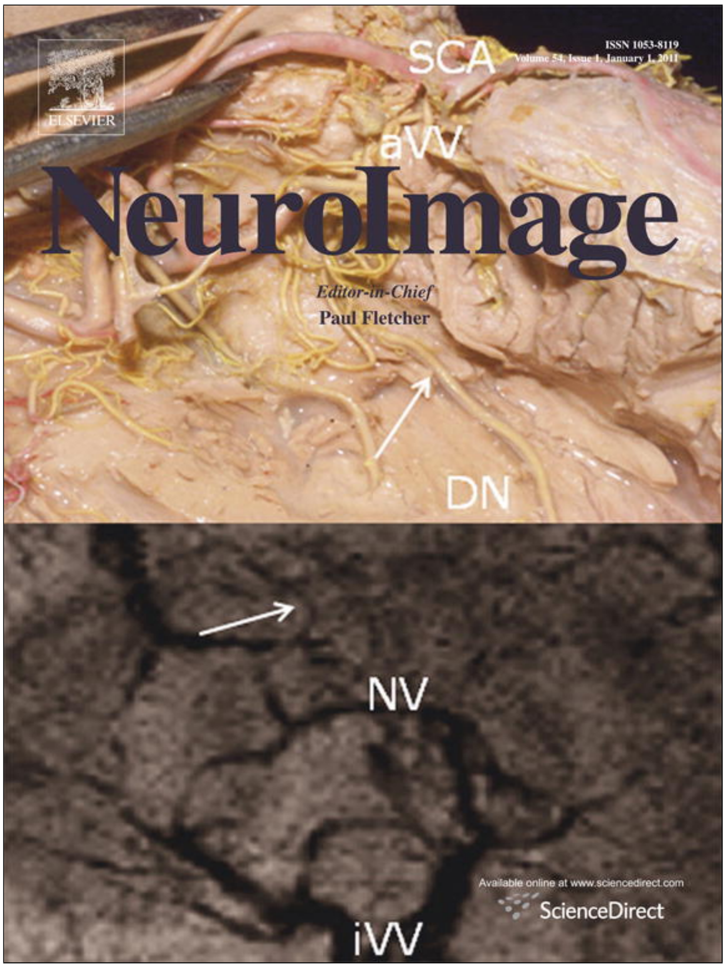
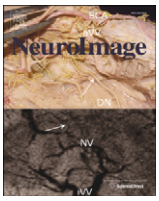

ELSEVIER
ISSN 1053-8119
Volume 54, Issue 1, January 1, 2011
SCA
aVV
NeuroImage
Editor-in-Chief
Paul Fletcher
DN
NV
Available online at www.sciencedirect.com
ScienceDirect
iVV

NeuroImage 54 (2011) 671–680

Contents lists available at ScienceDirect

NeuroImage

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



Beyond revenge: Neural and genetic bases of altruistic punishment

Alexander Strobel [a,b,*,1], Jan Zimmermann [a,c,1], Anja Schmitz [d,e], Martin Reuter [f], Stefanie Lis [g], Sabine Windmann [a,1], Peter Kirsch [g,h,*,1]

[a] Institute of Psychology, Goethe University Frankfurt/Main, Germany
[b] Department of Psychology, Technische Universität Dresden, Germany
[c] Department of Cognitive Neuroscience, Maastricht University, The Netherlands
[d] Institute of Psychology, University of Giessen, Germany
[e] National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[f] Institute of Psychology, University of Bonn, Germany
[g] Centre for Psychiatry, University of Giessen, Germany
[h] Department of Clinical Psychology, Central Institute of Mental Health, Mannheim, Germany

ARTICLE INFO

Article history:
Received 12 January 2010
Revised 28 June 2010
Accepted 21 July 2010
Available online 27 July 2010

Keywords:
Altruistic punishment
Functional magnetic resonance imaging
Nucleus accumbens
Nucleus caudatus
Catechol-O-methyltransferase
Polymorphism

ABSTRACT

It is still debated how altruistic punishment as one form of strong reciprocity has established during evolution and which motives may underlie such behavior. Recent neuroscientific evidence on the activation of brain reward regions during altruistic punishment in two-person one-shot exchange games suggests satisfaction through the punishment of norm violations as one underlying motive. In order to address this issue in more detail, we used fMRI during a one-shot economic exchange game that warrants strong reciprocity by introducing a third party punishment condition wherein revenge is unlikely to play a role. We report here that indeed, reward regions such as the nucleus accumbens showed punishment-related activation. Moreover, we provide preliminary evidence that genetic variation of dopamine turnover impacts similarly on punishment-related nucleus accumbens activation during *both* first person and third party punishment. The overall pattern of results suggests a common cognitive-affective-motivational network as the driving force for altruistic punishment, with only quantitative differences between first person and third party perspectives.

Introduction

While altruistic behavior can be observed in several species apart from humans, accounts for such behavior are typically based on genetic relatedness (kin-altruism) or benefits arising from altruistic behavior in repeated interactions (Bowles and Gintis, 2004; Fehr and Gächter, 2002). These theories, however, cannot explain why humans show altruistic behavior and cooperation even in anonymous and unattended interactions with strangers.

The theory of strong reciprocity (Bowles and Gintis, 2004) provides an explanation based on so-called altruistic punishment, which may have developed via gene–culture coevolution (Gintis, 2003). Indeed, experimental and simulation studies have shown that cooperation can be maintained even in larger groups and in one-shot interactions, if there is the possibility to punish defectors (Fehr and Gächter, 2002; Boyd et al., 2003; Fehr and Fischbacher, 2004; see, however, Ohtsuki et al., 2009). Such strongly reciprocal behavior is commonly termed *altruistic punishment* and is defined as the costly punishment of norm violations, which does not involve any overt benefit for the punisher. However, even in one-shot interactions, there may be some covert benefits of punishment (e.g., the satisfaction of revenge, the experience of power, the expectation of future rewards). Recent neuroscientific studies have begun to shed more light on the brain processes during altruistic punishment in order to provide an additional level of evidence for the discussion of the motives underlying this behavior.

In a functional magnetic resonance imaging (fMRI) study, Sanfey et al. (2003) scanned their participants while they had to decide whether to accept or to reject fair or unfair splits of a sum of money by a proposer in sequential one-shot interactions in the context of an Ultimatum Game. Critically, if recipients in the Ultimatum Game reject the proposal, neither the proposer nor the participants themselves receives money. Hence, the rejection of unfair offers bears a cost without a benefit and can thus be conceived of as a form of altruistic punishment. The main findings of this study were that right dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex

* Corresponding authors. A. Strobel is to be contacted at Department of Psychology, Technische Universität Dresden, Mommsenstr. 13, D-01062 Dresden, Germany. Fax: +49 351 463 36993. P. Kirsch, Department of Clinical Psychology, Central Institute of Mental Health, J 5, D-68159 Mannheim, Germany. Fax: +49 621 1703 706501.

E-mail addresses: alex.strobel@gmx.de (A. Strobel), peter.kirsch@zi-mannheim.de (P. Kirsch).

[1] These authors contributed equally to this work.

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.07.051

(ACC), and anterior insula exhibited stronger activation during unfair offers, with the activity of the anterior insula being positively correlated with rejection rates for unfair offers. The authors interpreted their findings based on the prominent role of the DLPFC in cognitive control processes, of the ACC in monitoring (cognitive-affective) conflict, and of the insula in subserving emotional processing via representations of signals of (especially aversive) internal states, highlighting the importance of emotions in economic decision-making. Indeed, recent evidence supports the view that insular representations of emotional states (Singer et al., 2009) may serve as bias or error signals in economic decision-making, which drive the motivation to reject unfair offers, and thereby to punish norm violations (see Montague and Lohrenz, 2007).

Another brain region that might be implicated in the rejection of unfair offers in the Ultimatum Game was highlighted in a lesion study by Koenigs and Tranel (2007). These authors showed that compared to a control group, individuals with lesions in the ventromedial prefrontal cortex–with the region of the strongest overlap of lesions tapping into the orbitofrontal cortex (OFC)–were more likely to reject unfair offers. As OFC damage has been associated with emotional dysregulation and failures in emotion-guided decision making (e.g., Bechara et al., 2000), Koenigs and Tranel (2007) argued that OFC lesions might impair emotion down-regulation (i.e., anger down-regulation when facing unfair offers) which may lead to economically irrational behavior (i.e., the rejection of unfair, but non-zero offers). Thus, this interpretation substantiates the role of emotional processes in altruistic punishment, but also suggests that non-altruistic motives can drive costly punishment.

This view is underscored by the results of a more direct investigation of the neural processes underlying altruistic punishment: De Quervain et al. (2004) performed positron emission tomography (PET) scans while their participants could use part of their reimbursement to punish defectors in a Trust Game. Stronger activation of the nucleus caudatus (NCd) was observed during effective as compared to symbolic punishment. The authors suggest that–as the NCd has been implicated in reward processing (e.g., Delgado et al., 2003)–the motivation to punish defectors could be partly due to feelings of satisfaction when social norm violations are punished and justice is reestablished.

