Behavioral/Cognitive

Neurobiological Mechanisms of Responding to Injustice

Mirre Stallen,[1,2,4] **Filippo Rossi,**[3] **Amber Heijne,**[1] **Ale Smidts,**[2] **Carsten K.W. De Dreu,**[4,5] **and Alan G. Sanfey**[1,6]

[1]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, 6500 HB, The Netherlands, [2]Rotterdam School of Management, Erasmus University, Rotterdam, 3062 PA, The Netherlands, [3]Institute for Neural Computation, University of California, San Diego, California, [4]Institute of Psychology, Leiden University, Leiden, 2300RA Leiden, The Netherlands, [5]Center for Experimental Economics and Political Decision-Making, University of Amsterdam, Amsterdam, 1018 WB, The Netherlands, and [6]Behavioral Science Institute, Radboud University, Nijmegen, 6525 HR, The Netherlands

People are particularly sensitive to injustice. Accordingly, deeper knowledge regarding the processes that underlie the perception of injustice, and the subsequent decisions to either punish transgressors or compensate victims, is of important social value. By combining a novel decision-making paradigm with functional neuroimaging, we identified specific brain networks that are involved with both the perception of, and response to, social injustice, with reward-related regions preferentially involved in punishment compared with compensation. Developing a computational model of punishment allowed for disentangling the neural mechanisms and psychological motives underlying decisions of whether to punish and, subsequently, of how severely to punish. Results show that the neural mechanisms underlying punishment differ depending on whether one is directly affected by the injustice, or whether one is a third-party observer of a violation occurring to another. Specifically, the anterior insula was involved in decisions to punish following harm, whereas, in third-party scenarios, we found amygdala activity associated with punishment severity. Additionally, we used a pharmacological intervention using oxytocin, and found that oxytocin influenced participants' fairness expectations, and in particular enhanced the frequency of low punishments. Together, these results not only provide more insight into the fundamental brain mechanisms underlying punishment and compensation, but also illustrate the importance of taking an explorative, multimethod approach when unraveling the complex components of everyday decision-making.

Key words: compensation; computational modeling; neuroimaging; oxytocin; punishment; social norms

Significance Statement

The perception of injustice is a fundamental precursor to many disagreements, from small struggles at the dinner table to wasteful conflict between cultures and countries. Despite its clear importance, relatively little is known about how the brain processes these violations. Taking an interdisciplinary approach, we combine methods from neuroscience, psychology, and economics to explore the neurobiological mechanisms involved in both the perception of injustice as well as the punishment and compensation decisions that follow. Using a novel behavioral paradigm, we identified specific brain networks, developed a computational model of punishment, and found that administrating the neuropeptide oxytocin increases the administration of low punishments of norm violations in particular. Results provide valuable insights into the fundamental neurobiological mechanisms underlying social injustice.

Introduction

People typically have strong reactions to what they perceive as injustice. In fact, violations of social norms, standards of behavior that are based on widely shared beliefs about how one should behave, not only trigger an emotional response in the "victims" of such violations, but can also lead to strong responses from third-parties observing the situation (Fehr and Fischbacher, 2004; Chavez and Bicchieri, 2013).

Recent neuroimaging work has greatly advanced our understanding of how the brain responds to social norm violations. These studies highlight the role of reward- and emotion-related processes in punishment as well as in compensatory decision-making, where the latter refers to the allocation of resources to someone who has been the victim of a violation (Sanfey et al., 2003; Treadway et al., 2014). To date, however, most research has focused on punishment and very few studies have directly compared punishment and compensation decisions at the neural level (Hu et al., 2015, 2016). Therefore, the extent to which the biological mechanisms underlying punishment and compensation are either similar to, or different from, each other remains elusive. Our first aim, accordingly, is to unravel the neural mechanisms underlying decisions to punish transgressors and compensate victims.

Received May 5, 2017; revised Dec. 26, 2017; accepted Jan. 4, 2018.

Author contributions: M.S., A.S., C.D.D., and A.G.S. designed research; M.S. and A.H. performed research; M.S. and F.R. analyzed data; M.S., A.S., C.D.D., and A.G.S. wrote the paper.

This work was supported by a Grant from the European Research Council (ERC313454) and support from the Erasmus Research Institute of Management (ERIM). We would like to thank Jason Roos for his helpful input.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Mirre Stallen, Leiden University, Leiden, P.O. Box 9555, 2300RA Leiden, The Netherlands. E-mail: m.stallen@fsw.leidenuniv.nl.

DOI:10.1523/JNEUROSCI.1242-17.2018

The next aim of this study focuses more closely on punishment of transgressors. Recent neuroscientific work suggests that punishment is not a single, unitary, process, but rather comprises distinct subcomponents (Krueger and Hoffman, 2016). However, studies to date have predominantly focused on disentangling these subcomponents in so-called third-party punishment situations where others are in the role of victims, and participants administer punishments to the transgressor. Research examining second-party punishment, where participants are directly affected, is exceedingly scarce (Strobel et al., 2011). This aim therefore will compare the neural processes underlying both second- and third-party punishment, and specifically identify the motives underlying decisions to punish and, subsequently, of how much to punish, across both types of contexts.

