Title

The Emerging Neuroscience of Social Punishment: Meta-Analytic Evidence

Abbreviated title

Meta-analytic evidence of the social punishment

Author names and affiliations

Gabriele Bellucci [1*], Julia A. Camilleri [2,3], Vijeth Iyengar [4], Simon B. Eickhoff [2,3], Frank Krueger [5,6*]

[1] Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2] Institute for Neuroscience and Medicine (INM-7), Research Center Jülich, Germany
[3] Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Germany
[4] Administration for Community Living/Administration on Aging, U.S. Department of Health and Human Services, Washington DC, USA
[5] School of Systems Biology, George Mason University, Fairfax, VA, USA
[6] Department of Psychology, George Mason University, Fairfax, VA, USA

***Corresponding authors**

Gabriele Bellucci & Frank Krueger

GBellucc@gmail.com & FKrueger@gmu.edu

Manuscript Information

Number of pages: *26*

Number of figures/tables: *9/1*

Number of words for Abstract/Introduction/Discussion: *151/1,058/2,063*

Key words: social punishment, second-party punishment, third-party punishment, activation likelihood estimation, meta-analytic connectivity mapping, resting-state functional connectivity

Abstract

Social punishment (SOP)—third-party punishment (TPP) and second-party punishment (SPP)—sanctions norm-deviant behavior. The hierarchical punishment model (HPM) posits that TPP is an extension of SPP and both recruit common processes engaging large-scale domain-general brain networks. Here, we provided meta-analytic evidence to the HPM by combining the activation likelihood estimation approach with connectivity analyses and hierarchical clustering analyses. Although both forms of SOP engaged the dorsolateral prefrontal cortex and bilateral anterior insula (AI), a functional differentiation also emerged with TPP preferentially engaging social cognitive regions (temporoparietal junction) and SPP affective regions (AI). Further, although both TPP and SPP recruit domain-general networks (salience, default-mode, and central-executive networks), some specificity in network organization was observed. By revealing differences and commonalities of the neural networks consistently activated by different types of SOP, our findings contribute to a better understanding of the neuropsychological mechanisms of social punishment behavior—one of the most peculiar human behaviors.

Introduction

In a highly complex social world, norms are necessary to govern and organize the multifaceted dynamics of interpersonal interactions (Bicchieri, 2014). However, establishing a set of norms is not sufficient to guarantee everybody's compliance (Fehr and Fischbacher, 2004a, b). As opposed to self-punishment, social punishment (SOP) sanctions deviant behavior that violates the group's social norms. Individuals punish transgressors to enforce these social norms even when punishment is costly (Dawes et al., 2007; Seymour et al., 2007). SOP takes two essential forms: second-party punishment (SPP) and third-party punishment (TPP) (Fehr and Fischbacher, 2003; Fehr and Gächter, 2002). Both require that the norm-enforcer recognizes the intention of the offender and the harm inflicted onto the victim. These evaluations are then integrated into estimations of the transgressor's blameworthiness to assign the appropriate punishment (Buckholtz and Marois, 2012). However, TPP and SPP differ in the target of the wrongdoing. In SPP, victim and punisher are the same person, while in TPP the victim is another person than the impartial, third-party judge. Previous work has suggested this difference is reflected in the different neuropsychological processes engaged by SPP and TPP.

Psychophysiological evidence suggests the emotional reaction of the victim/punisher to the inflicted harm is an essential component of SPP. For instance, punishing others for their unfairness increases skin conductance response (SCR, a measure of emotional activation), and higher emotional states are reported during punishment of unfair behaviors (van 't Wout et al., 2006). Interestingly, SCR increases only when the punisher is also the target of the unfair act but not when the unfair act affects someone else (Civai et al., 2010). These results concur with neuroimaging research pointing to neural activations during punishment of unfairness in the anterior insula (AI) (Sanfey et al., 2003)—a brain region associated with aversive experiences (Craig, 2002; Damasio et al., 2000). Crucially, activity in the AI is linearly related to punishment of

unfair behaviors, suggesting that this region plays an essential role in SPP (Tabibnia et al., 2008).

On the contrary, the intentions of a wrongdoer are central to TPP to determine the transgressor's responsibility for the appropriate punishment. Thus, impartial third-party judges punish intentional unfair behaviors more than unintentional unfair behaviors (Blount, 1995; Falk et al., 2008) and rely on mentalizing regions during their punishment decisions, such as the posterior temporoparietal junction (pTPJ)—a region involved in inferences on others' intentions (Igelstrom and Graziano, 2017; Igelström et al., 2015; Saxe and Kanwisher, 2003; Saxe and Powell, 2006). For instance, the pTPJ plays an important role in third-party decisions to punish in- and outgroup members (Baumgartner et al., 2012; Baumgartner et al., 2014). Furthermore, impartial third-party judges recruit the lateral prefrontal cortex (LPFC) when they assess responsibility in norm violations (Zhong et al., 2016) or need to distinguish between contextual situations on the basis of criminal responsibility (Buckholtz et al., 2008). This evidence suggests that the LPFC is involved in the decision on the punishment that best aligns with the transgressor's blameworthiness.

However, this evidence is at odds with other findings. For instance, second-party punishers also evaluate the transgressor's responsibility before a punishment decision and third-party punishers respond emotionally to a transgression as well (Civai, 2013; de Quervain et al., 2004; Egas and Riedl, 2008; Fehr and Gächter, 2002). Moreover, AI activations have been observed for norm violations in both SPP and TPP irrespective of the target of the violation (Civai et al., 2012; Corradi-Dell'Acqua et al., 2013) and even for one's own wrongdoing (Güroğlu et al., 2010). These findings indicate that the AI rather signals the violation of a norm —a computation required for the determination of the proper penalty in both SPP and TPP. Similarly, the LPFC has been linked to enforcement of social norm compliance in SPP (Ruff et al., 2013) and disruption of the

dorsal LPFC of second-party punishers reduces punishment of unfairness (Knoch et al., 2006). This evidence indicates that the LPFC takes a role in the implementation of norm-enforcing behaviors —a process required in both forms of SOP.

An alternative view on the processes underlying punishment behaviors has been put forward by the *hierarchical punishment model* (HPM) (Krueger and Hoffman, 2016). The HPM posits that SPP draws on a rudimental form of first-party punishment (i.e., conscience and guilt), while TPP emerges as an extension of SPP, allowing for a generalized norm-enforcing behavior in genetically heterogeneous societies. Accordingly, TPP is supposed to piggyback on a set of processes already engaged by SPP, and SPP relies on core processes already engaged by first-party punishment. Previous studies have provided preliminary evidence on neural commonalities underlying different forms of social punishment (Stallen et al., 2018; Zinchenko, 2019). The difference between TPP and SPP does not rely in their different cognitive processes but in how these processes are engaged. SPP places more weight on the harm of a norm violation engaging affective processes to signal its aversive and threatening nature. On the contrary, TPP relies more on the intentions behind a norm violation requiring perspective-taking abilities to mentally represent internal states and external circumstances of a self-unrelated situation. This hypothesis is consistent with phylogenetic and ontogenetic evidence that TPP is rare or non-existent in non-human primates and small-scale societies (Guala, 2012; Riedl et al., 2012) and emerges in humans after age six when mentalizing abilities are fully developed (Frith and Frith, 2003; McAuliffe et al., 2015; Mendes et al., 2018).

On the neural level, the HPM proposes three domain-general large-scale networks as underlying SOP. The salience network (involving the AI and anterior cingulate cortex, ACC) signals the norm violation and weights the severity of the inflicted harm. The default-mode network (including the medial PFC and pTPJ) evaluates the perpetrator's intentions and integrates harm and intent for assessment of blameworthiness. Finally, the

central-executive network (anchored in the LPFC) converts blameworthiness evaluations into a punishment decision.

In this study, we investigated whether TPP and SPP engage different brain mechanisms associated with putatively different cognitive processes. First, we identified the meta-analytic brain regions consistently activated by SPP and TPP, implementing the activation likelihood estimation (ALE) method (Eickhoff et al., 2009). Second, we determined the consensus connectivity networks of the emerging meta-analytic brain regions underlying SPP and TPP and their sub-network compositions, employing connectivity analyses (i.e., task-based meta-analytic connectivity mapping, MACM, and task-free resting-state functional connectivity, RSFC) and hierarchical analysis (Eickhoff et al., 2018; Goodkind et al., 2015; Hardwick et al., 2015; Kolling et al., 2016; Wang et al., 2015). These analyses allowed us to first delineate the connectivity profiles of TPP and SPP brain regions, their overlap and specificity, and then their functional roles with the help of functional decoding analyses (Genon et al., 2018).