However, three issues in these seminal studies require further investigation. Firstly, in these investigations, participants were directly affected by the unfair behavior of the other players, whom they could punish. In such situations, punishment may be driven by anger and revenge-like motives, thus presumably reflecting conditions where punishment is subjectively beneficial via satisfaction through revenge. To test this interpretation, it is important to contrast such conditions with others where the punisher is not directly affected by unfair behavior, so that revenge-like motives cannot account for punishment. Secondly, in the De Quervain et al. (2004) study, NCd activation was stronger in two conditions where punishment was effective. Yet, only one of these conditions was an operationalization of altruistic–i.e., costly–punishment, which raises the question whether the NCd might be implicated in *effective* rather than in *altruistic* punishment. Thirdly, the pattern of neural activation in the above-mentioned studies may provide some hint for the empirical examination of the assumption of genetic factors impacting on the development of altruism (Gintis, 2003). The activity of the regions highlighted above, i.e., DLPFC, ACC, Insula, and NCd is critically modulated by dopaminergic projections from the midbrain, which have also been implicated in reward processing and prediction error signaling (Schultz, 1998). Therefore, genetic variation in dopamine function could be expected to impact on neural responses to norm violations in which punishment of defectors is possible.

In the present study, we addressed these issues in order to increase our knowledge of those brain regions involved in altruistic punishment in order to provide a more differentiated basis for a neuroscientifically based account of the motives underlying altruistic punishment. To this end, we employed an economic exchange game based on the Dictator Game (DG). In its original first person version, one player A, the dictator, has to decide how to split a pie (usually a sum of money or monetary units) between him- or herself and another player B (the recipient), who has no means to reject this decision, even in the case of very unfair assignments. In our version of the game, we were interested in the behavior of player B whom we gave the opportunity to punish the dictators at the cost of reducing their own payoff. Hence, the operationalization of altruistic punishment was whether and how much the recipients punished the dictators in terms of punishment points invested.

We were further interested in whether there were differences in brain activation during punishment acts where individuals might pursue some subjective benefit such as satisfaction through revenge (i.e., punishment for norm violations affecting *one self*, or first person punishment) as compared to punishment acts where individuals seemingly do not pursue any subjective benefit (i.e., punishment for norm violations affecting *other people*, or third party punishment). Hence, we compared two conditions where players B were either the recipients themselves or were “watching” interactions between the dictators and some players C.

Our second question was whether NCd activation would be associated with *effective* rather than with *altruistic* punishment. Hence, we compared two conditions where the punishment was either highly effective (strong punishment, resulting in a substantial reduction of the dictator's payoff) or rather less effective (weak punishment, resulting in marginal reduction).

Finally, we addressed our third question of a possible role of genetic factors in the modulation of neural responses during altruistic behavior by examining whether molecular genetic variation in dopaminergic function might explain part of the variance in neural activation during altruistic punishment. A widely studied genetic variation of dopamine function is a G to A single nucleotide polymorphism in the gene encoding the dopamine-degrading enzyme catechol-O-methyltransferase (COMT) resulting in the substitution of the amino acid valine (Val) by methionine (Met) at amino acid position 158 of the COMT enzyme. The *COMT* Val158Met polymorphism impacts on this enzyme's thermostability: Met allele homozygotes exhibit only ¼ of the COMT activity than Val/Val homozygotes, and hence, presumably have higher levels of synaptic dopamine (Lachman et al., 1996; Chen et al., 2004). The *COMT* Val158Met polymorphism has been associated with prefrontally modulated cognitive and affective processing (for review and meta-analysis see Mier et al., 2009), but has recently also been implicated in reward processing, with Met allele carriers exhibiting higher activation in the ventral striatum and the DLPFC during reward anticipation (Dreher et al., 2009). Hence, the *COMT* Val158Met polymorphism might explain part of the variance in neural activation in the brain regions of interest in the present study.

We hypothesized (1) punishment-related activation in brain regions implicated in altruistic punishment in the literature (i.e., OFC, DLPFC, ACC, insula, NCd) or implicated in reward-related behavior in general (i.e., nucleus accumbens, NAc), with higher activation when subjects punished as compared to trials with no punishment; (2) differences in neural activation related to the players' B perspective, with reward-related areas (especially the NAc) being more strongly activated in the first person perspective, pointing to revenge-like behavioral tendencies, and with areas like the DLPFC and ACC presumably being more strongly activated in the third party perspective, pointing to elevated cognitive control demands and cognitive-affective conflict accompanying the decision to punish defectors in this situation; (3) differences in NCd activation related to the effectiveness of punishment, with higher nucleus caudatus activation for effective (strong) punishment as compared to less effective (weak) punishment; and (4) genetic variation in dopaminergic function to be associated with neural activation during altruistic punishment, with carriers of the *COMT* Met alleles showing stronger NAc and DLPFC activation in punishment vs. no punishment trials.

Materials and methods

Subjects and procedure

Twenty-four students of the University of Giessen (11 women, mean age ± SD: 23.8 ± 3.8 years, age range: 20–34 years) were recruited from the Giessen Gene Brain Behavior Project data bank based on their *COMT* Val158Met genotype, from which we selected an entirely Caucasian sample with almost equally distributed *COMT* genotypes (Val/Val: $n = 9$; Val/Met: $n = 7$; Met/Met: $n = 8$). Subjects were contacted via a telephone interview to rule out the usual MRI exclusion criteria (e.g., non-removable metal parts in/at the body, known neurological diseases, claustrophobia) and subsequently scheduled for the experiment. No further exclusion criteria were used. Upon arrival, subjects were first informed about the study, gave written informed consent, and completed several questionnaires including an assessment of personality characteristics. For the current study, we examined only the Altruism facet scale of the Revised NEO Personality Inventory (NEO-PI-R, Costa and McCrae, 1992). The study was carried out in accordance with the Declaration of Helsinki and both the imaging as well as the genetic parts of the experiment were approved by local ethics committees.

After a demonstration of the paradigm outside the scanner, subjects were placed in the fMRI scanner and performed a short training run to familiarize with the response button device. Measurements began with an anatomical scan, followed by two functional runs of the Dictator Game. In each run, subjects were in the role of a player B who faced 60 decisions of real players A (the "dictators") on how to split a sum of 20 € between themselves and a recipient. In one run, the recipients were the players B themselves (*first person* perspective, FP), whereas in the other run, the recipients were players C (see below), with the players B only "watching" (*third party* perspective, TP). The order of runs was counterbalanced across subjects. After presentation of each player A's decision for 2 s, players B had 6 s to decide whether or not to punish player A for his or her decision by assigning zero to four punishment points. In half of the trials of each run, players B had the opportunity for "strong punishment" (PS), where for every punishment point they invested, 2.50 € were subtracted from player A's outcome, resulting, e.g., in a 10 € reduction of player A's outcome after maximal punishment by investing all 4 punishment points. In the other half of the trials, opportunity for "weak punishment" (PW) was given; here, every punishment point invested by the players B resulted in a subtraction of only 0.5 €, and hence, in a maximal reduction of player A's outcome by 2 €. After each decision, subjects were presented a feedback screen to inform them about the outcome of all players. Then, after a mean inter-trial interval of 9 s (range 7.5 to 10.5 s), the condition of the next trial was revealed for 1 s, followed by the next decision of another player A (see Fig. 1A). After the two runs, which in total lasted about 50 min, subjects were debriefed and paid according to their decisions, i.e., according to their punishment points *not invested* plus a basic reimbursement, resulting in an outcome between 10.50 and 50.50 €. Overall, there were 30 trials each for the four conditions FP/PS, FP/PW, TP/PS, and TP/PW, being composed of the following numbers of player A decisions on dictator : recipient assignments (in €): eleven 10:10, one 11:9, two 12:8, one 13:7, two 15:5, one of each 17:3, 18:2, 19:1, and ten 20:0 assignments.

This distribution of assignments was chosen from a distribution of decisions of real "dictators" who had been asked in advance during the course of a psychology lecture at the University of Frankfurt to decide how they would split a sum of 20 € between themselves and another real person who they did not know and would not knowingly meet again. Further details on the procedure are given in the Supplementary Methods.

fMRI data acquisition

Imaging was performed on a GE Sigma 1.5 T scanner (General Electrics, Milwaukee, WI) using gradient echo planar imaging. 30 axial slices oriented according to AC-PC were acquired in interleaved order (TR = 3000 ms, TE = 50 ms, flip angle = 90°, slice thickness = 5 mm, FOV = 240 mm, matrix = 64 × 64). Prior to the functional imaging, a T1-weighted structural image (172 axial slices, $1 \times 1 \times 1$ mm^3 voxel size) was collected from each subject.

fMRI analyses

Processing of the images was performed using SPM5 (The Wellcome Trust Center for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm/). First, images were slice time corrected and then realigned to the first volume. Thereafter, the images were spatially normalized into a standard stereotactic space (MNI template), resliced to 3 mm isotropic voxels and smoothed with a 6 mm FWHM Gaussian kernel.

