution statement

Yancheng Tang: Data curation, Formal analysis, Investigation, Software, Writing – original draft, Writing – review & editing. **Yang Hu:** Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Jie Zhuang:** Data curation, Investigation, Project administration. **Chunliang Feng:** Formal analysis, Writing – review & editing. **Xiaolin Zhou:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Data availability

The pre-processed rs-fMRI time courses, behavioral performance, and scripts visualizing main results are available via OSF (https://osf.io/s8dvg/).

Acknowledgments

This work was supported by research grants from National Natural Science Foundation of China awarded to Y.H. (32200853), X.Z. (71942001), and C.F. (32271126), and the STI 2023 – Major Projects (2021ZD0200500) awarded to X.Z. and Y.H. Y.H. was also supported by National Science Foundation of Shanghai (23ZR1418400), and the Fundamental Research Funds for the Central Universities (2022ECNU-XWK-XK003). We thank Jiayi Zhao for helpful assistance with brain imaging data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120468.

References

Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S.B., Krueger, F., 2018. The role of the anterior insula in social norm compliance and enforcement: evidence from coordinate-based and functional connectivity meta-analyses. Neurosci. Biobehav. Rev. 92, 378–389.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34 (4), 537–541.

Brainard, D.H., Vision, S., 1997. The psychophysics toolbox. Spat. Vis. 10 (4), 433–436.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60 (5), 930–940.

Buckholtz, J.W., Marois, R., 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat. Neurosci. 15 (5), 655–661.

Buckholtz, J.W., Martin, J.W., Treadway, M.T., Jan, K., Zald, D.H., Jones, O., Marois, R., 2015. From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron 87 (6), 1369–1380.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10 (3), 186–198.

Chen, P.A., Jolly, E., Cheong, J.H., Chang, L.J., 2020. Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. Neuroimage 216, 116851.

Civai, C., Huijsmans, I., Sanfey, A., 2019. Neurocognitive mechanisms of reactions to second- and third-party justice violations. Sci. Rep. 9 (9271), 1–11.

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S.S., Yan, C., Li, Q., Lurie, D., Vogelstein, J., Burns, R., 2013. Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (c-pac). Front. Neuroinform. 42, 10–3389.

Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Dayan, P., Dolan, R.J., 2017. Moral transgressions corrupt neural representations of value. Nat. Neurosci. 20 (6), 879–885.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., Greene, J.D., 2012. Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. Soc. Cogn. Affect. Neurosci. 7 (8), 888–895.

Dhaliwal, N.A., Patil, I., Cushman, F., 2021. Reputational and cooperative benefits of third-party compensation. Organ. Behav. Hum. Decis. Process. 164, 27–51.

Dubois, J., Adolphs, R., 2016. How the brain represents other minds. Proc. Natl. Acad. Sci. 113 (1), 19–21.

Fehr, E., Fischbacher, U., 2004a. Social norms and human cooperation. Trends Cogn. Sci. (Regul. Ed.) 8 (4), 185–190.

Fehr, E., Fischbacher, U., 2004b. Third-party punishment and social norms. Evol. Hum. Behav. 25 (2), 63–87.

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J.J., Phelps, E.A., 2014. Fairness violations elicit greater punishment on behalf of another than for oneself. Nat. Commun. 5, 5306.

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 (2), 591–602.

Feng, C., Zhu, Z., Cui, Z., Ushakov, V., Dreher, J.C., Luo, W., Gu, R., Wu, X., Krueger, F., 2021. Prediction of trust propensity from intrinsic brain morphology and functional connectome. Hum. Brain Mapp. 42 (1), 175–191.

Feng, C., Zhu, Z., Gu, R., Wu, X., Luo, Y.J., Krueger, F., 2018. Resting-state functional connectivity underlying costly punishment: a machine-learning approach. Neuroscience 385, 25–37.

Finn, E.S., Corlett, P.R., Chen, G., Bandettini, P.A., Constable, R.T., 2018. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. Nat. Commun. 9 (1), 2043.

Finn, E.S., Glerean, E., Khojandi, A.Y., Nielson, D., Molfese, P.J., Handwerker, D.A., Bandettini, P.A., 2020. Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. Neuroimage 215, 116828.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. 8 (9), 700–711.