Understanding the motives underlying these two decision components—whether, and how much, to punish—is nontrivial. One account of punishment in response to unfairness posits that punishment is driven by social preferences for equitable outcomes (Fehr and Schmidt, 1999). However, alternative accounts suggest that affective experiences, such as anger, envy, or retaliation, are key drivers in sanctioning norm violators (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Pedersen et al., 2013). Although these motives are not mutually exclusive (i.e., violations of fairness can induce emotional arousal), neuroimaging evidence can help in formalizing the specific role of fairness preferences and negative emotions in punishment. Indeed, previous studies suggest that these motivational processes are linked with different brain regions. Perceived unfairness has been consistently associated with activation of anterior insula (Corradi-Dell'Acqua et al., 2013; Zhong et al., 2016), whereas the amygdala has been associated with negative arousal (Gospic et al., 2011; Treadway et al., 2014). To disentangle the role of fairness preferences and negative arousal in punishment, we developed a computational model of punishment decision-making. This formal model allowed us to estimate individual parameter values reflecting individuals' willingness to punish, as well as the subsequent severity of their punishment.

Additionally, there has been much recent interest in the role of hormones in cognition (Rilling et al., 2012), in particular oxytocin. Many studies first demonstrated oxytocin as promoting specifically prosocial behavior, but recent work has suggested a more complex role (Ma et al., 2016), potentially also relevant to negative, punishment, behavior (Krueger et al., 2013; Daughters et al., 2017). Therefore, our final aim was to explore the role of oxytocin in both compensation and punishment contexts in response to norm violations. We were particularly interested in whether oxytocin impacted punishment differently in second- and third-party contexts, and what neural mechanism might underlie these effects. If oxytocin impacted responses to norm violations, this would suggest that oxytocin is involved in social behavior more generally, potentially playing a role in shifting focus from self-interest to group norms (De Dreu and Kret, 2016).

Materials and Methods

To achieve these multifaceted goals, we used an exploratory and multi-methodological approach. We combined functional magnetic resonance imaging (fMRI) with a pharmacological intervention of oxytocin, and developed a novel experimental decision-making task termed the Justice Game. We recruited 55 males (28 oxytocin) for a randomized, double-blind, placebo-controlled between-subject experimental design. One participant in the oxytocin group was excluded because he reported not performing the task seriously, and therefore behavioral analyses were conducted with 54 participants (mean age = 21.2 years, SD = 2.4 years; 27 oxytocin). Neuroimaging data and computational model analyses were conducted with 53 participants (27 oxytocin), as neuroimaging data of one participant in the oxytocin group was lost due to technical problems.

Procedure

Upon arrival at the laboratory, participants provided informed consent and received instructions about the task. Participants received a standard payment of €35 for participation plus an additional bonus payment of between €7 and €15 based on their decisions (average earnings of €47, SD = €1.29). Exclusion criteria for participation were significant medical or psychiatric illness, medication, smoking more than five cigarettes per day, or drug/alcohol abuse. Participants were instructed to refrain from smoking, eating, and drinking (except for water) for 3 h before the experiment, and all were tested in the afternoon to minimize effects of circadian rhythm. The experiment was approved by the local university ethics committee.

At the start of the experiment, participants completed several practice trials, and were subsequently quizzed to check whether instructions had been understood correctly. Next, they self-administered the medication (Syntocinon spray, Novartis; 24 IU; 3 puffs per nostril, each with 4 IU of either oxytocin or placebo) under experimenter supervision. To avoid pharmacological effects other than those caused by oxytocin, the placebo contained all active ingredients with the exception of the neuropeptide. While in the MRI scanner, participants completed one unrelated task before beginning the present task. About 40 min (M = 41.5, SD = 21.7) after substance administration (depending on how quickly participants completed the previous task), participants started with the experiment of interest. This timing conforms to other studies showing that effects of oxytocin are present between 30 and 60 min after intranasal administration (Baumgartner et al., 2008; De Dreu et al., 2010, 2011; Stallen et al., 2012). At both the beginning and the end of the experiment (after oxytocin/placebo administration and task completion), participants completed a questionnaire to assess their perceptions of fairness and to enhance the credibility of the task. Participants were told that the actions they witnessed in the experiment reflected questionnaire answers given by fellow participants whose final payment depended on participants' choices in the experiment.

The Justice Game

The Justice Game allows us to measure participants' responses to norm violations concerning distributional fairness in terms of both punishment and compensation, and was specifically designed for use in an MRI environment (Fig. 1). In contrast to studying individuals' willingness to reject unfair offers, as is often done in the study of punishment behavior (Feng et al., 2015), the Justice Game examined how people perceive and respond to a situation in which a third-party deliberately (and unjustly) takes resources from either oneself or someone else. Experimental work in economics has shown that this deliberate act elicits very negative emotions in the person from whom the resources are taken, and that this is typically experienced as a stronger norm violation than receiving an unfair share of resources (Bosman and Van Winden, 2002; Ben-Shakhar et al., 2007; Korenok et al., 2014), hence our use of this manipulation.