Statistical analyses at the first level were conducted for each participant separately by means of a general linear model. The model consisted of two separate sessions (FP, TP). For both sessions the following events were modeled: fair offers, unfair offers, weak punishment, weak non-punishment, strong punishment, strong non-punishment, and feedback. A synthetic hemodynamic response function was used to model brain responses to the particular events. To reduce error variance, the realignment parameters (3 translations, 3 rotations) as derived from preprocessing as well as two noise regressors consisting of time series from white matter and cerebrospinal fluid regions were included into the analyses as covariates. For each participant, contrasts reflecting activation to each event in each session were calculated. These contrasts were then included into a three-way factorial model with 2 (player B perspective: FP vs. TP) × 2 (punishment; punishment vs. non punishment) × 2 (punishment effectiveness: weak vs. strong) second level mixed effects model. We decided not to include the fairness factor into the model because it was confounded with punishment (participants almost always punished during unfair trials and did not punish during fair trials). Separate analyses of the fairness conditions showed that the fairness factor did not substantially explain additional variance. For the second level model, main effects as well as interactions were calculated.

To test for genotype effects on brain activation, difference contrasts were calculated on the first level (punishment vs. non punishment, FP vs. TP, and PS vs. PW). These contrasts were then included into second-level regression analyses using the *COMT* Val158Met genotype as regressor. For all second level analyses, age and gender were used as nuisance covariates.

Statistical interference was conducted in two ways: For exploratory whole brain analysis, we used a threshold of $P<0.0005$ and a cluster size of $k \geq 10$. To test our hypotheses, we used region of interest (ROI) analysis using small volume correction at a threshold of $P<0.05$ false discovery rate (FDR) corrected for multiple testing (in the following denoted as P_{FDR}). In addition, we accounted for a possible underestimation of type I error by FDR-correction and additionally calculated minimum cluster sizes within regions of interest using AlphaSim as implemented in the AFNI software package. AlphaSim corrects for multiple testing using Monte Carlo simulations taking into account the spatial correlation of voxels, the size of the region of interest, and a pre-defined level of significance at the voxel level (Ward, 2000). Cluster sizes were calculated for a predefined P of 0.001 and by running 1000 simulations per ROI. Significance levels for cluster sizes are reported as P_{MC}.

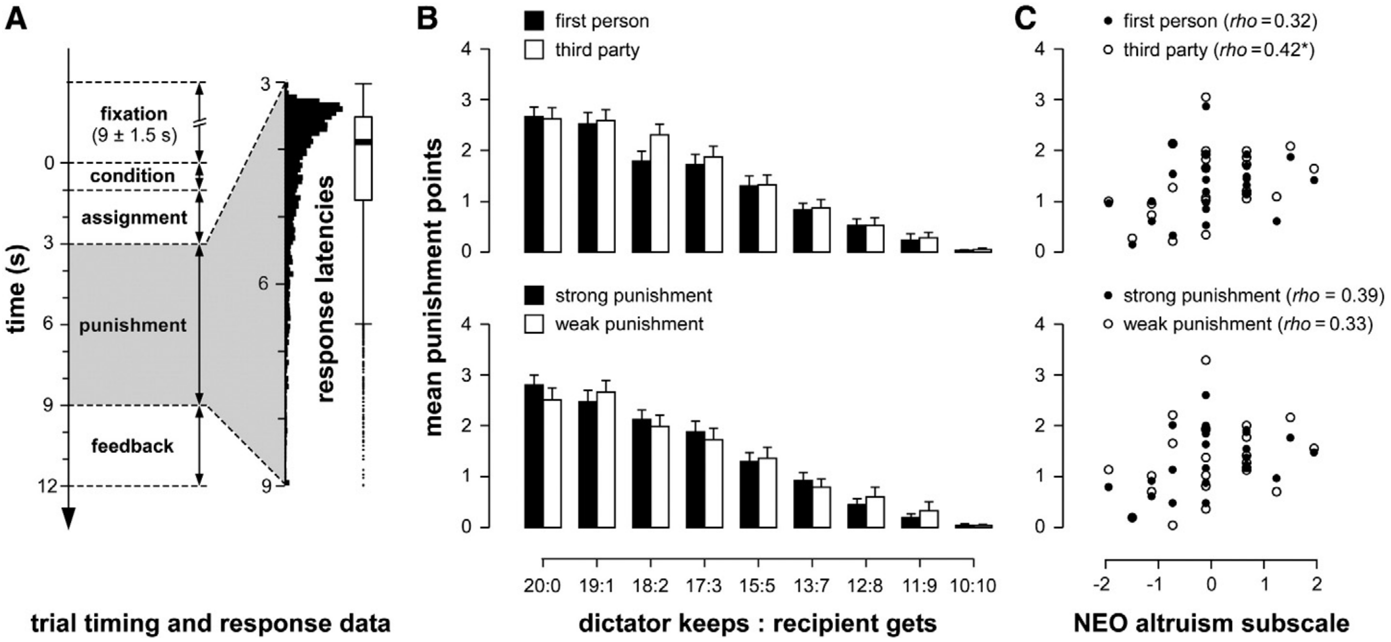


Fig. 1. Summary of the paradigm and behavioral results. (A) Overview over one trial of the Dictator Game used in the present study, together with histogram an box plot of the overall response latencies. (B) Mean punishment points (± S.E.M.) assigned to player A by player B, separately for first person vs. third party conditions (upper panel) and strong vs. weak punishment conditions (lower panel). (C) Correlation between NEO Altruism scores and mean punishment points, separately for first person vs. third party conditions (upper panel) and strong vs. weak punishment conditions (lower panel), together with the results of nonparametric correlation analyses (Spearman's *rho*; note that for NEO Altruism, normalized scores are presented for illustrative purposes only, the normalization does not affect the results of the correlation analyses), $^{*}P<0.05$.

ROIs were defined a priori based on studies on neural correlates of altruistic punishment. Following findings of the fMRI study by Sanfey et al. (2003) on an involvement of the insula (Ins), the cingulate gyrus (CG), and the dorsolateral prefrontal cortex (DLPFC, with the involvement of especially the right DLPFC also being substantiated by findings of a study by Knoch et al., 2006), these brain areas were included as ROIs. The medial OFC was included based on the findings by Koenigs and Tranel (2007). Furthermore, following the results of de Quervain et al. (2004), the nucleus caudatus (NCd) was also included and in addition another reward-related brain area, the nucleus accumbens (NAc). ROI masks were constructed using the SPM toolbox *WFU Pick-Atlas* version 2.4 (Maldjian et al., 2003). Details on ROI construction are given in the Supplementary Methods.

Additional statistical analyses

Additional statistical analyses were mainly carried out using SPSS 17 (SPSS Inc., Chicago, USA), with several pre-processing steps or minor statistical analyses such as correlation analyses being performed using MATLAB (The Mathworks, Natick, MA, USA). To analyze whether there were differences between the experimental conditions in the amount of punishment points invested, nonparametric tests were performed, as due to the low variation in punishment points invested for punishing rather fair offers (i.e., player A : player B assignments of 12:8, 11:9, and 10:10), these variables significantly deviated from the normal distribution (Kolmogorov–Smirnov tests, $P<0.05$). To examine the expected increase of punishment points invested with decreasing fairness of the assignments, the Friedman test was used, with the dependent variables being the means of punishment points invested facing the different levels of player A : player B assignments (9 levels from 10:10 to 20:0), averaged across the conditions “player B perspective” or “punishment effectiveness.” To examine whether there were differences in punishment related to “player B perspective” and “punishment effectiveness,” the Wilcoxon test was used, with the dependent variables being the means of punishment points invested in the first person vs. the third party perspective (FP vs. TP), and in the strong vs. weak punishment conditions (PS vs. PW), respectively, averaged across the different levels of assignments. Two-tailed *P*-values are reported.

To examine influences of individual differences in *COMT* genotype, age, sex, and NEO Altruism on players' B punishment behavior, correlations were calculated between these variables and the mean punishment points across assignments in the FP and the TP conditions, and in the PS and PW conditions, respectively. Nonparametric correlations (Spearmans' *rho*) were used to account for the non-normality of some of the entered variables and for potential outliers.