Frith, C.D., Frith, U., 2006. The neural basis of mentalizing. Neuron 50 (4), 531–534.

Ginther, M.R., Bonnie, R.J., Hoffman, M.B., Shen, F.X., Simons, K.W., Jones, O.D., Marois, R., 2016. Parsing the behavioral and brain mechanisms of third-party punishment. J. Neurosci. 36 (36), 9420–9434.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. 100 (1), 253–258.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P, 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. 105 (18), 6741–6746.

He, Y., Evans, A., 2010. Graph theoretical modeling of brain connectivity. Curr. Opin. Neurol. 23 (4), 341–350.

Hill, C.A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J.P., Ruff, C.C., 2017. A causal account of the brain network computations underlying strategic social behavior. Nat. Neurosci. 20 (8), 1142.

Hu, J., Li, Y., Yin, Y., Blue, P.R., Yu, H., Zhou, X., 2018. How do self-interest and other-need interact in the brain to determine altruistic behavior? Neuroimage 157, 598–611.

Hu, Y., Fiedler, S., Weber, B., 2020. What drives the (un)empathic bystander to intervene? Insights from eye tracking. Br. J. Soc. Psychol. 59 (3), 733–751.

Hu, Y., Hu, C., Derrington, E., Corgnet, B., Qu, C., Dreher, J.C., 2021. Neural basis of corruption in power-holders. eLife 10, e63922.

Hu, Y., Strang, S., Weber, B., 2015. Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. Front. Behav. Neurosci. 9, 24.

Hutcherson, C., Bushong, B., Rangel, A., 2015. A neurocomputational model of altruistic choice and its implications. Neuron 87 (2), 451–462.

Iyer, S., Collier, E., Finn, E.S., & Meyer, M.L. (2023). Negative affect homogenizes and positive affect diversifies social memory consolidation across people *bioRxiv*.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - an S4 package for kernel methods in R. J. Stat. Softw. 11 (9), 1–20.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. (Regul. Ed.) 17 (8), 401–412.

Krueger, F., Hoffman, M., 2016. The emerging neuroscience of third-party punishment. Trends Neurosci. 39 (8), 499–501.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., Bischl, B., 2019. mlr3: a modern object-oriented machine learning framework in R. J. Open Source Softw. 4 (44), 1903.

Lee, M.H., Smyser, C.D., Shimony, J.S., 2013. Resting-state fMRI: a review of methods and clinical applications. Am. J. Neuroradiol. 34 (10), 1866–1872.

Leliveld, M.C., Dijk, E., Beest, I., 2012. Punishing and compensating others at your own expense: the role of empathic concern on reactions to distributive injustice. Eur. J. Soc. Psychol. 42 (2), 135–140.

Li, T., Pei, Z., Zhu, Z., Wu, X., Feng, C., 2022a. Intrinsic brain activity patterns across large-scale networks predict reciprocity propensity. Hum. Brain Mapp. 43 (18), 5616–5629.

Li, T., Yang, Y., Krueger, F., Feng, C., Wang, J., 2022b. Static and dynamic topological organizations of the costly punishment network predict individual differences in punishment propensity. Cereb. Cortex 32 (18), 4012–4024.

Li, Z., Dong, Q., Hu, B., Wu, H., 2023. Every individual makes a difference: a trinity derived from linking individual brain morphometry, connectivity and mentalising ability. Hum. Brain Mapp. 44 (8), 3343–3358.

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., Fetchenhauer, D., 2011. Individual differences in third-party interventions: how justice sensitivity shapes altruistic punishment. Negot. Conflict Manag. Res. 4 (4), 297–313.

Lu, X., Li, T., Xia, Z., Zhu, R., Wang, L., Luo, Y.J., Feng, C., Krueger, F., 2019. Connectome-based model predicts individual differences in propensity to trust. Hum. Brain Mapp. 40 (6), 1942–1954.

McAuliffe, K., Dunham, Y., 2021. Children favor punishment over restoration. Dev. Sci. 24 (5), e13093.

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), 1114–1121.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 (4), e1003553.

Pelli, D.G., Vision, S., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10, 437–442.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59 (3), 2142–2154.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320–341.

Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112, 267–277.

R Core Team. (2014). R: a language and environment for statistical computing.