The Justice Game is a single-shot game involving two players, one of whom is randomly assigned the role of Taker. At the start of the game, both the Taker and the Partner receive an endowment of 200 chips. The Taker is then given the opportunity to take chips from the Partner and add these chips to his own endowment, or he can choose to leave the distributions as they are. The Taker can take a maximum of 100 chips from the Partner. However, after the Taker has indicated his decision, the Partner is given the option of punishing the Taker by spending chips of his own. For each chip the partner spends, the income of the Taker is reduced by 3 chips. Thus, punishment in this task is both costly and effective. The Partner can use a maximum of 100 chips from his own endowment for punishment purposes.

Participants played three conditions of the Justice Game: second-party punishment games, third-party punishment games, and third-party compensation games. The second-party punishment games were identical to the Justice Game as described above, with participants always assigned to the role of Partner. In third-party punishment games, participants were assigned the role of Observer, where they watched a Taker

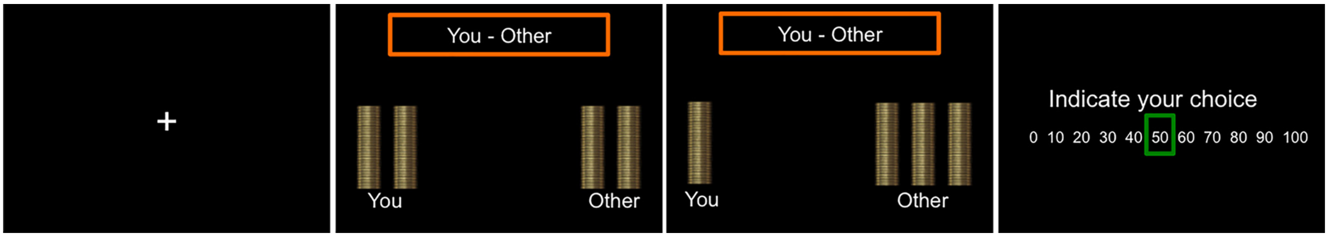

Figure 1. Trial outline of a second-party punishment game in the Justice Game. In this second-party punishment sample trial, a Taker takes 100 chips from the participant and the participant can decide how much, if any, chips he wants to spend on punishment. Fixation screen: 2–5 s; Start screen: 2 s; Taker decision: 4 s; Response screen: 6.5 s. The 4 s window indicating the number of chips taken (Taker decision) was entered into the general linear models used for our fMRI data analysis.

decide whether to take chips from another participant playing the role of Partner. In this role as Observer, the participant also received an endowment of 200 chips at the beginning of each game. Following the decision of the Taker to either take chips or not, the Observer had the option of punishing the Taker for his action toward the Partner. Observers could use a maximum of 100 chips from their endowment for punishment purposes. Each punishment chip spent by the Observer decreased the income of the Taker by 3 chips. Observers did not know whether the Partner had also punished the Taker or not. Finally, third-party compensation games were identical to the third-party punishment games, except that in third-party compensation games the Observer had the option of compensating the Partner (as opposed to punishing the Taker). One chip spent by an Observer in third-party compensation games increased the income of the Partner by 3 chips.

The identity of Takers was never revealed to the participants, and on each trial they were told they were paired with different Takers. Thus, all games were anonymous, independent, and single-shot. Furthermore, we emphasized to participants that they should make decisions based on their own preferences, and that the punishment of the Taker would be based on the decision of one player only. Therefore, participants should punish a Taker even if they thought that another player would also punish this Taker. To control the number of trials of interest for neuroimaging analyses, Takers' choices were preprogrammed. Half of the trials were programmed to be "fair" trials, in which the Taker did not take any chips from the Partner. The other half of the trials consisted of "unfair" trials in which the Taker took 25, 50, 75, or 100 chips from the Partner. For each game type, there were 24 fair trials and 24 unfair trials, with 6 trials of each unfair trial type (i.e., 6 trials in which a Taker took 25 chips, 6 trials in which a Taker took 50 chips, etc.). To ensure participants experienced punishment and compensation to be real and consequential, six trials were randomly selected for payment, and participants were remunerated accordingly at the end of the experiment.