Results

Behavioral results

Participants indeed invested in punishing the dictators (see Fig. 1B for descriptive results), and did so the more money the dictators kept for themselves (Friedman test: $\chi^2=172.57$, $df=8$, $P<0.001$). There were no differences in the punishment points invested in the strong vs. weak punishment conditions (PS vs. PW; Wilcoxon test: $Z=-0.14$, $P=0.886$), but players B slightly differed in their punishment points invested when facing the dictators' decisions themselves compared to “watching” them interacting with third persons (first person perspective, FP, vs. third party perspective, TP; Wilcoxon test: $Z=-1.92$, $P=0.049$). As descriptively, the latter result seemed to be due to some deviation for 18:2 assignments (see Fig. 1B, upper panel), Wilcoxon tests were performed comparing the punishment points invested in FP and TP conditions separately for the different levels of assignment. Indeed, only for the 18:2 assignment, significantly more punishment points were invested in the TP vs. the FP condition ($P=0.004$ and still significant after Bonferroni correction accounting for the number of tests, i.e., 9; all other $P\geq0.337$).

In order to explain individual differences in players' B punishment behavior, we took into account their *COMT* genotype as well as age, sex, and NEO Altruism and calculated nonparametric (Spearman) correlations between the mean punishment points across assignments in the FP and the TP conditions, and in the PS and PW conditions, respectively. Neither *COMT* genotype nor age and sex were significantly related to the mean punishment points in the different conditions (all $P > 0.20$; the insignificant effect of *COMT* genotype was also confirmed by Kruskal–Wallis tests). However, Altruism correlated significantly with the mean punishment points in the TP condition (Spearman's $rho = 0.42$, $P = 0.041$) and also showed medium-sized, though not or only marginally significant correlations with punishment points in the other conditions (all other $r \geq 0.32$, all $P \leq 0.126$; see Fig. 1C). These results indicate that punishment behavior in our paradigm indeed to some extent reflected (self-reported) habitual altruistic behavioral tendencies, thereby justifying the interpretation of this behavior as “altruistic punishment.” Nevertheless, the medium effect sizes of the relation between NEO Altruism and punishment behavior point to further sources of variation in “altruistic punishment.” Here, the fMRI results on neuronal activation during punishment can be informative.

fMRI results

The main effect *punishment>no punishment* was significant for almost all regions of interest (ROI, see Table 1 and Fig. 2B, left panel), i.e., bilateral cingulate gyrus (CG), DLPFC, insula, NCd, and NAc showed stronger activation in trials where subjects punished as compared to trials where they did not punish. Only for the OFC ROIs, we did not observe punishment- related activation (generally, we did not observe OFC activation differences for any of the contrasts calculated, see Discussion). For the CG ROIs, a differentiation was revealed into anterior and posterior clusters of activation in both hemispheres, corresponding to anterior and posterior cingulate gyrus, while for the right DLPFC ROI, there was also a differentiation into two clusters, one of wide-spread lateral activation, and one of posterior middle frontal gyrus activation near the frontal eye fields, which descriptively was observed also in the left hemisphere, were it was, however, connected with the main lateral cluster of activation. Whole-brain analyses corresponded well with this pattern of activation (see Fig. 2B, right panel), but also pointed to additional clusters of activation, notably parietal regions (intraparietal lobule and precuneus), as well as, among others, medial occipital gyrus, fusiform gyrus, and cerebellum, and thalamus (see Supplementary Table 1 and Supplementary Fig. 1 for a full account). For the reverse contrast *no punishment>punishment*, ROI analyses indicated only one cluster of activation in left posterior insula ($k = 131$, $P_{MC} < 0.001$, peak voxel at -38, -14, 14: $T_{182} = 3.81$, $P_{FWE} = 0.024$, $P_{FDR} = 0.027$, $P_{unc} < 0.001$). Whole-brain analyses confirmed this region, but yielded also additional, wide-spread, but somewhat scattered activation, most prominently in temporal areas, hippocampus/parahippocampal gyrus, and cuneus (see Supplementary Table 1).

For the contrast *FP>TP*, right NAc and bilateral CG exhibited significantly activated clusters, with the latter again showing a differentiation into anterior-to-middle and posterior clusters (see Table 1 and Fig. 2C, left panel). Whole-brain analyses confirmed these clusters of activation (see Fig. 2C, right panel), with the posterior CG extending to the precuneus, and with several smaller clusters of activation in other brain areas (see Supplementary Table 1). For the reversed contrast *TP>FP*, no significant clusters of activation were found in the ROI analyses. In the whole-brain analyses, only occipital clusters of activation reached significance, namely right middle occipital gyrus and bilateral lingual gyrus, perhaps reflecting the more complex computer display in this condition (see Supplementary Table 1 and Supplementary Fig. 1).

For the contrast *PS>PW*, ROI analyses revealed only the NCd to exhibit a significant cluster of activation (see Table 1 and Fig. 2D, left panel). Whole-brain analyses confirmed this region and additionally pointed to significant activation in Brodmann area 10. For the reverse contrast *PW>PS*, no substantial effects were observed (see Supplementary Table 1 and Supplementary Fig. 1).

Both of the ROIs reflecting brain areas implicated in reward processing, i.e., NAc and the NCd, showed significant activation differences not only related to punishment, but also–as for NAc–related to the factor “player B perspective,” and–as for NCd–related to the factor “punishment effectiveness.” Therefore, we examined more closely the activation-deactivation features of these effects, as positive

Table 1
Full factorial model: Regions of interest (ROI) analyses for main effects *punishment>no punishment*, *first person>third party*, and *strong>weak punishment.*

ROI				Peak voxel					
Name		*A priori* size	*Significant voxels k*	T_{182}	P_{FWE}	P_{FDR}	*x*	*y*	*z*
Punishment>no punishment									
Cingulate gyrus	L	3803	582	6.46	<0.001	<0.001	−6	26	36
			249	4.47	0.005	<0.001	0	−34	32
	R	3850	953	6.91	<0.001	<0.001	4	26	40
			436	4.65	0.003	<0.001	2	−34	32
Insula	L	1858	603	6.74	<0.001	<0.001	−40	16	−2
	R	1770	658	6.59	<0.001	<0.001	32	24	−4
DLPFC	L	8430	5090	9.06	<0.001	<0.001	−42	40	28
	R	8654	3852	9.27	<0.001	<0.001	38	8	30
			458	6.05	<0.001	<0.001	32	2	56
N. accumbens	L	70	66	4.52	<0.001	<0.001	−10	8	−2
	R	67	66	4.98	<0.001	<0.001	10	10	−4
N. caudatus	L	564	207	4.85	<0.001	<0.001	−8	4	8
	R	566	130	4.78	<0.001	<0.001	10	10	−2
First person>third party									
Cingulate gyrus	L	3803	157	3.94	0.031	0.023	0	10	34
			241	3.67	0.070	0.023	−6	−44	52
	R	3850	185	4.00	0.025	0.006	2	10	34
			248	4.50	0.004	0.006	14	−34	46
N. accumbens	R	67	42	3.61	0.002	0.005	8	10	−10
Strong punishment>weak punishment									
N. caudatus	R	566	227	4.52	0.001	0.001	12	14	10

Note. Age and sex were entered as covariates into the full factorial model; T_{182} = *T*-statistic and degrees of freedom; P_{FWE} = corrected significance of peak voxel based on family-wise error rate; P_{FDR} = corrected significance of peak voxel based on false discovery rate; only clusters with $P_{FDR} < 0.05$ and $k \geq 10$ significant voxels per cluster are given (all $P_{unc} \leq 0.0002$). All cluster sizes were significant at $P_{MC} < 0.001$; for reverse contrasts, no significant activation was observed in any of the ROIs except for left posterior Insula, were $k = 131$ voxels showed higher activation during no punishment trials vs. punishment trials (peak voxel at coordinates -38, -14, 14: $T_{182} = 3.81$, $P_{FWE} = 0.024$, $P_{FDR} = 0.027$, $P_{unc} = 0.0001$); *x*, *y*, and *z* coordinates of peak voxels are MNI coordinates; DLPFC = dorsolateral prefrontal cortex.