Materials and Methods

Literature search and selection

A systematic online database search was performed on PubMed and Google Scholar by entering various combinations of relevant search items referring to punishment behaviors (up to the December 2, 2018). The following keywords were used for the search: 'altruistic punishment', 'third-party punishment', 'second-party punishment', 'punishment', 'modified Ultimatum Game' and 'modified Dictator Game', in combination with 'fMRI', 'magnetic resonance imaging', and 'neuroimaging', 'PET', 'positron emission tomography'. Note that economic games and vignettes were included for the TPP studies. In economic game studies, participants in general face an unfair monetary distribution executed by another player (the transgressor) that they can punish by spending some of their endowment to diminish the transgressor's payoffs (i.e., costly punishment). On the contrary, vignettes studies present in general participants with descriptions of various legal/moral transgressions and participants must decide how much to punish the transgressor (i.e., hypothetical punishment). To be able to capture neural activity specifically related to the different forms of punishment, this meta-analysis included only contrasts during the decision phase that singled out the neural underpinnings of a third-party/second-party decisions (e.g., punishment vs. no punishment/baseline condition, TPP vs. SPP or SPP vs. TPP, reject vs. accept in economic games) as opposed to other types of decisions as well as correlations of neural signal with the amount of punishment/rejection rates (see *Supplemental Material. List of Studies*).

In addition, several other sources were explored, including (a) the BrainMap database (http://brainmap.org), (b) work cited in review papers, and (c) direct searches on the names of frequently occurring authors. The searched studies were further assessed according to the following criteria: (i) participants were free from psychiatric or neurological diagnoses; (ii) participants were adults; (iii) no pharmacological modulations were reported; (iv) fMRI was

used as the imaging modality (no PET studies were found under the searched terms); (v) whole-brain analyses were applied (excluding region of interest [ROI] analyses) to reduce the inclusion of false positives; (vi) fMRI results were derived from a general linear model based on either a binary contrast or parametric analyses; and (vii) activations were presented in a standardized stereotaxic space (Talairach or Montreal Neurological Institute, MNI). Note that for studies reporting Talairach coordinates, a conversion to the MNI coordinates was implemented in the GingerALE software (https://www.brainmap.org/ale/) with the Brett's algorithm.

Activation likelihood estimation (ALE) method

The ALE meta-analysis follows previous work published by our group (Bellucci et al., 2017b; Bellucci et al., 2018). The ALE algorithm (using in-house MATLAB scripts) was employed to investigate the coordinate-based, consistent, meta-analytic activations across studies examining neural responses associated with SOP decisions (Eickhoff et al., 2012; Eickhoff et al., 2009; Eickhoff et al., 2016). ALE determines the convergence of foci reported from different functional (e.g., blood-oxygen-level dependent [BOLD] contrast imaging) neuroimaging studies with published foci in Talairach or MNI space (Laird et al., 2005; Turkeltaub et al., 2002). Reported foci are interpreted as spatial probability distributions in the ALE framework. Their widths refer to the empirical estimates of the spatial uncertainty based on between-subjects and between-templates variability of the neuroimaging data (Eickhoff et al., 2009). To weight the between-subject variability, the number of subjects analyzed in studies is considered by the ALE algorithm. The assumption is that more reliable approximation to the 'true' activation are given by larger sample sizes. Thus, these samples are modelled with smaller Gaussian distributions (Eickhoff et al., 2009).

An ALE map across studies is obtained by calculating the union of the individual modulated activation maps created from the maximum probability associated with any one

focus (always the closest one) for each voxel (Turkeltaub et al., 2012). This ALE map is determined against a null-distribution of random spatial association between studies employing a non-linear histogram integration algorithm (Eickhoff et al., 2012; Turkeltaub et al., 2012). Results were thresholded for significance using a cluster-level family-wise error (FWE) correction at $P < 0.05$ with a cluster defining threshold of $P < 0.001$ and 10,000 permutations (Eickhoff et al., 2012; Eklund et al., 2016). Moreover, to meet criteria of robust unbiased results, clusters were only considered significant if the most dominant experiment (MDE) contributed to the significant cluster on average less than 50% and the two MDEs (2MDEs) contributed on average less than 80% (Bellucci et al., 2017b; Bellucci et al., 2018; Eickhoff et al., 2016). For experiments' contributions, the fraction of the ALE value accounted for by each experiment contributing to the cluster was computed. This average non-linear contribution of each experiment to the ALE value was computed from the ratio of the ALE values at the location of the cluster with and without the experiment in question (Eickhoff et al., 2016).

A total of 47 experiments (see *Supplemental Material. List of Studies*) examining SOP with a total of 312 foci across 1,188 subjects were identified, including a total of 22 experiments for SPP (139 foci, 598 subjects, average of 27.2 subjects per experiment) and a total of 25 experiments for TPP (173 foci, 590 subjects, average of 23.6 subjects per experiment). Among these, eight experiments were from articles investigating both SPP and TPP in the same sample. Three main effect analyses for SOP (the pooled analysis of SPP and TPP), SPP and TPP, and two contrast analyses for SPP > TPP and TPP > SPP were performed.

Task-based, meta-analytical connectivity modeling (MACM) analysis

To investigate the meta-analytic co-activation profiles of punishment decisions, MACM analyses were conducted using the peak coordinates of each significant brain region identified

from the three previous ALE analyses for SOP, SPP and TPP as seed regions (i.e., sphere radius = 5 mm as in previous studies) (Camilleri et al., 2018; Langner et al., 2014).

The BrainMap database (http://www.brainmap.org/) was used (Laird et al., 2009a), which at the time of assessment contained coordinates of reported activation foci and associated meta-data of approximately 14,720 neuroimaging experiments pertaining to "normal mapping" analyses. For SOP, the MACM analyses were based on the following experiments, foci, and number of subjects for each of the following seed regions: left AI (618 experiments | 9,282 foci | 9,318 subjects), right AI (547 | 8,053 | 8,220), and left dorsolateral PFC (DLPFC; 116 | 1,655 | 1,818). The MACM analyses for SPP were based on the left AI (557 | 8,422 | 8,465) and the right AI (510 | 7,715 | 7,731) as seed regions, whereas for TPP on the left pTPJ (98 | 1,400 | 1,506) and the left ventrolateral PFC (VLPFC; 182 | 2,310 | 2,747) as seed regions.

In brief, whole-brain peak coordinates of all those studies from BrainMap that reported at least one focus of activation within the respective ROIs were downloaded. Next, coordinates were analyzed with the ALE algorithm to detect areas of convergence of co-activations with those seeds. Finally, the ALE maps were thresholded at $P < 0.05$ cluster-level corrected (cluster-forming threshold: $P < 0.001$ at voxel-level) and converted into *z*-scores for display (Bellucci et al., 2018; Camilleri et al., 2018).

Task-free, resting-state functional connectivity (RSFC) analysis

To investigate the FC profiles of punishment decisions, RSFC analyses were conducted using the peak coordinates of each significant brain region identified from the three previous ALE analyses for SOP, SPP and TPP as seed regions (i.e., sphere radius = 5 mm as in previous studies). RS-fMRI images of 192 healthy volunteers were obtained from the Enhanced Nathan Kline Institute – Rockland Sample (Nooner et al., 2012). Images were acquired on a Siemens TimTrio 3T scanner using BOLD contrast [gradient-echo EPI pulse sequence, TR = 1.4 s, TE

= 30 ms, flip angle = 65, voxel size = 2.0 mm × 2.0 mm × 2.0 mm, 64 slices]. Physiological and movement artifacts were removed from the resting-state data by using FIX (FMRIB's ICA-based Xnoiseifier, version 1.061 as implemented in FSL 5.0.9) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) and data were further preprocessed using SPM8 (Wellcome Trust Centre for Neuroimaging, London) and in-house MATLAB scripts, following previously employed processing procedures (Camilleri et al., 2018; Satterthwaite et al., 2013).

The processed time-course of each seed (sphere radius = 5 mm) was then correlated with the (identically processed) time-series of all other gray-matter voxels in the brain using linear (i.e., Pearson) correlation. The resulting correlation coefficients were transformed into Fisher's *z*-scores, which were entered in a second-level ANOVA for group analysis including age and gender as covariates of no interest. The data was then subjected to non-parametric permutation based inference and thresholded at $P < 0.05$ corrected for multiple comparisons on the cluster level.