Participants played 48 trials of each of the three conditions of the Justice Game, resulting in 144 trials in total. Games and trial types were randomized within each experiment. After the presentation of a fixation screen (2–5 s), a start screen (2 s) appeared indicating the beginning of a trial. This screen displayed the initial endowments of both the Taker and the Partner (200 chips each), and its trial type (second- or third-party punishment or third-party compensation). We avoided directly using the terms "punishment" and "compensation" to prevent influencing individuals' actual fairness norms (Pedersen et al., 2013). Instead, games were introduced as "You-Other" games (second-party punishment games), "Other-Other, You: Take" games (third-party punishment games), and "Other-Other, You: Give" games (third-party compensation games). Next, a screen was shown that revealed the decision of the Taker, i.e., the number of chips the Taker actually took from the Partner (4 s). At the conclusion of each trial, participants indicated the amount of chips they wished to use to punish or compensate respectively on a response scale, which consisted of a row of numbers from 1 to 100, in steps of 10 (6.5 s). Responses were indicated by scrolling to the number of their choice and pressing a confirmation button. The starting position of the scrolling cursor was placed randomly on each trial. To ensure participants' attention, three self-paced breaks were included. We assessed participants' expectations about the task by asking them to report the number of chips they expected a Taker to typically take from a Partner. This measure was completed both before the administration of oxytocin/placebo, and again at the conclusion of the experiment (M = 1.5 h after treatment, SD = 6 min).

Neuroimaging data acquisition

For each participant, we acquired a T1-weighted MPRAGE high-resolution image preceded by six functional scans. During the first two functional scans, participants performed an unrelated decision-making task. During functional scans 3–6, participants played the Justice Game. Functional scans were acquired using 5-shot multiecho planar imaging GRAPPA (TR = 2390 ms; TE1 = 9.4 ms, TE2 = 21.2 ms, TE3 = 33 ms, TE4 = 45 ms, TE5 = 56 ms; matrix = 64 × 64; FOV = 224 mm; slice thickness = 3 mm; 31 axial slices; voxel size: 3.5 mm × 3.5 mm × 3.5 mm).

Neuroimaging data preprocessing

Images were preprocessed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) and FSL 6.0 (http://fsl.fmrib.ox.ac.uk). First, we combined separate echoes, using a weighted image computed from scans 5–30 acquired during the third functional run (first scan of the experiment). Subsequently, images were corrected for slice acquisition delay (ascending procedure), realigned to the mean image (6-parameter rigid body transformation), and resliced to correct for motion artifacts. The high-resolution T1 image was then coregistered to the mean functional image, and segmented into tissue types. Parameters from segmentation were then used to normalize all images to the MNI 152 brains template (12-parameter affine transformation). Functional scans were smoothed using an 8 mm Gaussian Kernel. The resulting images were corrected for temporal autocorrelation using an order 1 auto-regressive model, and periods longer than 100 s were filtered out.

General linear model specifications neuroimaging data

We estimated two separate general linear models for the analyses of the neural correlates associated with the perception and response to unfairness. The first model focused on the unfairness of the Taker's action and modeled BOLD signal from subject *i*, run *r*, and voxel *v* using the following equation (TAKER_model) :

$$y_{i,v,r} = \alpha_r + \beta_1 * \mathrm{FAIR}_{G1} + \beta_2 * \mathrm{UNFAIR}_{G1} + \beta_3 * \mathrm{UNFAIR}_{G1} * \mathrm{AT} + \beta_4 * \mathrm{SW}_{G1} + \beta_5 * \mathrm{DW}_{G1} \ldots + \mathrm{R}\gamma + \varepsilon_i. \quad (1)$$

The TAKER_model focused on the 4 s window when participants were informed how many chips had been taken by the Taker; namely 0, 25, 50, 75, or 100. We used two boxcar regressors to model this window: FAIR (trials in which the amount taken = 0) and UNFAIR (trials in which the amount taken > 0). We also used the actual amount taken per trial as a parametric modulator for the UNFAIR boxcar regressor (UNFAIR × AT). We included these three covariates for each game separately (total of 3 × 3 = 9 regressors). We also included a boxcar regressor modeling the first window of each trial (SW; duration = 2 s) regardless of game type, and a boxcar regressor covering the decision window (DW; duration = reaction time). G1 indicates Game 1 (second-party punishment game). The "…" indicates that we included the same covariates for Game 2 (third-party punishment game) and Game 3 (third-party compensation game). R is a 6-column matrix with realignment parameters and γ is the vector of coefficients that multiplies **R**. The TAKER_model was used to explore the neural correlates of fairness and unfairness.

Given that the unfairness of the Taker's action highly correlated with the amount of punishment/compensation administered, we used a second model to examine the neural correlates of decision-making separately (SPEND_model). The SPEND_model modeled BOLD signal from subject *i*, run *r*, and voxel *v* using the following equation:

$$y_{i,v,r} = \alpha_r + \beta_1 * \mathrm{NSPN}_{G1} + \beta_2 * \mathrm{SPN}_{G1} + \beta_3 * \mathrm{SPN}_{G1} * \mathrm{AI} + \beta_4 * \mathrm{FAIR}_{G1} + \beta_5 * \mathrm{SW}_{G1} + \beta_6 * \mathrm{DW}_{G1} \ldots + \mathrm{R}\gamma + \varepsilon_i. \quad (2)$$