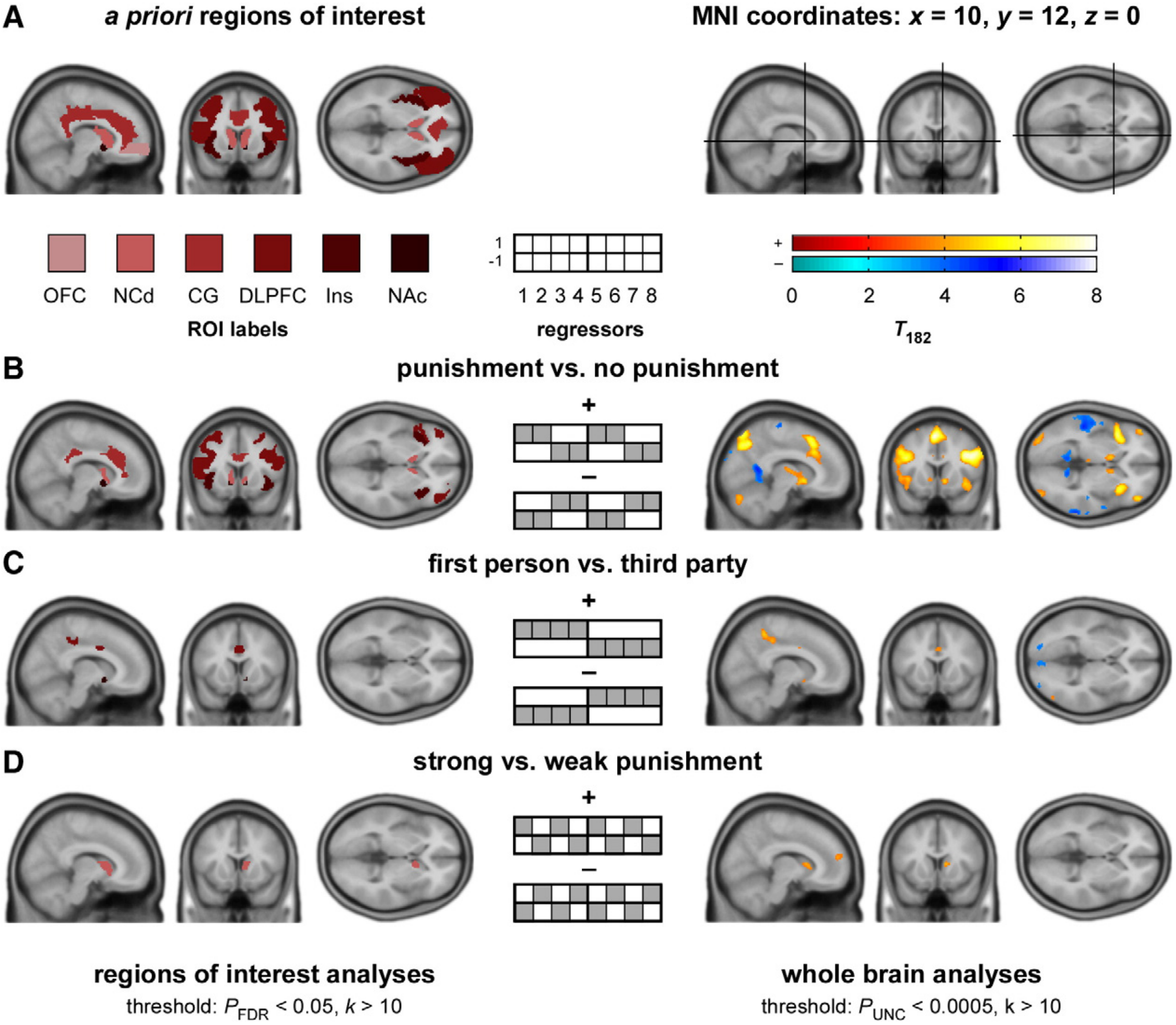

A
a priori regions of interest
MNI coordinates: x = 10, y = 12, z = 0
OFC
NCd
CG
DLPFC
Ins
NAc
ROI labels
1
-1
1 2 3 4 5 6 7 8
regressors
+
–
0
2
4
6
8
T_{182}
B
punishment vs. no punishment
+
–
C
first person vs. third party
+
–
D
strong vs. weak punishment
+
–
regions of interest analyses
threshold: P_{FDR} < 0.05, k > 10
whole brain analyses
threshold: P_{UNC} < 0.0005, k > 10

Fig. 2. fMRI results: regions of interest (ROI) analyses and corresponding whole-brain-analyses. (A) *A priori* masks for ROI analyses and legend to panels B–D. (B–D) Results of ROI analyses (left) and corresponding whole-brain analyses (right) using a full factorial model with the three within-subjects factors (B) *punishment* (regressors 1, 2, 5, and 6) *vs. no punishment* (3, 4, 7, and 8), (C) *first person* (1–4) *vs. third party* (5–8), and (D) *strong punishment* (1, 3, 5, and 7) *vs. weak punishment* (2, 4, 6, and 8); results for all positive and negative main effects are presented at the coordinates given in (A) at a threshold of $P_{FDR}<0.05$, $k \geq 10$ for the ROI analyses, and at $P_{unc}<0.0005$, $k \geq 10$ for the whole-brain analyses; please refer to text for full information; note on image orientation: right is displayed right; see also Supplementary Table 1 and Supplementary Fig. 1.

contrasts may be due to stronger activation in punishment trials, stronger deactivation in no punishment trials, or both. Fig. 3 depicts the comparison between neuronal activation related to “player B perspective” and “punishment effectiveness,” stratified for trials with vs. without punishment (this figure also illustrates the findings of anterior and posterior CG). As can be seen descriptively, prominent NAc activation was found only when subjects were recipients themselves and chose to punish dictators, while NCd activation was especially observed when subjects assigned punishment points in trials where punishment was effective.

However, these effects could not be identified as *interactive* effects in the statistical sense, and generally, interaction effects were sparse throughout the ROI analyses. Only the following two interaction effects reached significance (see Fig. 4): Firstly, activation in a cluster of $k=243$ ($P_{MC}<0.001$) in the left DLPFC revealed an interaction between “punishment” and “player B perspective” (peak voxel at MNI coordinates −40, 22, 46: $T_{182}=4.61$, $P_{FWE}=0.003$, $P_{FDR}=0.006$, $P_{UNC}<0.001$), showing that in the FP condition, there was weaker activation in punished than in not punished trials, whereas for TP, the reverse pattern emerged. Secondly, activation in a cluster of $k=11$ in the right DLPFC ROI showed an interaction between “punishment” and “punishment effectiveness” (peak voxel at 46, 26, 8: $T_{182}=4.32$,

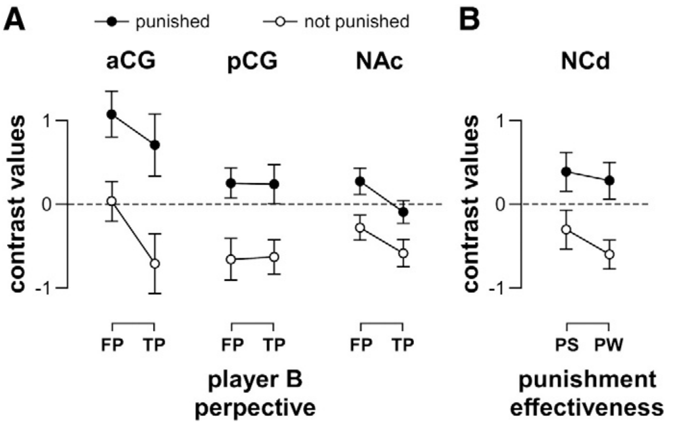


Fig. 3. Comparison between neuronal activation related to player B perspective and punishment effectiveness. Comparison between neuronal activation in trials where players punished (filled circles) as compared to trials without punishment (open circles): (A) main effect of player's perspective: first person (FP) > third party (TP) at the peak voxels of right aCG = anterior cingulate gyrus (MNI coordinates: 2, 10, 34); right pCG = posterior cingulate gyrus (14, −34, 46; for both CG ROIs, results for left hemisphere are analogous); and right NAc = nucleus accumbens (8, 10, −10); (B) main effect of punishment effectiveness: strong punishment (PS) > weak punishment (PW) at the peak voxel of right NCd = nucleus caudatus (12, 14, 10); see Table 1 and text for further details.

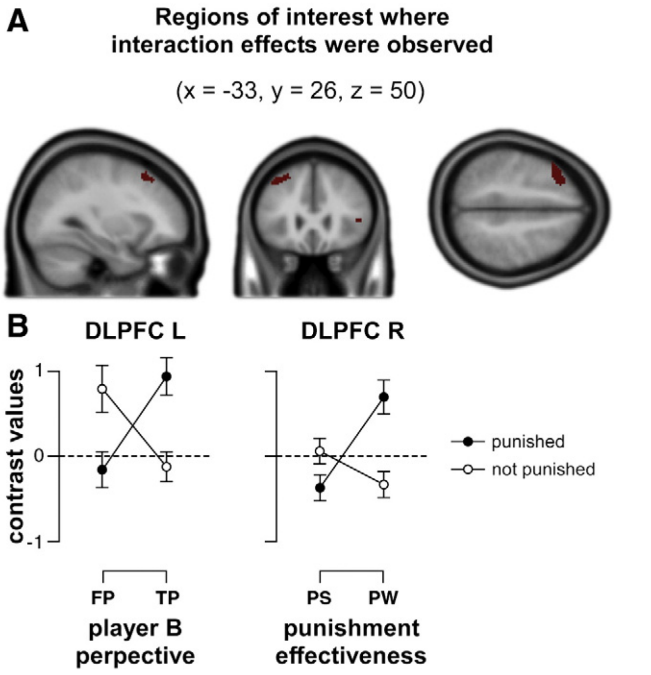


Fig. 4. Regions of interest, where interaction effects were observed. (A) significant clusters of activation for the interaction between player's perspective (first person, FP vs. third party, TP) and punishment (punishment vs. no punishment) in left DLPFC, and for the interaction between punishment effectiveness (strong punishment, PS, vs. weak punishment, PW) and punishment (punishment vs. no punishment) in right DLPFC; (B) Comparison of neuronal activation between trials with punishment (filled circles) as compared to trials without punishment (open circles); mean contrast values (± S.E.M) were derived from the respective peak voxels of left DLPFC (MNI coordinates: −40, 22, 46) and right DLPFC (46, 26, 8); note on image orientation: right is displayed right; see also Supplementary Table 2.