Consensus connectivity map

After having identified brain areas showing task-based co-activation (i.e., MACM) and task-free FC (i.e., RSFC) FC our seed regions, conjunction analyses were performed across the MACM and RSFC maps for each seed using the minimum statistic (Nichols et al., 2005). This resulted in three consensus connectivity maps (i.e., SOP, SPP, TPP) that yielded brain regions consistently interacting with each seed across different brain states (Clos et al., 2014; Hardwick et al., 2015). An extent-threshold of 10 continuous voxels was applied to exclude smaller regions of putatively spurious overlaps. The decision to use this exact number of voxels was indeed arbitrary but reflects standard procedures used in previous work (Camilleri et al., 2018).

Hierarchical cluster analysis of SPP and TPP regions

To identify potential cliques among the networks of each brain region for SPP and TPP, hierarchical cluster analyses were performed using their RSFC patterns (Camilleri et al., 2018). Using the FSLNets toolbox (*http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets*), RSFC between all regions of the identified SPP and TPP networks was computed. Partial temporal correlations between all regions' time series data were computed to estimate pairwise functional connectivity (Marrelec et al., 2006). For each pairwise connection, Fisher's *z*–transformed functional connectivity values were submitted to one-sample *t*-tests. The resulting *t* values, reflecting connection strength as well as consistency across the sample, were z-transformed (into units of the standard normal distribution). This connectivity matrix was then fed into the WARD clustering.

Of note, all features entered the analyses without any thresholding for significance, which is a distinction from the analyses described above but necessary to preserve the full pattern of the respective connectional and functional profiles. The concept behind hierarchical clustering is to group the initial elements (regions) in a stepwise fashion such that elements within a cluster have features that are as homogeneous as possible while different clusters are maximally distinct from each other. This was achieved through an agglomerative approach in which clusters initially formed by individual regions that are subsequently merged according to their similarity using standardized Euclidean distances and Ward's incremental sum of squares method (Eickhoff et al., 2011; Timm, 2002). This hierarchical approach then revealed cliques of SPP and TPP regions at different levels of granularity. Average Euclidean distances (EDs) and standard deviations (SDs) were provided for each hierarchical cluster.

Functional characterization of punishment brain regions

The functional profiles of the observed meta-analytic brain regions were characterized based on the behavioral domain (BD), which describes the categories (i.e., action, cognition, emotion, interception, and perception) and subcategories (e.g., reward, language, social

cognition, pain, among others) of the mental operations likely isolated by the experiments in the BrainMap database (Fox and Lancaster, 2002; Fox et al., 2014; Laird et al., 2009b). This functional characterization was based on forward inference with the aim to identify all experiments that engage a particular region of interest, and then analyze the experimental meta-data describing the experimental settings employed in these experiments (Müller et al., 2013; Rottschy et al., 2012). This allows statistical inference on the type of tasks that evoke activations in a brain region.

In this forward inference approach, the functional profile was determined by identifying BD for which the probability of finding activation in the respective region/set of regions was significantly higher than the overall (a priori) chance across the entire database. That is, it was tested whether the conditional probability of activation given a particular BD, i.e., P(Activation|Domain), was higher than the baseline probability of activating the region(s) in question *per se*, i.e., P(Activation). Significance was established using a binomial test using the standard $\alpha = 0.05$, corrected for multiple comparisons using false discovery rate (FDR).

Results

ALE analysis

A total of 47 experiments were identified, 22 for SPP and 25 for TPP. We first investigated the consistent neural patterns activated for SOP (i.e., pooled analysis across SPP and TPP). The analysis revealed three regions: left AI (17 contributing experiments; i.e., 36.2% of the total experiments, MDE=10.34%, 2MDEs=20.17%), right AI (13 contributing experiments; i.e., 27.7% of the total experiments, MDE=14.87%, 2MDEs=28.48%), and left DLPFC (6 contributing experiments; i.e., 12.8% of the total experiments, MDE=32.36%, 2MDEs=54.21%) (**Tab. 1 & Fig. 1A**). Examining the experiments contributing to the activation in the left and right AI as well as left DLPFC, 8 out of 18 (44.4%), 5 out of 13 (38.5%), and 4 out of 6 (66.7%) experiments investigated TPP, respectively (**Tab. S1**). These results indicated a certain degree of meta-analytic convergence of both SPP and TPP studies on common brain regions in the insular and lateral prefrontal cortices. Next, separate meta-analytic analyses were run to identify the specific meta-analytic clusters for SPP and TPP.

.....

Insert Figure 1 & Table 1 about here

.....

The single meta-analysis for SPP revealed consistent activations in bilateral AI (left: 10 experiments, i.e., 45.5% of total experiments, MDE=16.38%, 2MDEs=32.49%; right: 8 experiment, i.e., 36.4% of total experiments, MDE=19.06%, 2MDEs=37.29%) (**Tab. 1 & Tab. S2 & Fig. 1B**), while the single meta-analysis for TPP showed consistent activations in the left VLPFC (5 experiments, i.e., 20.0% of total experiments, MDE=31.65%, 2MDEs=60.25%) and pTPJ (5 experiments, i.e., 20.0% of total experiments, MDE=32.36%, 2MDEs=59.33%) (**Tab. 1 & Tab. S2 & Fig. 1C**). Follow-up contrast analyses revealed that the right AI was more strongly activated by SPP than TPP. These single meta-analyses mirror

the results of the previous pooled meta-analysis. In fact, in the latter the bilateral AI revealed a higher proportion of contributions from SPP studies, while the left LPFC from TPP studies. Moreover, a conjunction analysis revealed that the SOP clusters in bilateral AI largely overlap with the insular clusters observed in the SPP analysis (**Fig. 1D**). Hence, both analyses support the hypothesis that punishment decisions rely on common brain regions differently engaged by SPP and TPP.

MACM (task-based co-activation) and RSFC (task-free functional connectivity) analysis

To characterize the functional profile of the observed meta-analytic patterns, we first analyzed their functional connectivity fingerprinting. To this end, we analyzed the connectivity profiles both at rest and across different tasks, as the good match between resting-state connectivity patterns and activation patterns across tasks seems to reveal the underlying functional hierarchy of specific brain regions, which is highly informative of their functional role (Cole et al., 2014; Raichle, 2015; Smith et al., 2009; Tavor et al., 2016). The task-based co-activation (i.e., MACM) analyses revealed similar neural patterns for the bilateral AI (identified as seed regions in both the pooled analysis for SOP and the single meta-analysis for SPP), including lateral frontoparietal brain regions (e.g., DLPFC, inferior parietal lobule, IPL, pTPJ [peak in the supramarginal gyrus], medial prefrontal regions (e.g., middle cingulate cortex , MCC, temporal areas), subcortical brain regions (e.g., striatum) (**Tab. S3 & Fig. S1 & Fig. S2**).

The task-free functional connectivity (i.e., RSFC) analyses demonstrated similar findings with additional connectivity pattern in somatosensory and motor brain regions (**Tab. S4 & Fig. 2 & Fig. 3**). Further, for the left DLPFC, analyses revealed consistent connectivity patterns with MCC, subcortical brain regions (e.g., right AI, thalamus, striatum), and frontoparietal brain regions (e.g., bilateral DLPFC, bilateral frontopolar cortex, bilateral IPL, left superior parietal lobule, SPL, and right angular gyrus).

.....

Insert Figure 2 & Figure 3 about here

.....

For the brain regions identified in the single meta-analysis for TPP, the MACM showed that the pTPJ was functionally coupled not only with dorsomedial prefrontal cortex (DMPFC, superior medial gyrus, BA 10), posterior cingulate cortex and middle temporal gyrus, but also with lateral brain areas of the prefrontal cortex (e.g., DLPFC). The VLPFC showed, in addition, co-activation with the MCC, striatum, and SPL (**Tab. S3 & Fig. S3**). The RSFC analyses revealed similar results with further connections to somatosensory and motor brain regions such as the primary somatosensory cortex (i.e., postcentral gyrus, BA 1), primary motor cortex (BA 4) and supplementary motor cortex (BA 6) (**Tab. S4 & Fig. 4**).

.....

Insert Figure 4 about here

.....

Consensus connectivity maps of MACM and RSFC profiles

Consensus functional connectivity maps of the SOP brain regions from the pooled analysis were determined on the basis of the connectivity profiles emerged from MACM and RSFC analyses. This analysis identified a consensus connectivity network, including clusters in the LFPC (e.g., bilateral DLPFC and frontopolar cortex), the medial frontal regions (e.g., MCC), the parietal cortex (e.g., right angular gyrus and left SPL), and subcortical regions (e.g., left AI and a cluster in the thalamus extending to the striatum) (**Tab. S5 & Fig. 5A**).