In this equation, "NSPN" and "SPN" are dummy variables indicating whether participants punished or compensated (SPN) or did not punish or compensate (NSPN) in an unfair trial, i.e., when >0 chips were taken. "AI" is the amount of chips used for punishment and compensation and is the parametric modulator for the regressor SPN. "FAIR" concerns the 4 s window of trials in which the amount of chips taken was 0. SW and DW are the same dummy variables as in the TAKER model. Contrast images for the contrast punishment versus compensation were computed with 42 participants because coefficients for the parametric modulator could not be estimated when participants never, or only one time, spent chips on punishment or compensation per functional run or when participants consistently spent the same amount on punishment or compensation per run. Contrast images for the contrast deciding to not punish versus deciding to punish included 20 participants, as punishment amounts per run varied sufficiently in both second- and third-party punishment games for 20 participants only.

We estimated both models using restricted maximum likelihood. The SPEND_model is used to examine the neural correlates underlying the decision to not respond to the violation and to compare punishment versus compensation decisions. We analyzed the data using a two-level mixed-effect general linear model. At the first level, we analyzed data from each participant. Error due to separate functional runs was accounted for by a fixed effect model. All covariates of interest in the models were convolved with a double-gamma canonical hemodynamic response function. Additionally, we included first derivatives for each covariate of interest. We also included the six head-motion parameters as covariates of no-interest.

Computational punishment model development and estimation procedures

Our second aim was to better understand the neural correlates associated with punishment decisions in particular. To this end, we fitted a formal computational decision utility model to the behavioral punishment data. This model aimed to disentangle the processes underlying participants' willingness to punish and the severity of any subsequent punishment. Formally, modeling participants' decisions had three benefits. First, parameters corresponding to individuals' willingness to punish and the severity of their punishment were calculated separately for second- and third-party punishment games. This allowed us to assess the impact of personal involvement (i.e., being hurt or observing someone else being hurt) separately in both of the subcomponent processes. Second, this approach enabled us to disentangle whether the effect of oxytocin on punishment levels reflected a modulation of individuals' decision to punish, of their decision to punish harshly, or indeed of both. Third, having distinct individual parameters for both the decision to punish and the severity of punishment allowed us to examine the neural underpinnings of punishment decisions in a more specific manner than by use of standard brain maps alone.

Our computational utility model incorporated two parameters. Parameter θ represented the utility participants derive from punishment relative to the utility they derive from keeping chips to themselves. This value was constrained between 0 and 1, with values close to 1 reflecting a high propensity to punish, and values close to 0 denoting participants who rarely punish. Parameter α represented whether participants punished harshly in response to a violation (i.e., how eager is one to spend a large number of chips to punish?). The contribution of this parameter to the decision utility was conditional on parameter θ being >0 (i.e., if participants never punished, punishment amount is irrelevant). We formalized the model as follows, with U_i being participant *i*'s decision utility of spending chips on punishment ($s_{p,i}$):

For individual *i*:

$$U_i(s_{p,i}|\alpha_{i,k}, \theta_{i,k}, s_t) = (1 - \theta_{i,k})\pi_i + \theta_{i,k} f(s_{p,i}, s_t, \alpha_{i,k}), \quad (3)$$

with:

$$f(s_{p,i}, s_t, \alpha_{i,k}) = -\max(\alpha_{i,k}\, s_t - s_{p,i}, 0). \quad (4)$$

The term s_t represented the number of chips originally taken by the Taker, and the term π represented participant's material payoff, with π being $200 - s_t - s_p$ for second-party games and $200 - s_p$ for third-party games. The term *k* represented game type, with *k* being either second- or third-party punishment game. Equation 4 shows how participants' utility is impacted by parameter α_k. If *x* chips are taken from the participant, he will spend up to αx chips to punish the taker, as this will maximize his utility.

The decision of how much chips to spend on punishment depends on how many chips participants have available after the Taker's action. Therefore, $U_i(s_{p,i}|\alpha_{i,k}, \theta_{i,k}, s_t)$ is subject to the constraint of $s_{p,i} < 200 - s_t$ in second-party games and $s_{p,i} < 200 - s_t - s_{p,i}$ in third-party games.

The Utility model generates an optimal number of chips spent on punishment, which we label $s^*_{p,i}$:

$$s^*_{p,i} = \arg\max\_x\ U_i[x|\alpha_{i,k}, \theta_{i,k}, s_t] \text{ subject to}$$
$$x < 200 - s_t \text{ in second-party games and}$$
$$x < 200 \text{ in third-party games, with } x \in \{10, 20, \ldots\}. \quad (5)$$