$P_{\mathrm{FWE}} = 0.018$, $P_{\mathrm{FDR}} = 0.029$, $P_{\mathrm{UNC}} < 0.001$), reflecting a strong increase of activation in the PW as compared to the PS condition when subjects punished and a rather slight decrease of activation in the PW as compared to the PS condition when they did not punish. These effects were confirmed in whole-brain analyses (see Supplementary Table 2, with further interactions summarized).

Genetic analyses

Subsequent analyses focused on the impact of variation in dopaminergic function due to genetic variation in the *COMT* gene, namely the Val158Met genotype. An additional inclusion of the *COMT* genotype into the factorial model applied for the examination of the effects described above likely would have resulted in a rather "fragile" model, which would have been quite complicated to evaluate. Moreover, as based on our hypotheses, we were mainly interested in a genetic impact on punishment-related neuronal activity, potential genotype-specific differences in neuronal activation within the *a priori* ROIs were examined using multiple regression with *COMT* genotype (Met/Met>Val/Met > Val/Val) as predictor, the first-level contrast images for the contrasts *punished*>*not punished* as criterion, and age and sex as covariates, again employing a threshold of $P_{\mathrm{FDR}} < 0.05$ and $k \geq 10$. Only for the contrasts *punished*>*not punished*, three of the *a priori* ROIs exhibited significant clusters of activation (see Fig. 5): Firstly, a cluster of $k = 15$ ($P_{\mathrm{MC}} < 0.03$) in the left anterior cingulate gyrus (peak voxel at MNI coordinates -2, 50, 10: $T_{20} = 6.04$, $P_{\mathrm{FDR}} = 0.010$), secondly, a cluster of $k = 52$ ($P_{\mathrm{MC}} < 0.002$) in the right posterior insula (peak voxel at 44, −12, 6: $T_{20} = 5.03$, $P_{\mathrm{FDR}} = 0.021$), and thirdly, a cluster of $k = 15$ ($P_{\mathrm{MC}} < 0.01$) in the right nucleus accumbens (peak voxel at 12, 12, −6: $T_{20} = 3.36$, $P_{\mathrm{FDR}} = 0.011$). No significant effects were observed for other contrasts or for the reverse *COMT* genotype (Val/Val>Val/Met>Met/Met) as predictor.

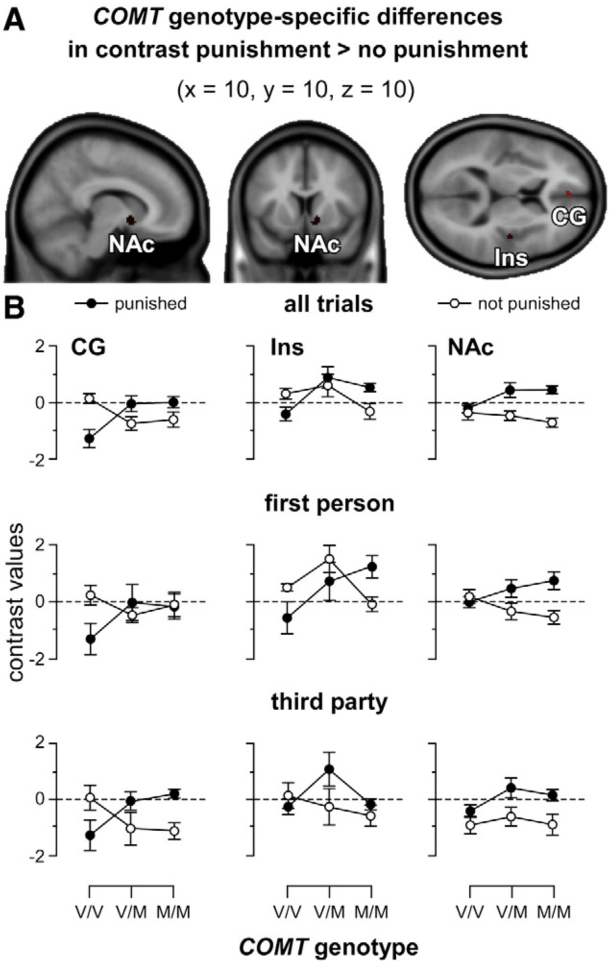


Fig. 5. *COMT* genotype effects. (A) Multiple regression with *COMT* genotype (Met/Met>Val/Met > Val/Val) as predictor (and age and sex as covariates): Regions of interest analyses (threshold: $P_{\mathrm{FDR}} < 0.05$, $k \geq 10$) for contrast *punishment*>*no punishment* revealed significant COMT genotype-specific activation differences in CG = cingulate gyrus (peak voxel coordinates: −2, 50, 10), Ins = Insula (44, −12, 6), and NAc = nucleus accumbens (12, 12, -6); note on image orientation: right is displayed right; (B) Illustration of *COMT* genotype-dependent neuronal activation in punishment trials (filled circles) compared to no punishment trials (open circles), separately for all trials (top), first person perspective (middle), and third party perspective (bottom): for this illustration, the mean contrast values (± S.E.M) of the *full factorial model* were derived from the respective peak voxels of the multiple regression results and were stratified for *COMT* genotype (V = Val, M = Met).

Finally, as sex was included in the regression analyses above, we also examined possible sex effects. However, we did not find such effects with regard to activation differences related to punishment, player B perspective, or punishment effectiveness at a threshold of $P_{\mathrm{FDR}} < 0.05$.

Discussion

In the present study, we addressed the question of the neural and genetic basis underlying strongly reciprocal behavior such as altruistic punishment in more detail by employing fMRI during a one-shot economic exchange game. This enabled us to contrast conditions of strong reciprocity with and without personal involvement by comparing first person with third party punishment. As recent evidence on the activation of brain reward regions during altruistic punishment suggests satisfaction through the punishment of norm violations as one underlying motive for altruistic punishment, we specifically concentrated on reward-related regions, but also took into account further brain regions implicated in cognitive and affective processing during social interactions. The results of our investigation will be summarized and discussed along the lines of our *a priori* hypotheses.

Punishment-related activation in cognition-, emotion-, and motivation-related regions

Our behavioral data indicate that punishment behavior covaried with the unfairness of the dictators' assignments, but was rather independent from punishment effectiveness or personal involvement (although it has to be noted that we observed significantly higher third party than first person punishment for the 18:2 assignment, the reasons for which remain speculative until replication of this punishment pattern). Importantly, we observed medium-sized positive correlations between punishment points invested by our participants and their Altruism scores as assessed using the NEO-PI-R. It has to be noted, that only the correlation between Altruism and third party punishment reached significance, but exactly in the third party condition, altruistic behavioral tendencies should manifest most clearly. Hence, this result supports the notion that punishment behavior in economic exchange games indeed covaries to some extent with self-reported altruistic behavioral tendencies, and hence, can be used as an experimental operationalization of altruistic behavior.

Similar to the behavioral results, the fMRI data show that altruistic punishment involves comparable processes irrespective of personal involvement or punishment effectiveness. Almost all ROIs hypothesized to exhibit punishment-related activation patterns–DLPFC, (especially dorsal) ACC, insula, NCd, and NAc, see our first hypothesis–showed stronger activation during punishment trials than during trials without punishment. Only for the OFC ROIs, we did not observe punishment-related activation (nor did we observe OFC activation for any other contrast). Although image acquisition covered the entire brain, EPI distortion artifacts were quite severe adjacent to perinasal sinuses (see exemplary EPI images in Supplementary Figure 2). On the used hardware, we were unable to employ better sequences like z-shim GE-EPI (see Weiskopf et al., 2006). Hence, it is likely that the lack of OFC activation might be related to the observed signal loss in this area.

While we modeled the fairness of the assignments in separate regressors (see Materials and methods), we cannot rule out that some of the observed activation during punishment may result from a temporal overlap with processes of fairness evaluation. Nevertheless, the interpretation of the observed activation patterns as correlates of punishment-related processes includes aspects of fairness evaluation, and future studies might be able to examine the complexity of these processes in more detail.