Finally, the neural convergence of the consensus connectivity maps separately yielded by the SPP and TPP single meta-analyses was determined with a conjunction analysis. Convergence was observed in a set of brain regions clustered into three main sub-networks: a central-executive network involving bilateral DLPFC, a mentalizing network involving the

pTPJ, temporal cortex and temporal pole, and a salience network involving AI, MCC, and striatum (**Tab. S6 & Fig. 5B**).

.....

Insert Figure 5 about here

.....

Hierarchical cluster analysis of SPP and TPP regions

Hierarchical clustering analyses based on the RSFC profile of the identified meta-analytic brain regions were performed to provide insights into functionally coherent sub-networks or "cliques" underlying punishment behaviors. First, SOP brain regions from the pooled analysis clustered into three main sub-networks (**Fig. 6A**): a salience sub-network (i.e., left AI, MCC, thalamus, caudate; ED = 8.13, SD = 0.96), a frontoparietal sub-network (i.e., right angular gyrus, bilateral inferior frontal gyrus, left SPL; ED = 8.29, SD = 1.05) and a frontal sub-network (i.e., left inferior frontal gyrus, bilateral DLPFC; ED = 9.98, SD = 2.08).

Second, the SPP clusters from the single analysis grouped into four main sub-networks (**Fig. 6B**): a salience sub-network (i.e., bilateral AI, inferior frontal gyrus [pars orbitalis] ; ED = 6.79, SD = 1.42), a subcortical sub-network (i.e., putamen, thalamus, cerebellum; ED = 8.86, SD = 0.91), a latero-medial prefrontal sub-network (i.e., bilateral MCC and middle frontal gyrus; ED = 10.16, SD = 0.64), and a central-executive sub-network (i.e., bilateral SPL, DLPFC; ED = 9.51, SD = 1.76).

Third, the TPP clusters from the single analysis clustered into two sub-networks (**Fig. 6C**): a frontotemporal sub-network (i.e., inferior frontal gyrus, inferior temporal gyrus; ED = 10.96, SD = 1.03) and a frontoparietal sub-network (i.e., IPL, ACC; ED = 11.24, SD = 0). Interestingly, no subcortical sub-networks were recruited by TPP, although previous MACM and RSFC analyses revealed some TPP-related brain areas in subcortical areas such as the thalamus and striatum.

Finally, the clustering profile of the common neural patterns yielded by the conjunction analysis of SPP and TPP revealed four sub-networks (**Fig. 6D**): a salience sub-network (i.e., AI, putamen, and bilateral frontal orbital cortex; ED = 9.51, SD = 1.67), a default-mode sub-network (i.e., bilateral pTPJ and temporal pole; ED = 10.57, SD = 1.43), a lateral frontotemporal sub-network (i.e., bilateral DLPFC and inferior frontal gyrus; ED = 10.52, SD = 1.60) and a smaller mediofrontal sub-network (i.e., MCC and premotor cortex; ED = 7.26, SD = 0).

.....

Insert Figure 6 about here

.....

Functional characterization of punishment brain regions

Finally, the functional profile of the meta-analytic clusters was characterized using forward inference analyses based on the meta-data included in the BrainMap database. The goal was to determine differences and convergences of the functional roles undertaken by the identified meta-analytic clusters. First, analyses of the brain region from the *pooled SOP analysis* revealed that the left AI (**Fig. 7A**) was functionally associated with both cognitive and affective domains involving language, pain and reward (**Fig. 7D**). The left DLPFC (**Fig. 7B**) was associated only with processes of the cognitive domain such as reasoning, working memory and explicit memory (**Fig. 7D**). Finally, the right AI (**Fig. 7C**) was particularly associated with the interoceptive domain involving in particular pain processing (**Fig. 7D**). Comparing the likelihood ratios of the two AI clusters, the left AI was more likely related to cognition, whereas the right AI to affective processing.

.....

Insert Figure 7 about here

.....

Second, analyses of the brain regions from the *single meta-analysis for SPP* demonstrated that the left AI (**Fig. 8A**) was associated with processes of the cognitive and affective domains such as language, reward and pain (**Fig. 8C**), while the right AI (**Fig. 8B**) with processes of the interoceptive and affective domains such as pain, gustation, disgust, and anxiety (**Fig. 8C**), although only pain survived correction for multiple comparisons. Comparing the likelihood ratios of the two AI clusters, an opposite functional pattern was observed for SPP than for the previous SOP analysis: the left AI was here more likely related to both cognition and affective processing than the right AI.

Finally, analyses of the brain regions from the *single meta-analysis for TPP* showed that the left pTPJ (**Fig. 9A**) was associated with affective and social cognitive domains (**Fig. 9C**), while the left VLPFC (**Fig. 9B**) only with processes of a cognitive domain such as language and social cognition (**Fig. 9C**). Given these results, SPP and TPP do not specifically engage different psychological domains. On the contrary, both are associated with similar affective and cognitive processes, despite the different brain regions related to these processes.

.....

Insert Figure 8 & Figure 9 about here

.....

Discussion

SOP represents an important mechanism for social behavior control, enabling cooperation within large-scale societies among genetically heterogeneous individuals (Fehr and Fischbacher, 2004a, b). The HPM posits that SPP and TPP engage similar cognitive processes for blameworthiness recognition but in different manners. Here, we investigated whether TPP and SPP engage different brain mechanisms associated with putatively different cognitive processes. As the basis for all subsequent analyses, we first identified the meta-analytic brain regions consistently activated by SPP and TPP (ALE method). We next determined the consensus connectivity networks of those SPP and TPP brain regions (using task-based MACM and task-free RSFC) and their sub-network compositions (using hierarchical analyses). Those steps allowed us to determine commonalities and differences of the neural patterns associated with SPP and TPP. We finally characterized the functional roles of these neural activations employing functional decoding analyses. Overall, we demonstrated that similar affective and cognitive processes are associated with the two forms of SOP. Their common neural patterns clustered into four functional networks: the salience, default-mode, frontotemporal, and medial prefrontal networks. However, we also observed a certain degree of neural and functional specificity for the two forms of SOP —bilateral AI for SPP, and VLPFC and pTPJ for TPP— with partially diverging neural network configurations.

The psychological processes of brain regions underlying punishment

We first tested the HPM hypothesis that TPP and SPP involve similar cognitive processes but while the SPP puts more weight on the harm of the transgression, TPP focuses on the intentions of the transgressor (Krueger and Hoffman, 2016). Our results showed that across forms of SOP, a set of common brain regions—the bilateral AI and the left DLPFC—were consistently activated by punishment decisions. While the right AI was more likely related to affective processing, the left AI and DLPFC were more strongly associated with the cognitive

domain. These findings mirror the results of the SOP contribution analyses showing that more TPP studies contributed to the cluster in the DLPFC associated with cognition than SPP studies, whereas the reverse was true for the AI, especially for the cluster in the right AI associated with affective processing.

Nonetheless, separate, single analyses for each form of SOP revealed a certain degree of specificity in the SPP and TPP neural patterns. On the one hand, SPP consistently activated the bilateral AI and was supported by salience (including the AI), subcortical (e.g., putamen and thalamus) and lateromedial prefrontal (e.g., MCC and DLPFC) networks. On the contrary, TPP consistently activated the left pTPJ and VLPFC and was supported by frontotemporal and frontoparietal networks (including IPL and inferior temporal gyrus). Both clusters were associated with affective and cognitive domains, although the SPP clusters loaded more on emotion processing and the TPP clusters on cognition. Finally, SOP did not only engage the cognitive and affective domains but also an interoceptive domain associated with negative emotions that might help map the victim's affective state for empathic concern. Thus, even though both forms of SOP recruit a common cognitive-affective-motivational network (Strobel et al., 2011), still differences remain. Such differences might be traced back to the different role dynamics between victim and punisher in SPP and TPP.

In SPP, the harm of a norm violation might be more salient, as the punisher and the victim are the same person. Indeed, norm-deviant behaviors are judged as more severe by those who are penalized by them. For example, individuals are more averse to disadvantageous than advantageous inequality (Bechtel et al., 2018; Fehr and Schmidt, 1999; Gao et al., 2018; Loewenstein et al., 1989). This might suggest an egocentric bias in evaluations of the severity of a norm violation, which might lead to a weighting imbalance between the transgression's severity and the transgressor's responsibility, resulting in harsher sanctions (Schiller et al., 2014; Sommers and Ellsworth, 2000). Future studies using, for instance, computational modeling to estimate subject-specific weighting of a transgression

might consider testing whether individuals weight a transgression's severity more when they are the target of the transgression, and whether this stronger weighting correlates with harsher punishments.