In Equation 5, $s^*_{p,i}$ is the number of chips that optimizes participant *i*'s utility. Note that $s^*_{p,i}$ does not depend on the observed level of punishment, $s_{p,i}$, but rather on s_t and the participant-specific parameters α_k and θ_k. Parameters α_k and θ_k are estimated by minimizing the sum of squared errors, with *t* representing trial:

$$RSS_{k,t} = \sum_t \sum_i (s_{p,i,t} - s^*_{p,i,t})^2 \quad (6)$$

To assess whether our computational punishment model provided a compelling representation of participants' behavior, we compared it to five other models. First, we assessed the prediction accuracy based on the Nash equilibrium of both games, which is to never invest chips (Nash, 1951) and thereby reflects classical economic theory. Second, we tested performance of Inequity Aversion model (AI), a prominent model of social preferences that argues that people value equality and demonstrate a preference that everyone receives the same amount, even if this is zero (Fehr and Schmidt, 1999). Finally, we estimated three versions of our model, one that sets $\theta_{3pp} = \theta_{2pp}$ and $\alpha_{3pp} = \alpha_{2pp}$; one that allows different θ values across games but same α values; and the model as presented in the paper in which both parameters can differ between game types (Model 4).

It is important to note that our goal here is not to test whether our model is necessarily better or worse than the Nash equilibrium or the AI model, but rather to assess the extent to which our model fits our data relative to these alternative theories. Thus, we present these comparisons to provide a reference point to evaluate model fit. Another advantage of the development of the computational model is that this allows us to test whether the willingness to punish and the decision of how severely to punish are separable processes. In the experiment itself, participants make only one decision (that is, a punishment choice of 0–100 chips). However, the output of the model will demonstrate whether these processes are in fact separable, even though this process distinction is not directly observable in the behavior itself. Furthermore, we believe that, in principle, it is useful to define a model when trying to understand complex behavior, such as responses to unfairness, as this model can potentially be applied to other paradigms. Modeling behavior can therefore be a principled way forward to better understand and represent responses to unfairness in a broader context.

For all games, we computed parameters to minimize the residual sum of squares (See Eq. 6). Parameters were estimated independently for each

Table 1. Punishment model summary and fit

Model	No. of parameters	RSS	LL	AIC	R^2
Nash	0	92,511	1431	2862	NAN
AI	3	43,704	505	1017	0.303
Model 2	2	23,015	243	489	0.526
Model 3	3	19,170	198	403	0.616
Model 4	4	15,687	162	333	0.678

The column labeled RSS reports the average of the residual sum of squares across participants. LL indicates the average of the log-likelihood across participant. AIC was computed per participant using Equation 7, and the average is reported here. R^2 is the squared correlation coefficient between true and predicted response. NAN = not a number.

participant. We started by constructing a coarse grid over the parameter space to select the best initial estimates. Then, we used a constraint optimization procedure implemented in MATLAB to obtain the final estimates. The Akaike Information Criterion (AIC) for each model was computed as follows:

$$\mathrm{AIC}^M = 2\mathrm{k} - 2LL + 2\mathrm{k}(\mathrm{k}+1)/(\mathrm{N}-\mathrm{k}-1), \qquad (7)$$

where LL is the log-likelihood for model M, k the number of parameters, and N the number of observations. We computed LL using the estimated utilities per each strategy and a softmax function. Although there are no dramatic differences in the fit of the three versions of the model, Model 4 is the best performing ($LL = 162$, AIC = 333, $R^2 = 0.68$, accuracy = 46%; Table 1). Therefore, we used this parameterization to investigate the motivations underlying punishment behavior.

Region-of-interest analyses for model parameters

With regard to the neural correlates associated with the model parameters and the effect of oxytocin on parameter estimates, we were primarily interested in brain areas that had been previously shown to play a key role in processes associated with punishment decision-making. These prespecified brain regions included the amygdala, the anterior insula (AIns) and the dorsolateral prefrontal cortex (DLPFC). Previous research showed that the amygdala and the AIns were associated with the signaling of fairness violations (Corradi-Dell'Acqua et al., 2013; Zhong et al., 2016), and the encoding of harm respectively (Buckholtz et al., 2008; Gospic et al., 2011; Treadway et al., 2014), whereas the final determination of appropriate punishment actions has been suggested to critically depend on the DLPFC (Knoch et al., 2006; Buckholtz et al., 2008, 2015; Treadway et al., 2014; Zhong et al., 2016). We conducted our analyses only on these three brain regions, as we had strong a priori hypothesis about these areas in particular. Regions-of-interest (ROIs) in AINs and DLPFC were defined using the MarsBaR toolbox for Statistical Parametric Mapping (SPM). AIns and DLPFC masks were created by centering 10 mm radius spheres at the peak of activation clusters reported previously in the neuroimaging literature on social norm violations (Table 2) (Sanfey et al., 2003). The amygdala mask was based on the Automated Anatomical Labeling (AAL) atlas.

Experimental design and statistical analysis

Our experimental design was a randomized, double-blind, placebo-controlled between-subject design.