Concerning the functional roles of the regions identified here as punishment-related, there is a vast body of evidence that implicates the DLPFC in the implementation of executive, or cognitive control as well as in the temporal organization of goal-directed action (e.g., Fuster, 2000; Miller and Cohen, 2001). The ACC has been proposed to be involved in monitoring the need for control implementation to promote behavioral adjustments in the presence of conflict (e.g., Botvinick et al., 2001). In addition, it has been suggested that different divisions of the ACC are more strongly related to emotional conflict (ventral/subgenual), and to cognitive conflict (dorsal), respectively (Bush et al., 2000). Insular activation can be assumed to reflect representations of (in the present context presumably negative) emotional states (Singer et al., 2009) and norm violations (Montague and Lohrenz, 2007), while neuronal activation in ventral (i.e., NAc) and dorsal (i.e., NCd) striatal areas are likely to be associated with the reward-oriented integration and transformation of these input signals into motor outputs (see e.g., Depue and Collins, 1999; Robbins and Everitt, 1996). It could therefore be speculated that insular representations of negative emotional states due to norm violations would provide a bias signal, which interferes with signals of immediate individual financial reward if no punishment is exerted. The resulting behavioral conflict, monitored and signaled by the ACC, would result in DLPFC-mediated implementation of cognitive control, which would impact on the striatal integration of the input signals in favor of the decision to punish, given that more future reward would be anticipated following such behavior due to learned contingencies between norm-conform or norm-enforcing behavior and social rewards.

Interestingly, this interpretation resembles a recent model of instrumental punishment put forward by Seymour and colleagues (Seymour et al., 2007), which aims at explaining also reciprocity-based punishment. Of course, there are also notable differences (e.g., we could not focus on OFC regions due to technical reasons, and the mentioned model does not include the ACC), and moreover, any interpretation of the activation pattern observed here (including alternative accounts, see end of this section) needs to be tested more thoroughly in future studies, ideally also by focusing on the *effective* connectivity between the mentioned regions (e.g., by means of Granger Causality mapping, Goebel et al., 2003), an issue, which again we could not address adequately due to technical limitations associated with the use of a 1.5 T scanner. Nevertheless, our results further underscore the complexity of the interplay of cognitive, emotional, and motivational processes and their neural correlates during social decision-making.

Interestingly, there were no interactions between punishment and personal involvement, or punishment effectiveness, respectively, with the exception of two subregions of the DLPFC: first, an interaction effect between punishment and personal involvement was present in a more superior cluster in the left DLPFC (see Fig. 4), with stronger activation during both first person *non*-punishment and third party punishment. This result could indicate that this region is involved in cognitive control processes, which modulate the decision not to punish defective behavior towards oneself (i.e., not to pursue revenge-like behavior), but to punish defective behavior towards other individuals (i.e., to behave altruistically in a psychological sense). Second, a more inferior cluster in the right DLPFC showed stronger activation during punishment trials with weak punishment (see Fig. 4). This activation pattern could reflect cognitive control processes necessary to overcome the tendency not to punish defective behavior when punishment has only a limited impact. However, further evidence and replication is necessary before more definite conclusions can be drawn.

Activation differences related to punishers' perspective and punishment effectiveness

In our second hypothesis, we expected that reward-related areas (especially the NAc) would be more strongly activated in the first person perspective, pointing to revenge-like behavioral motives, while DLPFC and ACC would presumably be more strongly activated in the third party perspective due to the assumed volitional and conflict-laden nature of the decision to punish unfair behavior without being directly affected by such behavior. We could not confirm the latter part of this hypothesis—in fact, two anterior, and posterior cingulate subregions, respectively, were more strongly activated in the first person perspective. However, we observed that indeed, the NAc was more strongly activated during first person punishment as compared to third party punishment. Furthermore, we hypothesized differences in NCd activation related to the effectiveness of punishment, with higher activation for strong punishment as compared to weak punishment. As expected, only in the NCd ROI, an effect of punishment effectiveness could be observed. These patterns of activation were generally confirmed by whole-brain analyses. These results suggest that reward-related regions are more involved when the punisher is directly affected (NAc) or when punishment has a strong effect (NCd). Hence, our study supports the general interpretation of altruistic punishment as rewarding (De Quervain et al., 2004), but in addition enables us to specify that this should be even more the case for highly effective punishment in direct interactions. While this may at first glance emphasize revenge-like motives as driving forces of altruistic punishment, it is necessary to point out that both the NAc and the NCd were significantly involved in punishment *per se* (i.e., the contrast punishment vs. no punishment), with no statistical interaction with personal involvement or punishment effectiveness. Therefore, the activation patterns of these reward-related regions suggest only quantitative differences in their activation during punishment in different conditions, and hence, a common motivational mechanism underlying the punishment of defectors, perhaps arising from learned contingencies between norm-abiding behavior and social rewards.

Furthermore, the shared and almost equally strong activations observed for both conditions, first person and third party, point to the possibility that altruistic punishment may not mainly be driven by personal involvement, but could share common emotional or cognitive components. Common emotional processes that might prompt altruistic punishment could be anger about unfair offers or enjoyment of revenge. Activations observed in limbic regions like the insula or the striatum might support this assumption. From a cognitive point of view, in both tasks the salience of losses associated with unfair offers might be stronger than that of losses related to the own punishment decision. Clearly, such an interpretation remains speculative, but it would be worthwhile to test potential variations of loss aversion in a future study.

Finally, the fact that both conditions produced activations in a common network could alternatively be explained by a simulation process. One could argue that observing someone else being treated unfairly could initiate a simulation process which requires the same network as the processing of an unfair offer itself. It has been shown that the representation of the self and others is processed in a common representation network (Decety and Sommerville, 2003) and that observing someone else in pain leads to activation in the same areas as experiencing pain (Singer et al., 2004). Interestingly some of the regions found in this study in the context of empathy for pain, namely the insula and the cingulate cortex, were activated in the present study as well. Therefore, the perspective-independent activation pattern observed in the present study could also reflect such a representation network.

Genetic variation in dopamine function associated with punishment-related activation

Based on the results of Dreher et al. (2009), we further hypothesized that carriers of the *COMT* Met alleles showed stronger NAc and DLPFC activation in punishment trials as compared to trials without punishment. For the DLPFC, we could not identify such genotypic differences in neural activation, which at first glance is intriguing given the strong evidence for *COMT* genotype effects on prefrontal activation, but may be explained by the fact that the decision-making process in our paradigm required both cognitive and emotional control processes, for which opposing genotypic effects have been demonstrated (for a comprehensive overview, see Mier et al., 2009).

However, in line with our hypothesis, we found an effect of the *COMT* genotype on punishment-related neural activation in the NAc. While we also observed genotypic effects in two clusters in anterior cingulate and the right insula, the activation patterns of these two clusters were rather difficult to interpret (see Fig. 5). In contrast, the NAc showed a clear activation pattern, with carriers of at least one *COMT* Met allele, which has been associated with reduced COMT enzyme function and presumably elevated synaptic dopamine availability (Lachman et al., 1996; Chen et al., 2004), exhibiting higher punishment-related NAc activation. It is important to stress that due to the comparably small sample size for molecular genetic studies, these results have to be regarded as preliminary. In fact, given our sample size, a nominal alpha level of 0.05, and a power of 0.80, only strong effects could be detected. Although such strong effects cannot be readily expected for genetic variations even at the neuronal level, the location and direction of the observed effect in our view underscore the validity of our finding. Furthermore, it corresponds with recent findings from other groups. Comparable Met allele-related activation patterns in the NAc–or more general: the ventral striatum–have been observed in several studies examining neural activation during reward anticipation (Dreher et al., 2009; Yacubian et al., 2007), although conflicting evidence exists (Forbes et al., 2009; Schmack et al., 2008). Interestingly, the *COMT* Met allele was recently also shown to be associated with a higher ability to experience reward in daily life (Wichers et al., 2008), further strengthening its role in reward anticipation. Hence, *COMT* Met allele-related stronger activation in the ventral striatum due to elevated synaptic dopamine availability may bias ventral striatal integration of input signals from DLPFC, ACC, insula, and other regions based on a higher reward anticipation associated with the decision to punish unfair behavior.

This interpretation would require, as mentioned above, that individuals have experienced contingencies between norm-conform or norm-enforcing behavior and social rewards, with *COMT* Met allele carriers being more inclined to adapt their behavior in order to attain such rewards. Our results could therefore be interpreted as a first empirical support for the assumption of gene–culture co-evolution (Fehr and Fischbacher, 2004; Gintis, 2003), with *COMT* Met allele carriers being particularly susceptible to social signals of reward. Given that the *COMT* Met allele appears to be the evolutionary recent allele unique to humans (Palmatier et al., 1999), a gene–culture coevolution–favoring the Met allele in societies where strongly reciprocal behavior is the social norm–might explain why strongly reciprocal behavior seems to be shown mainly by humans.