On the contrary, third-party punishers (as long as they have no relationship with the victim) are in the impartial position to carefully consider and equally weight the transgression's severity and the transgressor's responsibility. Thus, they might show a reduced bias in integration of harm evaluations and inferences on the transgressor's intentions for the determination of blameworthiness (Zhong et al., 2016).

Such differences were also reflected by the different networks the SPP and TPP were observed to engage. In particular, SPP engaged networks of regions such as the AI, MCC, DLPFC, SPL and subcortical areas involved in detection of a variety of norm violations, such as unfairness, dishonesty, defection of cooperation and betrayal (Bellucci et al., 2019a; Feng et al., 2015; Yang et al., 2019). On the contrary, TPP engaged networks of higher-order regions such as the IPL, middle frontal gyrus and the temporal cortex involved in transgression evaluations and responsibility attributions (Bellucci et al., 2017a; Berthoz et al., 2002). These results corroborate the hypothesis that norm violation is judged more severely in SPP than TPP, while both norm violation and assessment of responsibility are more equally weighted in TPP as opposed to SPP.

The neural networks underlying punishment

Next, we tested the HPM prediction that SOP recruits a specific set of domain-general networks. The salience network (e.g., AI) is supposed to detect the presence of a harmful act, signaling a norm violation; the default-mode network (e.g., pTPJ) is required to assign intentions and integrate harm signal to determine the blameworthiness of the wrong-doer; and, finally, the central-executive network (e.g., DLPFC) is hypothesized to sustain the last steps

of a punishment choice, namely, the infliction of the adequate sanction (Krueger and Hoffman, 2016).

We found only partial evidence to this hypothesis. SOP brain regions clustered indeed into three networks. First, we observed the salience network encompassing the AI, MCC and caudate (Dosenbach et al., 2007; Seeley et al., 2007). The MCC might monitor the contextual situation and urge to take action, representing negative emotions associated with the transgression (Hoffstaedter et al., 2014; Lieberman and Eisenberger, 2015). The caudate may reflect a desire to seek revenge for the suffered norm violation (Singer et al., 2006) or integrate behaviorally-relevant information for belief updating about the character of those who show norm-deviant behaviors, as this region is activated during interactions with unfair or immoral others (Harle et al., 2012; Servaas et al., 2015; Wardle et al., 2013).

The AI might be recruited to signal the unexpected norm violation implied by the harmful act, as this region signals violations of expectations and prediction errors in the aversive domain (Allen et al., 2016; Farrer and Frith, 2002; Koelsch et al., 2002). In particular, activity in the AI is rather related to expectancies of negative events than to the encoding of negative events as such (Lin et al., 2013). In the social domain, the AI might be recruited in response to actual or hypothetical violations of social expectancies, such as norm violations (Feng et al., 2015; Zinchenko and Arsalidou, 2018). For instance, activations in the AI are elicited by (hypothetical) defections of trust violating a reciprocity norm (Bellucci et al., 2017b; Delgado et al., 2005; van den Bos et al., 2009), and a norm violation more strongly engages the AI when perpetrated by an in-group member for whom expectations of social norm compliance are stronger (Wu et al., 2018).

Second, we found a frontoparietal network that largely overlaps with the domain-general central executive network encompassing the SPL and DLPFC. Both SPL and DPLFC are associative brain regions that allow the integration of cognitive and affective evaluations for the formation of abstract, conceptual knowledge that might sustain evaluation processes

related to the determination of the proper sanction (Carter and Huettel, 2013; Culham and Valyear, 2006; Wood and Grafman, 2003). In particular, the DLPFC has been suggested to encode social utility signals that reflect social preferences informative of an individual's propensity to engage in prosocial behaviors or to enforce social norm compliance (Holper et al., 2018; Ruff et al., 2013). This region is activated during norm-compliant behaviors triggered by punishment threats (Spitzer et al., 2007) and might take a specific role in the execution and selection of the appropriate punishment beyond evaluations of blameworthiness. Indeed, stimulation-induced disruption of the DLPFC impairs norm-enforcing behaviors leaving the recognition of the wrong-doing intact (Buckholtz et al., 2015; Knoch et al., 2006).

However, the HPM also posits that to determine the perpetrator's blameworthiness, weighting the severity of the norm violation is not sufficient for the choice of the proper punishment. Inferences on the transgressor's intentions need to be made, which are supposed to be carried out by the default-mode network, especially the pTPJ. Indeed, a basic tenet in criminal law poses that the act makes a person guilty only if the mind is also guilty (Shen et al., 2011). Accordingly, momentary disruption of the pTPJ via transcranial magnetic stimulation makes participants judge attempted harms as more permissible (Young et al., 2010).

We found evidence for the involvement of the default-mode network only in the conjunction analysis between the separate functional profiles of SPP and TPP. Given the heavy engagement of mentalizing brain regions by TPP, the results of this conjunction analysis might likely be driven by the importance of mentalizing regions in TPP. In fact, our ALE analysis revealed that TPP strongly engaged the pTPJ—a central region for inferences on the others' intentions (Saxe and Kanwisher, 2003; Saxe and Powell, 2006)— and the VLPFC —a pivotal region in regulatory processes for prosocial behaviors (Fouragnan et al., 2013; Souza et al., 2009; Yang et al., 2019). In particular, in TPP, the pTPJ might weight the intentions and beliefs of the transgressors during the wrongdoing, while the VLPFC might

dampen the harm-driven urge of harsh punishments promoting fairer sanctions. On the contrary, no mentalizing brain regions were found for SPP.

Limitations

This meta-analysis provided an overview of the psychological processes of brain regions and neural networks involved in different types of SOP. However, there are some limitations to our study that deserve discussion. First, variations in the intentionality of a norm violation (e.g., accidental vs. attended harm) might help better understand how the transgression's severity and the transgressor's responsibility are weighted for the determination of blameworthiness. This might, for instance, clarify the role of the default-mode network in SPP as well, as this neural network was preferentially engaged by TPP in the current work. Furthermore, it might elucidate the role of the amygdala and medial PFC, which, contrary to the HPM predictions, were not found in the current study. The HPM proposes that the amygdala signals the severity of the inflicted harm, whereas the medial PFC evaluates the transgressor's blameworthiness integrating information about the transgression's severity and the transgressor's responsibility. One reason for this null finding might lie in the fact that fMRI studies have not so far disentangled evaluations of the transgression's harm from evaluations of the transgressor's blameworthiness.

Second, it is still an open question whether different psychological processes and neural patterns are evoked by hypothetical and actual punishment decisions. In particular, punishment decisions have been studied using either vignettes where participants are asked to make hypothetical punishment decisions or economic games where participants make actual, costly punishment decisions. Due to the paucity of studies, we were not able to address this open question, but we here notice that there are already conflicting results in the literature that might be due to these two different paradigms. For instance, in one study, impairment of punishment for wrongful acts could be experimentally achieved only via disruption of the

right DLPFC (Knoch et al., 2006), whereas in a more recent study, no lateralization effects were found after bilateral DLPFC disruption, despite successful punishment reduction (Buckholtz et al., 2015). The use of different paradigms might well explain these different findings, since the former study used an economic game while the latter asked participants to make hypothetical decisions. Thus, participants in the economic game might have faced a more conflicting situation that required reliance on the right DLPFC, which is central to control adjustments in high conflict situations (Mansouri et al., 2009). Future studies are still needed to better understand how these different paradigms and their associated psychological components and neural signatures interact with each other to bring about a punishment decision.

Despite these limitations, we identified a consensus connectivity network that entails candidate brain regions for the representations of the core functional mechanisms underlying SOP. This network might be used in future studies to test how it instantiates the processes that bring about a punishment decision. For instance, neural activity within this network might be computed for predictions of individual decisions to punish wrongdoing. Predictive models based on this network might yield better performance than models based on whole-brain activity (Bellucci et al., 2019b) or domain-general networks that are likely unrelated to the punishment phenomenon. Finally, the identification of this network might provide insights into investigations of individual differences in norm violations and clinical traits such as psychopathology and social phobia (Blair et al., 2010; Veit et al., 2010).