Analysis behavioral data. To assess whether punishment and compensation behavior differed across game types and whether punishment and/or compensation levels were impacted by oxytocin administration, we conducted a 2 (treatment: placebo/oxytocin, between-subjects factor) × 3 (game type: second-party punishment game/third-party punishment game/third-party compensation game, within-subjects factor) repeated-measures multivariate ANOVA with participants' mean levels of punishment and compensation behavior as dependent variables. To assess to what degree the magnitude of the norm violation impacted the frequency and severity of punishment per game type, we conducted a 2 (treatment: placebo/oxytocin, between-subjects factor) × 4 (number of chips taken by Taker: 25/50/75/100 chips, within-subjects factor) × 10 (number of chips spent per punishment or compensation: 10/20/30/40/50/60/70/80/90/100, within-subjects factor) repeated-measures multivariate ANOVA for each game type with participants' summed choices as dependent variables. For all ANOVAs, Greenhouse–Geisser corrections were used to correct for violations of sphericity if required, and partial η^2 values are reported as a measure of effect size. To assess the effect of game type and treatment on parameters of the computational model, two 2 (treatment: placebo/oxytocin, between-subjects factor) × 2 (game type: second-party punishment game/third-party punishment game, within-subjects factor) repeated-measures multivariate ANOVAs were conducted with participants' parameter estimates (θ or α) as the dependent variable. Results of the self-report measures were analyzed using a χ^2 test. Behavioral data were analyzed with IBM SPSS statistics software.

Table 2. MNI coordinates used to create 10 mm sphere masks for ROI analyses

Brain region	Hemisphere	x	y	z	
AIns	R	39	18	−4	Sanfey et al., 2003
AIns	L	−35	16	7	Sanfey et al., 2003
DLPFC	R	39	37	22	Sanfey et al., 2003
DLPFC	L	−34	45	16	Sanfey et al., 2003

Analysis neuroimaging data. All neuroimaging data were analyzed with SPM software. As explained above, we estimated two separate general linear models for the analyses of the neural correlates associated with the perception and response to unfairness. Neuroimaging results are reported here using a common cluster-corrected threshold at $p < 0.05$ familywise error (FWE) on the basis of an initial whole-brain voxelwise threshold of $p < 0.001$. Exceptions are when specific clusters of brain activity are better visualized by using a more stringent whole-brain FWE correction for multiple comparisons (for Neural correlates of fairness of unfairness). Neural correlates of our model parameters were analyzed using a priori defined ROIs. To explore the role of the amygdala, AIns, and DLPFC in the decision to punish and in punishment severity, we performed multiple regression analyses for second- and third-party games separately. ROI activations were used as dependent variables in SPM's multiple-regression analysis, whereas model parameters, oxytocin treatment, and the interaction between treatment and parameters, served as predictors. Contrast values are reported to indicate effect size and directionality of the effect. To directly compare correlations between parameter estimates and ROI activity in the different game types, Pearson r values were transformed into Z-scores, and these Z-scores were compared using one-tailed tests for nonoverlapping correlations based on dependent groups (Silver et al., 2004). Correlations were compared using R's package cocor (Diedenhofen and Musch, 2015).

Results

Results for decisions to punish violators and compensate victims

First, we examined participants' decision-making in the task, exploring differences between punishment and compensation games, as well as the effect of oxytocin on behavior.

Punishment and compensation. Repeated-measures multivariate ANOVAs with participants' mean levels of punishment and compensation behavior as dependent variables, confirmed that participants were willing to incur costs to both punish and compensate in response to unfairness, and also that behavior differed across game types ($F_{(2,51)} = 24.05$, $p < 0.001$, partial $\eta^2 = 0.49$). Specifically, participants were more likely to punish than to compensate ($t_{(53)} = 6.06$, $p < 0.001$) and were willing to spend more chips in the second-party punishment condition (when they themselves were victims) than in the third-party punishment condition (when someone else was hurt; $t_{(53)} = 5.99$, $p < 0.001$; $\mathrm{M}_{\text{second punish}} = 38.07$, $\mathrm{SD}_{\text{second punish}} = 23.55$; $\mathrm{M}_{\text{third punish}} = 23.52$, $\mathrm{SD}_{\text{third punish}} = 18.47$; $\mathrm{M}_{\text{comp}} = 13.93$, $\mathrm{SD}_{\text{comp}} = 11.63$; Fig. 2). Oxytocin administration did not affect mean levels of punishment or compensation ($F_{(1,52)} = 0.12$, $p = 0.742$, partial $\eta^2 = 0.002$).