Alternative explanations

While our data provide further evidence on brain regions activated during altruistic punishment, and thus, may aid the discussion on the motives underlying altruistic punishment, we cannot directly address this issue due to the lack of emotion ratings. This is a clear limitation of the present study, as this would have informed us whether our participants' felt anger (as we expected at least fort he first person condition), dominance (which could have played a more important role than anger especially in the third party condition) or rather other emotions such as disgust towards the dictators.

Indeed, disgust may offer an alternative interpretation for part of our findings. The observed insula activation during punishment may point in this direction, as the insula has been implicated in processing of disgust, just as–interestingly–the NCd (Calder et al., 2000). Another model of insula function by Craig (2009) integrates ACC and insula into a complementary system involved in voluntary motivation, or agency, and interoception of bodily conditions, which together constitute emotions. Together with evidence for direct and indirect connections between ACC and NCd, and also between NCd and DLPFC within parallel basal ganglia thalamocortical loops (Alexander et al., 1986), this could lead to the speculation that it is the mere interoceptive awareness of disgust due to the violation of social or moral norms (see also Schnall et al., 2008) as represented in the insula, which due to the close connection with the ACC gives rise to direct as well as indirect signals (via the NCd and thalamus, which indeed was also activated during punishment, see Fig. 2 and Supplementary Table 1) from the ACC to DLPFC, that encompass and/or generate voluntarily motivated sequential action. It has to be noted that at least for the NCd, our observation of higher activation for effective as compared to ineffective punishment just as in the de Quervain et al. (2004) study in our view favors the interpretation that the NCd in our paradigm was rather involved in reward processing. Nevertheless, the sketched alternative explanation for the punishment-related activation observed here has its appeal. Again, however, it is important that this interpretation requires measures of effective connectivity, an issue which future studies should address.

Conclusion

Taken together, our behavioral and fMRI data show that altruistic punishment correlates with self-reported altruistic tendencies and involves comparable neural processes irrespective of personal involvement or punishment effectiveness, supporting the assumption of its important role for the retention of cooperation in human societies. Highlighting the complex interplay of cognitive-affective processes supported by DLPFC and ACC with additional interoceptive-emotional and motivational bias signals conveyed by insula and striatum during social decision-making, our data provide further evidence on the neural bases of altruistic punishment and may thus aid the discussion of the motives underlying strongly reciprocal behavior. Notably, brain regions previously implicated in reward processing are more involved when punishment has a strong effect (NCd) or the punisher is directly affected (NAc), with neural activity in the latter region being associated with genetic variation in dopamine function. These results support the notion of altruistic punishment as a human trait that has developed via gene–culture coevolution due to its rewarding properties, which should be more pronounced when punishment results from violations of personal interest or when it has a strong effect.

Acknowledgments

The authors thank Karina Zygrodnik and Carina Sauer for their help during data acquisition, and Jürgen Hennig for access to the Giessen Gene Brain Behavior Project data bank.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.07.051.

References

Alexander, G.E., DeLong, M.R., Strick, P.L., 1986. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu. Rev. Neurosci. 9, 357–381.

Bechara, A., Damasio, H., Damasio, A.R., 2000. Emotion, decision making, and the orbitofrontal cortex. Cereb. Cortex 10, 295–307.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108, 624–652.

Bowles, S., Gintis, H., 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor. Popul. Biol. 65, 17–28.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Natl. Acad. Sci. U.S.A. 100, 3531–3535.

Bush, G., Luu, P., Posner, M.I., 2000. Cognitive and emotional influences in anterior cingulate cortex. Trends Cogn. Sci. 4, 215–222.

Calder, A.J., Keane, J., Manes, F., Antoun, N., Young, A.W., 2000. Impaired recognition and experience of disgust following brain injury. Nat. Neurosci. 3, 1077–1078.

Chen, J., Lipska, B.K., Halim, N., Ma, Q.D., Matsumoto, M., Melhem, S., Kolachana, B.S., Hyde, T.M., Herman, M.M., Apud, J., Egan, M.F., Kleinman, J.E., Weinberger, D.R., 2004. Functional analysis of genetic variation in COMT: effects on mRNA, protein, and enzyme activity in postmortem human brain. Am. J. Hum. Genet. 75, 807–821.

Costa, P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Psychological Assessment Resources, Odessa, FL.

Craig, A.D., 2009. How do you feel — now? The anterior insula and human awareness. Nat. Rev. Neurosci. 10, 59–70.

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.

Decety, J., Sommerville, J.A., 2003. Shared representations between self and other: a social cognitive neuroscience view. Trends Cogn. Sci. 7, 527–533.

Delgado, M.R., Locke, H.M., Stenger, V.A., Fiez, J.A., 2003. Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. Cogn. Affect. Behav. Neurosci. 3, 27–38.

Depue, R.A., Collins, P.F., 1999. Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation, and extraversion. Behav. Brain Sci. 22, 491–569.

Dreher, J.-C., Kohn, P., Kolachana, B., Weinberger, D.R., Berman, K.F., 2009. Variation in dopamine genes influences responsivity of the human reward system. Proc. Natl. Acad. Sci. U.S.A. 106, 617–622.

Fehr, E., Fischbacher, U., 2004. The nature of human altruism. Nature 425, 785–791.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Forbes, E.E., Brown, S.M., Kimak, M., Ferrell, R.E., Manuck, S.B., Hariri, A.R., 2009. Genetic variation in components of dopamine neurotransmission impacts ventral striatal reactivity associated with impulsivity. Mol. Psychiatry 14, 60–70.

Fuster, J.M., 2000. Executive frontal functions. Exp. Brain Res. 133, 66–70.

Gintis, H., 2003. The hitchhiker's guide to altruism: gene–culture coevolution, and the internalization of norms. J. Theor. Biol. 220, 407–418.

Goebel, R., Roebroeck, A., Kim, D.-S., Formisano, E., 2003. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magn. Reson. Imaging 21, 1251–1261.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E., 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314, 829–832.

Koenigs, M., Tranel, D., 2007. Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. J. Neurosci. 27, 951–956.

Lachman, H.M., Papolos, D.F., Saito, T., Yu, Y.M., Szumlanski, C.L., Weinshilboum, R.M., 1996. Human catechol-O-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. Pharmacogenetics 6, 243–250.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage 19, 1233–1239.

Mier, D., Kirsch, P., Meyer-Lindenberg, A., 2009. Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis. Mol. Psychiatry [Epub ahead of print].

Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Ann. Rev. Neurosci. 24, 167–202.

Montague, P.R., Lohrenz, T., 2007. To detect and correct: norm violations and their enforcement. Neuron 56, 14–18.

Ohtsuki, H., Iwasa, Y., Nowak, M.A., 2009. Indirect reciprocity provides a narrow margin of efficiency for costly punishment. Nature 457, 79–82.

Palmatier, M.A., Kang, A.M., Kidd, K.K., 1999. Global variation in the frequencies of functionally different catechol-O-methyltransferase alleles. Biol. Psychiatry 46, 557–567.

Robbins, T.W., Everitt, B.J., 1996. Neurobehavioral mechanisms of reward and motivation. Curr. Opin. Neurobiol. 6, 228–236.

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.

Schmack, K., Schlagenhauf, F., Sterzer, P., Wrase, J., Beck, A., Dembler, T., Kalus, P., Puls, I., Sander, T., Heinz, A., Gallinat, J., 2008. Catechol-O-methyltransferase val158met genotype influences neural processing of reward anticipation. NeuroImage 42, 1631–1638.

Schnall, S., Haidt, J., Clore, G.L., Jordan, A.H., 2008. Disgust as embodied moral judgment. Pers. Soc. Psychol. Bull. 34, 1096–1109.

Schultz, W., 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1–27.

Seymour, B., Singer, T., Dolan, R., 2007. The neurobiology of punishment. Nat. Rev. Neurosci. 8, 300–311.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D., 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303, 1157–1162.

Singer, T., Critchley, H.D., Preuschoff, K., 2009. A common role of insula in feelings, empathy and uncertainty. Trends Cogn. Sci. 13, 334–340.

Ward, B.D., 2000. Simultaneous inference for fMRI data. Online AFNI manual: http://afni.nimh.nih.gov/pub/dist/doc/manual/AlphaSim.pdf (last access: 2010-06-24).

Weiskopf, N., Hutton, C., Josephs, O., Deichmann, R., 2006. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage 33, 493–504.

Wichers, M., Aguilera, M., Kenis, G., Krabbendam, L., Myin-Germeys, I., Jacobs, N., Peeters, F., Derom, C., Vlietinck, R., Mengelers, R., Delespaul, P., van Os, J., 2008. The Catechol-O-methyltransferase Val158Met polymorphism and experience of reward in the flow of daily life. Neuropsychopharmacology 33, 3030–3036.

Yacubian, J., Sommer, T., Schroeder, K., Gläscher, J., Kalisch, R., Leuenberger, B., Braus, D.F., Büchel, C., 2007. Gene–gene interaction associated with neural reward sensitivity. Proc. Natl. Acad. Sci. U.S.A. 104, 8125–8130.