Conclusions

Taken together, our results demonstrated that different forms of SOP engage complementary neural networks with converging functional roles. These neural networks converged on common connectivity patterns revealing an extended, consensus connectivity network. However, given the complex and fine-grained network organization yielded by the separate

635 analyses for each form of SOP, the HPM in its preliminary formulation might be too coarse
636 and requires revision. Thus, future work is still needed to experimentally clarify the functional
637 role and interactions of these brain regions and networks. By highlighting the specific neural
638 and functional cliques that underlie SOP, our work will help future investigations in shaping
639 research hypotheses to shed light on one of the most peculiar human behaviors.
640

Acknowledgments

GB is supported by the Max Planck Society. SBE is supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/11-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement 785907 (HBP SGA2).

Conflict of Interest

The authors are unaware of any conflicts of interest, financial or otherwise.

References

Allen, M., Fardo, F., Dietz, M.J., Hillebrandt, H., Friston, K.J., Rees, G., Roepstorff, A., 2016. Anterior insula coordinates hierarchical processing of tactile mismatch responses. Neuroimage 127, 34-43.

Baumgartner, T., Götte, L., Gügler, R., Fehr, E., 2012. The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. Human Brain Mapping 33, 1452-1469.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L.R., Knoch, D., 2014. Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. Social cognitive and affective neuroscience 9, 653-660.

Bechtel, M.M., Liesch, R., Scheve, K.F., 2018. Inequality and redistribution behavior in a give-or-take game. Proceedings of the National Academy of Sciences of the United States of America 115, 3611-3616.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K.M., Grafman, J., Krueger, F., 2017a. Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. Soc Neurosci 12, 124-134.

Bellucci, G., Chernyak, S.V., Goodyear, K., Eickhoff, S.B., Krueger, F., 2017b. Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. Hum Brain Mapp 38, 1233-1248.

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.

Bellucci, G., Molter, F., Park, S.Q., 2019a. Neural representations of honesty predict future trust behavior. Nat Commun 10, 5184.

Bellucci, G., Münte, T.F., Park, S.Q., 2019b. Resting-state dynamics as a neuromarker of dopamine administration in healthy female adults. J Psychopharmacol, 269881119855983.

Berthoz, S., Armony, J.L., Blair, R.J., Dolan, R.J., 2002. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. Brain : a journal of neurology 125, 1696-1708.

Bicchieri, C., 2014. Norms, conventions, and the power of expectations, in: Cartwright, N., Montuschi, E. (Eds.), Philosophy of social science: A new introduction. Oxford University Press, Oxford, pp. 208-229.

Blair, K.S., Geraci, M., Hollon, N., Otero, M., DeVido, J., Majestic, C., Jacobs, M., Blair, R.J., Pine, D.S., 2010. Social norm processing in adult social phobia: atypically increased ventromedial frontal cortex responsiveness to unintentional (embarrassing) transgressions. Am J Psychiatry 167, 1526-1532.

Blount, S., 1995. When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. Organizational Behavior and Human Decision Processes 63, 131-144.

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, 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, 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, 1369-1380.

Camilleri, J.A., Muller, V.I., Fox, P., Laird, A.R., Hoffstaedter, F., Kalenscher, T., Eickhoff, S.B., 2018. Definition and characterization of an extended multiple-demand network. Neuroimage 165, 138-147.

Carter, R.M., Huettel, S.A., 2013. A nexus model of the temporal-parietal junction. Trends Cogn Sci 17, 328-336.

Civai, C., 2013. Rejecting unfairness: emotion-driven reaction or cognitive heuristic? Front Hum Neurosci 7, 126.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., Rumiati, R.I., 2010. Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. Cognition 114, 89-95.

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.

Clos, M., Rottschy, C., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2014. Comparison of structural covariance with functional connectivity approaches exemplified by an investigation of the left anterior insula. Neuroimage 99, 269-280.
Cole, M.W., Bassett, D.S., Power, J.D., Braver, T.S., Petersen, S.E., 2014. Intrinsic and task-evoked network architectures of the human brain. Neuron 83, 238-251.
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. Social cognitive and affective neuroscience 8, 424-431.
Craig, A.D., 2002. How do you feel? Interoception: the sense of the physiological condition of the body. Nature reviews. Neuroscience 3, 655-666.
Culham, J.C., Valyear, K.F., 2006. Human parietal cortex in action. Curr Opin Neurobiol 16, 205-212.
Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J., Hichwa, R.D., 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience 3, 1049-1056.
Dawes, C.T., Fowler, J.H., Johnson, T., McElreath, R., Smirnov, O., 2007. Egalitarian motives in humans. Nature 446, 794-796.
de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305, 1254-1258.
Delgado, M.R., Frank, R.H., Phelps, E.A., 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. Nature neuroscience 8, 1611-1618.
Dosenbach, N.U., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A., Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E., Schlaggar, B.L., Petersen, S.E., 2007. Distinct brain networks for adaptive and stable task control in humans. Proceedings of the National Academy of Sciences of the United States of America 104, 11073-11078.
Egas, M., Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society B: Biological Sciences 275, 871-878.
Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. Neuroimage 59, 2349-2361.
Eickhoff, S.B., Bzdok, D., Laird, A.R., Roski, C., Caspers, S., Zilles, K., Fox, P.T., 2011. Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. NeuroImage 57, 938-949.
Eickhoff, S.B., Constable, R.T., Yeo, B.T.T., 2018. Topographic organization of the cerebral cortex and brain cartography. Neuroimage 170, 332-347.
Eickhoff, S.B., Laird, A.R., Grefkes, C., Wang, L.E., Zilles, K., Fox, P.T., 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. Human brain mapping 30, 2907-2926.
Eickhoff, S.B., Nichols, T.E., Laird, A.R., Hoffstaedter, F., Amunts, K., Fox, P.T., Bzdok, D., Eickhoff, C.R., 2016. Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. Neuroimage 137, 70-85.
Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences of the United States of America 113, 7900-7905.
Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness—Intentions matter. Games and Economic Behavior 62, 287-303.
Farrer, C., Frith, C.D., 2002. Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. Neuroimage 15, 596-603.
Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425, 785-791.
Fehr, E., Fischbacher, U., 2004a. Social norms and human cooperation. Trends Cogn Sci 8, 185-190.
Fehr, E., Fischbacher, U., 2004b. Third-party punishment and social norms. Evolution and Human Behavior 25, 63-87.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137-140.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of