Raihani, N.J., Bshary, R., 2015. Third-party punishers are rewarded, but third-party helpers even more so. Evolution (N Y) 69 (4), 993–1003.

Sabbagh, C., Schmitt, M., 2016. Handbook of Social Justice Theory and Research. Springer.

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 300 (5626), 1755–1758.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34.

Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc. 12 (3), 506–518.

Skarlicki, D.P., O'Reilly, J., Kulik, C.T, 2015. The third-party perspective of (in) justice. In: Cropanzano, R., Ambrose, M. (Eds.), Oxford Handbook of Justice in Work Organizations. Oxford University Press, pp. 235–255.

Snyder, W., Uddin, L.Q., Nomi, J.S., 2021. Dynamic functional connectivity profile of the salience network across the life span. Hum. Brain Mapp. 42 (14), 4740–4749.

Song, K.R., Potenza, M.N., Fang, X.Y., Gong, G.L., Yao, Y.W., Wang, Z.L., Liu, L., Ma, S.S., Xia, C.C., Lan, J., Deng, L.Y., Wu, L.L., Zhang, J.T., 2021. Resting-state connectome-based support-vector-machine predictive modeling of internet gaming disorder. Addict. Biol. 26 (4), e12969.

Speer, S.P.H., Smidts, A., Boksem, M.A.S., 2022. Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest. Neuroimage 246, 118761.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C.K., Sanfey, A.G., 2018. Neurobiological mechanisms of responding to injustice. J. Neurosci. 38 (12), 2944–2954.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I.I., Tobler, P.N., Kalenscher, T., 2015. Social discounting involves modulation of neural value signals by temporoparietal junction. Proc. Natl. Acad. Sci. 112 (5), 1619–1624.

Treadway, M.T., Buckholtz, J.W., Martin, J.W., Jan, K., Asplund, C.L., Ginther, M.R., Jones, O.D., Marois, R., 2014. Corticolimbic gating of emotion-driven punishment. Nat. Neurosci. 17 (9), 1270–1275.

van Baar, J.M., Chang, L.J., Sanfey, A.G., 2019. The computational and neural substrates of moral strategies in social decision-making. Nat. Commun. 10 (1), 1483.

Van Doorn, J., Brouwers, L., 2017. Third-party responses to injustice: a review on the preference for compensation. Crime Psychol. Rev. 3 (1), 59–77.

van Doorn, J., Zeelenberg, M., Breugelmans, S.M., 2018. An exploration of third parties' preference for compensation over punishment: six experimental demonstrations. Theory Decis. 85 (3), 333–351.

Wang, H., Zhen, Z., Zhu, R., Yu, B., Qin, S., Liu, C., 2022a. Help or punishment: acute stress moderates basal testosterone's association with prosocial behavior. Stress 25 (1), 179–188.

Wang, R., Yu, R., Tian, Y., Wu, H., 2022b. Individual variation in the neurophysiological representation of negative emotions in virtual reality is shaped by sociability. Neuroimage 263, 119596.

Wu, M., Kujawa, A., Lu, L.H., Fitzgerald, D.A., Klumpp, H., Fitzgerald, K.D., Monk, C.S., Phan, K.L., 2016. Age-related changes in amygdala–frontal connectivity during emotional face processing from childhood into young adulthood. Hum. Brain Mapp. 37 (5), 1684–1695.

Wu, X., Zhang, L., Liu, B., Liao, J., Qiu, Y., Huang, R., 2023. Social navigation modulates the anterior and posterior hippocampal circuits in the resting brain. Brain Struct. Funct. 228 (3–4), 799–813.

Xie, E., Liu, M., Liu, J., Gao, X., Li, X., 2022. Neural mechanisms of the mood effects on third-party responses to injustice after unfair experiences. Hum. Brain Mapp. 43 (12), 3646–3661.

Yang, Q., Bellucci, G., Hoffman, M., Hsu, K.T., Lu, B., Deshpande, G., Krueger, F., 2021. Intrinsic functional connectivity of the frontoparietal network predicts inter-individual differences in the propensity for costly third-party punishment *Cognitive*. Psychobiology (Austin, Tex.) 21 (6), 1222–1232.

Zhong, S., Chark, R., Hsu, M., Chew, S.H., 2016. Computational substrates of social norm enforcement by unaffected third parties. Neuroimage 129, 95–104.