754 Economics 114, 817-868.
755 Feng, C., Luo, Y.-J., Krueger, F., 2015. Neural signatures of fairness-related normative decision making
756 in the ultimatum game: A coordinate-based meta-analysis. Human Brain Mapping 36, 591-602.
757 Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., Coricelli, G., 2013. Reputational Priors
758 Magnify Striatal Responses to Violations of Trust. J Neurosci 33, 3602-3611.
759 Fox, P.T., Lancaster, J.L., 2002. Mapping context and content: the BrainMap model. Nature Reviews
760 Neuroscience 3, 319-321.
761 Fox, P.T., Lancaster, J.L., Laird, A.R., Eickhoff, S.B., 2014. Meta-Analysis in Human Neuroimaging:
762 Computational Modeling of Large-Scale Databases. Annual Review of Neuroscience 37, 409-434.
763 Frith, U., Frith, C.D., 2003. Development and neurophysiology of mentalizing. Philosophical
764 transactions of the Royal Society of London. Series B, Biological sciences 358, 459-473.
765 Gao, X., Yu, H., Sáez, I., Blue, P.R., Zhu, L., Hsu, M., Zhou, X., 2018. Distinguishing neural correlates of
766 context-dependent advantageous- and disadvantageous-inequity aversion. PNAS 115, E7680-E7689.
767 Genon, S., Reid, A., Langner, R., Amunts, K., Eickhoff, S.B., 2018. How to Characterize the Function of a
768 Brain Region. Trends Cogn Sci 22, 350-364.
769 Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko,
770 Y.V., Roach, E.L., Korgaonkar, M.S., Grieve, S.M., Galatzer-Levy, I., Fox, P.T., Etkin, A., 2015.
771 Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 72, 305-315.
772 Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E.,
773 Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller,
774 K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved
775 resting state network imaging. Neuroimage 95, 232-247.
776 Guala, F., 2012. Reciprocity: weak or strong? What punishment experiments do (and do not)
777 demonstrate. Behav Brain Sci 35, 1-15.
778 Güroğlu, B., van den Bos, W., Rombouts, S.A., Crone, E.A., 2010. Unfair? It depends: neural correlates
779 of fairness in social context. Social cognitive and affective neuroscience 5, 414-423.
780 Hardwick, R.M., Lesage, E., Eickhoff, C.R., Clos, M., Fox, P., Eickhoff, S.B., 2015. Multimodal
781 connectivity of motor learning-related dorsal premotor cortex. Neuroimage 123, 114-128.
782 Harle, K.M., Chang, L.J., van 't Wout, M., Sanfey, A.G., 2012. The neural mechanisms of affect infusion
783 in social economic decision-making: a mediating role of the anterior insula. Neuroimage 61, 32-40.
784 Hoffstaedter, F., Grefkes, C., Caspers, S., Roski, C., Palomero-Gallagher, N., Laird, A.R., Fox, P.T.,
785 Eickhoff, S.B., 2014. The role of anterior midcingulate cortex in cognitive motor control: evidence
786 from functional connectivity analyses. Hum Brain Mapp 35, 2741-2753.
787 Holper, L., Burke, C.J., Fausch, C., Seifritz, E., Tobler, P.N., 2018. Inequality signals in dorsolateral
788 prefrontal cortex inform social preference models. Soc Cogn Affect Neurosci 13, 513-524.
789 Igelstrom, K.M., Graziano, M.S.A., 2017. The inferior parietal lobule and temporoparietal junction: A
790 network perspective. Neuropsychologia 105, 70-83.
791 Igelström, K.M., Webb, T.W., Graziano, M.S., 2015. Neural Processes in the Human Temporoparietal
792 Cortex Separated by Localized Independent Component Analysis. J Neurosci 35, 9432-9445.
793 Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E., 2006. Diminishing reciprocal fairness by
794 disrupting the right prefrontal cortex. Science 314, 829-832.
795 Koelsch, S., Gunter, T.C., v. Cramon, D.Y., Zysset, S., Lohmann, G., Friederici, A.D., 2002. Bach Speaks: A
796 Cortical “Language-Network” Serves the Processing of Music. NeuroImage 17, 956-966.
797 Kolling, N., Wittmann, M.K., Behrens, T.E., Boorman, E.D., Mars, R.B., Rushworth, M.F., 2016. Value,
798 search, persistence and model updating in anterior cingulate cortex. Nat Neurosci 19, 1280-1285.
799 Krueger, F., Hoffman, M., 2016. The Emerging Neuroscience of Third-Party Punishment. Trends
800 Neurosci 39, 499-501.
801 Laird, A.R., Eickhoff, S.B., Kurth, F., Fox, P.M., Uecker, A.M., Turner, J.A., Robinson, J.L., Lancaster, J.L.,
802 Fox, P.T., 2009a. ALE meta-analysis workflows via the brainmap database: progress towards a
803 probabilistic functional brain atlas. Frontiers in neuroinformatics 3, 23.
804 Laird, A.R., Eickhoff, S.B., Kurth, F., Fox, P.M., Uecker, A.M., Turner, J.A., Robinson, J.L., Lancaster, J.L.,
805 Fox, P.T., 2009b. ALE Meta-Analysis Workflows Via the Brainmap Database: Progress Towards A

Probabilistic Functional Brain Atlas. Front Neuroinform 3, 23.
Laird, A.R., Fox, P.M., Price, C.J., Glahn, D.C., Uecker, A.M., Lancaster, J.L., Turkeltaub, P.E., Kochunov, P., Fox, P.T., 2005. ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. Human brain mapping 25, 155-164.
Langner, R., Rottschy, C., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2014. Meta-analytic connectivity modeling revisited: controlling for activation base rates. NeuroImage 99, 559-570.
Lieberman, M.D., Eisenberger, N.I., 2015. The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference. Proceedings of the National Academy of Sciences of the United States of America 112, 15250-15255.
Lin, C.S., Hsieh, J.C., Yeh, T.C., Lee, S.Y., Niddam, D.M., 2013. Functional dissociation within insular cortex: the effect of pre-stimulus anxiety on pain. Brain Res 1493, 40-47.
Loewenstein, G.F., Thompson, L., Bazerman, M.H., 1989. Social Utility and Decision Making in Interpersonal Contexts. Journal of Personality and Social Psychology 57, 426-441.
Mansouri, F.A., Tanaka, K., Buckley, M.J., 2009. Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. Nat Rev Neurosci 10, 141-152.
Marrelec, G., Krainik, A., Duffau, H., Pelegrini-Issac, M., Lehericy, S., Doyon, J., Benali, H., 2006. Partial correlation for functional brain interactivity investigation in functional MRI. Neuroimage 32, 228-237.
McAuliffe, K., Jordan, J.J., Warneken, F., 2015. Costly third-party punishment in young children. Cognition 134, 1-10.
Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J., Singer, T., 2018. Preschool children and chimpanzees incur costs to watch punishment of antisocial others. Nat Hum Behav 2, 45-51.
Müller, V.I., Cieslik, E.C., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2013. Dysregulated left inferior parietal activity in schizophrenia and depression: functional connectivity and characterization. Front Hum Neurosci 7, 268.
Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.B., 2005. Valid conjunction inference with the minimum statistic. Neuroimage 25, 653-660.
Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedict, M.M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz, S.T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R.T., Kamiel, S.M., Anwar, A.R., Hinz, C.M., Kaplan, M.S., Rachlin, A.B., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Craddock, C.C., Calhoun, V., Courtney, W., King, M., Wood, D., Cox, C.L., Kelly, A.M., Di Martino, A., Petkova, E., Reiss, P.T., Duan, N., Thomsen, D., Biswal, B., Coffey, B., Hoptman, M.J., Javitt, D.C., Pomara, N., Sidtis, J.J., Koplewicz, H.S., Castellanos, F.X., Leventhal, B.L., Milham, M.P., 2012. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. Frontiers in neuroscience 6, 152.
Raichle, M.E., 2015. The restless brain: how intrinsic activity organizes brain function. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 370.
Riedl, K., Jensen, K., Call, J., Tomasello, M., 2012. No third-party punishment in chimpanzees. Proc Natl Acad Sci U S A 109, 14824-14829.
Rottschy, C., Caspers, S., Roski, C., Reetz, K., Dogan, I., Schulz, J.B., Zilles, K., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2012. Differentiated parietal connectivity of frontal regions for “what” and “where” memory. Brain Structure and Function 218, 1551-1567.
Ruff, C.C., Ugazio, G., Fehr, E., 2013. Changing social norm compliance with noninvasive brain stimulation. Science 342, 482-484.
Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449-468.
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, 1755-1758.
Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage 64, 240-256.
Saxe, R., Kanwisher, N., 2003. People thinking about thinking peopleThe role of the temporo-parietal

junction in “theory of mind”. NeuroImage 19, 1835-1842.
Saxe, R., Powell, L.J., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17, 692-699.
Schiller, B., Baumgartner, T., Knoch, D., 2014. Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. Evolution and Human Behavior 35, 169-175.
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.
Servaas, M.N., Aleman, A., Marsman, J.B., Renken, R.J., Riese, H., Ormel, J., 2015. Lower dorsal striatum activation in association with neuroticism during the acceptance of unfair offers. Cogn Affect Behav Neurosci 15, 537-552.
Seymour, B., Singer, T., Dolan, R., 2007. The neurobiology of punishment. Nature reviews. Neuroscience 8, 300-311.
Shen, F.X., Hoffman, M.B., Jones, O.D., Greene, J.D., Marois, R., 2011. Sorting Guilty Minds. New York U Law Rev 86, 1306-1360.
Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., Frith, C.D., 2006. Empathic neural responses are modulated by the perceived fairness of others. Nature 439, 466-469.
Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proceedings of the National Academy of Sciences of the United States of America 106, 13040-13045.
Sommers, S.R., Ellsworth, P.C., 2000. Race in the Courtroom: Perceptions of Guilt and Dispositional Attributions. Personality and Social Psychology Bulletin 26, 1367-1379.
Souza, M.J., Donohue, S.E., Bunge, S.A., 2009. Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. Neuroimage 46, 299-307.
Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G., Fehr, E., 2007. The neural signature of social norm compliance. Neuron 56, 185-196.
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.
Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., Kirsch, P., 2011. Beyond revenge: neural and genetic bases of altruistic punishment. Neuroimage 54, 671-680.
Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychological science 19, 339-347.
Tavor, I., Parker Jones, O., Mars, R.B., Smith, S.M., Behrens, T.E., Jbabdi, S., 2016. Task-free MRI predicts individual differences in brain activity during task performance. Science 352, 216-220.
Timm, N.H., 2002. Applied Multivariate Analysis. Springer-Verlag New York, New York.
Turkeltaub, P.E., Eden, G.F., Jones, K.M., Zeffiro, T.A., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. Neuroimage 16, 765-780.
Turkeltaub, P.E., Eickhoff, S.B., Laird, A.R., Fox, M., Wiener, M., Fox, P., 2012. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. Human brain mapping 33, 1-13.
van 't Wout, M., Kahn, R.S., Sanfey, A.G., Aleman, A., 2006. Affective state and decision-making in the Ultimatum Game. Exp Brain Res 169, 564-568.
van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S.A., Crone, E.A., 2009. What motivates repayment? Neural correlates of reciprocity in the Trust Game. Social cognitive and affective neuroscience 4, 294-304.
Veit, R., Lotze, M., Sewing, S., Missenhardt, H., Gaber, T., Birbaumer, N., 2010. Aberrant social and cerebral responding in a competitive reaction time paradigm in criminal psychopaths. Neuroimage 49, 3365-3372.
Wang, J., Fan, L., Wang, Y., Xu, W., Jiang, T., Fox, P.T., Eickhoff, S.B., Yu, C., Jiang, T., 2015.

910 Determination of the posterior boundary of Wernicke's area based on multimodal connectivity
911 profiles. Hum Brain Mapp 36, 1908-1924.
912 Wardle, M.C., Fitzgerald, D.A., Angstadt, M., Sripada, C.S., McCabe, K., Phan, K.L., 2013. The caudate
913 signals bad reputation during trust decisions. PLoS One 8, e68884.
914 Wood, J.N., Grafman, J., 2003. Human prefrontal cortex: processing and representational
915 perspectives. Nat Rev Neurosci 4, 139-147.
916 Wu, C.T., Fan, Y.T., Du, Y.R., Yang, T.T., Liu, H.L., Yen, N.S., Chen, S.H., Hsung, R.M., 2018. How Do
917 Acquired Political Identities Influence Our Neural Processing toward Others within the Context of a
918 Trust Game? Front Hum Neurosci 12, 23.
919 Yang, Z., Zheng, Y., Yang, G., Li, Q., Liu, X., 2019. Neural Signatures of Cooperation Enforcement and
920 Violation: A Coordinate-based Meta-analysis. Social cognitive and affective neuroscience.
921 Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R., 2010. Disruption of the right
922 temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral
923 judgments. Proc Natl Acad Sci U S A 107, 6753-6758.
924 Zhong, S., Chark, R., Hsu, M., Chew, S.H., 2016. Computational substrates of social norm enforcement
925 by unaffected third parties. Neuroimage 129, 95-104.
926 Zinchenko, O., 2019. Brain responses to social punishment: a meta-analysis. Sci Rep 9, 12800.
927 Zinchenko, O., Arsalidou, M., 2018. Brain responses to social norms: Meta-analyses of fMRI studies.
928 Hum Brain Mapp 39, 955-970.
929

930

Figure Legends

Figure 1. Meta-analytic results. Results of meta-analytic ALE analyses for SOP (**A**), SPP (**B**) and TPP (**C**). The depicted brain regions are consistently activated clusters across published fMRI studies that survived correction for multiple comparisons controlling for cluster-level familywise error (*cFWE* < .05). Finally, overlaps of the three meta-analytic maps (**D**) are shown to compare the anatomical extent of the observed clusters.

L, left; R, right; ALE, activation likelihood estimation; fMRI, functional magnetic resonance imaging; ∩, conjunction; DLPFC, dorsolateral prefrontal cortex; AI, anterior insula; VLPFC, ventrolateral prefrontal cortex; SOP, social punishment; SPP, second-party punishment; TPP, third-party punishment.

Figure 2. Task-free, resting-state functional connectivity (RSFC) for SOP. Regions significantly connected to the three clusters consistently activated for SOP, namely, the left AI (**red**), left DLPFC (**green**) and right AI **(blue)** based on RSFC analyses.

L, left; R, right; ∩, conjunction; SOP, social punishment; AI, anterior insula; DLPFC, dorsolateral prefrontal cortex; cluster-level familywise error (*cFWE*) < .05.

Figure 3. Task-free, resting-state functional connectivity (RSFC) for SPP. Regions significantly connected to the three clusters consistently activated for SPP, namely, the left AI (**red**) and right AI **(green)** based on RSFC analyses.

L, left; R, right; ∩, conjunction; SPP, second-party punishment; AI, anterior insula; cluster-level familywise error (*cFWE*) < .05.

Figure 4. Task-free, resting-state functional connectivity (RSFC) for TPP. Regions significantly connected to the three clusters consistently activated for TPP, namely, the left pTPJ (red) and left VLPFC (green) based on RSFC analyses.

L, left; ∩, conjunction; TPP, third-party punishment; pTPJ, posterior temporoparietal junction; VLPFC, ventrolateral prefrontal cortex; cluster-level familywise error (*cFWE*) < .05.

Figure 5. Consensus connectivity network. Results of consensus connectivity map for SOP **(A)** and overlaps of consensus connectivity maps for SPP and TPP (**B**).

SOP, social punishment; TPP, third-party punishment; SPP, second-party punishment; L, left; R, right; SPL, superior parietal lobule; DLPFC, dorsolateral prefrontal cortex; FPC, frontoparietal cortex; AI, anterior insula; MCC, middle cingulate cortex; AG, angular gyrus; Put, putamen; STG, superior temporal gyrus.

Figure 6. Hierarchical clustering analyses. Results from the clustering analyses based on the RSFC profiles of the meta-analytic clusters for SOP (**A**), SPP (**B**), TPP (**C**) and for the overlapping regions of the SPP and TPP results (**D**).

L, left; R, right; AG, angular gyrus; Thal, thalamus; Caud, caudate; Cbl, cerebellum; IPL, inferior parietal lobule; SPL, superior parietal lobule; IFG, inferior frontal gyrus; OrbC, orbital cortex; MTG, middle temporal gyrus; MCC, middle cingulate cortex; PreMC, premotor cortex; Put, putamen; STG, superior temporal gyrus; TP, temporal pole; MFG, middle frontal gyrus; PreG, precentral gyrus; ITG, inferior temporal gyrus.

Figure 7. Functional decoding analyses for SOP. Functional profiles of the left AI (**A**), left DLPFC (**B**) and right AI (**C**) and their functional decoding (**D**) based on meta-categories in

the BrainMap database. Around the spider plot are the behavioral domains yielded by forward inference, i.e., categories of mental operations likely to be isolated by the experiments in the BrainMap database. In parentheses are the subcategories that specify the behavioral domains. Depicted values are likelihood ratios.

L, left; R, right; SOP, social punishment; AI, anterior insula; DLPFC, dorsolateral prefrontal cortex; *FDR < .05.

Figure 8. Functional decoding analyses for SPP. Functional profiles of the left AI (**A**) and right AI (**B**) and their functional decoding (**C**) based on meta-categories in the BrainMap database. Around the spider plot are the behavioral domains yielded by forward inference, i.e., categories of mental operations likely to be isolated by the experiments in the BrainMap database. In parentheses are the subcategories that specify the behavioral domains. Depicted values are likelihood ratios.

L, left; R, right; SPP, second-party punishment; AI, anterior insula; *FDR < .05.

Figure 9. Functional decoding analyses for TPP. Functional profiles of the left AI (**A**) and right AI (**B**) and their functional decoding (**C**) based on meta-categories in the BrainMap database. Around the spider plot are the behavioral domains yielded by forward inference, i.e., categories of mental operations likely to be isolated by the experiments in the BrainMap database. In parentheses are the subcategories that specify the behavioral domains. Depicted values are likelihood ratios.

L, left; TPP, third-party punishment; pTPJ, posterior temporoparietal junction; VLPFC, ventrolateral prefrontal cortex; *FDR < .05.

Table Legends

Table 1. ALE meta-analysis results for punishment. ALE main-effect results for social punishment, second-party punishment and third-party punishment (*cFWE* < .05). The right anterior insula cluster was also more significantly activated for second-party punishment than third-party punishment in contrast analyses. DLPFC, dorsolateral prefrontal cortex; VLPFC, ventrolateral prefrontal cortex; IFG, inferior frontal gyrus; pTPJ, posterior temporoparietal junction; MTG, middle temporal gyrus; BA, Brodmann area; anatomical assignment based on the Anatomy toolbox in parentheses; L, left; ALE, activation likelihood estimation; MNI, Montreal Neurological Institute.