Frequency and severity of punishment. To assess to what degree the magnitude of the norm violation impacted the frequency and severity of punishment, repeated-measures multivariate ANOVAs that assessed the frequency of the various levels of punishment possible showed that, across game types, the more that was taken from the victim, the more chips were spent to either punish or compensate (second-party punishment

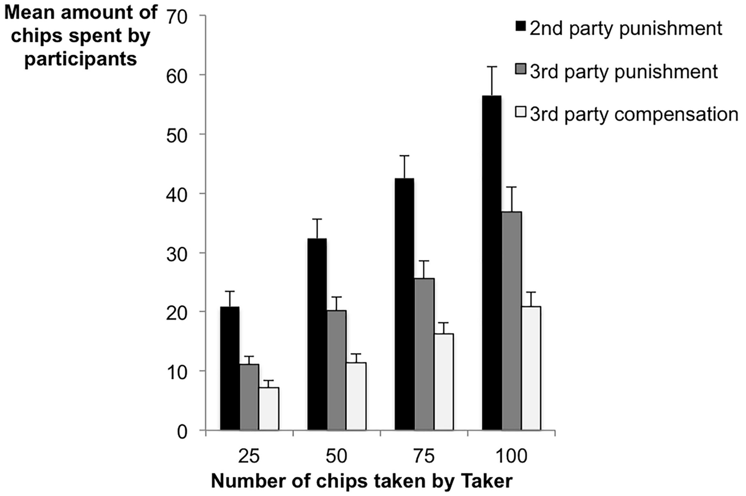

Figure 2. Mean amount of chips spent as a function of game type and number of chips taken by the Taker. Error bars are SEM. $n = 54$.

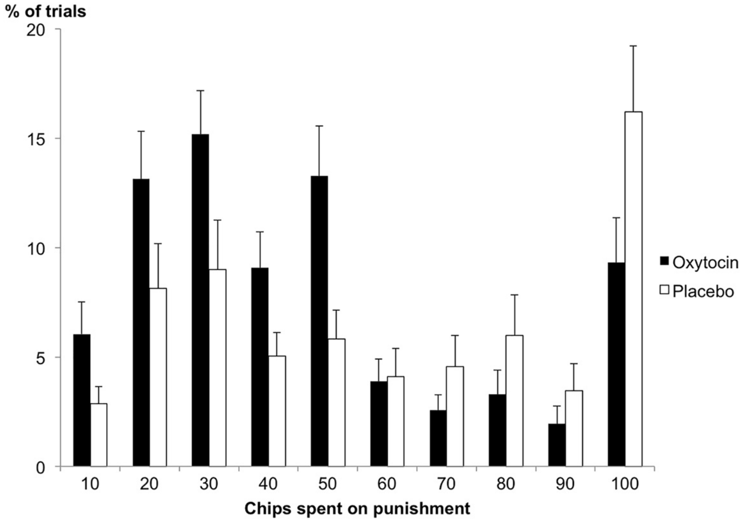

Figure 3. Frequency of punishment in the second-party punishment games. Participants in the oxytocin group administered smaller punishments (10–50 chips) more often than participants in the placebo group. Error bars are SEM. $n = 54$.

game: $F_{(27,26)} = 3.45$, $p = 0.001$, partial $\eta^2 = 0.78$; third-party punishment game: $F_{(27,26)} = 24.05$, $p = 0.016$, partial $\eta^2 = 0.71$; compensation game: $F_{(27,26)} = 2.41$, $p = 0.013$, partial $\eta^2 = 0.69$).

Effect of oxytocin. Interestingly, in the second-party punishment game, participants in the oxytocin group administered low punishments more often than participants in the placebo group (Punishment Amount × Treatment: $F_{(9,44)} = 2.10$, $p = 0.050$, partial $\eta^2 = 0.30$; Fig. 3). In the third-party punishment game, this effect also emerged, but was further qualified by the Taker's behavior (Punishment Amount × Treatment × Amount of chips taken: $F_{(27,26)} = 2.34$, $p = 0.017$, partial $\eta^2 = 0.71$). Thus, when participants observed someone else being hurt, the effect of oxytocin on punishment was stronger as the norm violation increased (when more was taken; Fig. 4). Self-report measures that were administered both before and after the experiment and assessed participants' beliefs about what behavior they expected from the other players, showed that participants who received oxytocin expected fair treatment more often than did participants in the placebo group. Specifically, 12 of 27 participants in the oxytocin group changed their beliefs after treatment, now expecting other players to take no chips at all, whereas only 3 of 27 participants in the placebo group changed their expectations ($\chi^2_{(1)} = 7.45$, $p = 0.006$).

Neuroimaging results

First, we examined the neural correlates of the perception of both fairness and unfairness in the Justice Game to compare our findings with previous reports in the literature on social norm violations. For these analyses, we collapsed across punishment and compensation conditions. Next, we explored the neural processes underlying responses to social norm violations as per our first aim, and report the neural activity underlying punishment versus compensation (findings for the opposite contrast were absent), as well as the neural correlates associated with the decision to refrain from either punishment or compensation in response to a violation. For these analyses, we collapsed across second- and third-party punishment conditions, and because oxytocin treatment had no effect on mean punishment and compensation levels, results here are additionally collapsed over treatment. Neural correlates of second- and third-party punishment were examined by exploring correlations between the computational model parameter estimates and the brain activity in the AINs, amygdala, and DLPFC.

Neural correlates of fairness and unfairness. Brain regions that showed greater activity for fairness versus unfairness were the ventral medial prefrontal cortex (VMPFC), posterior insula, posterior cingulate, and superior temporal sulcus (Table 3; Fig. 5; $p < 0.05$, FWE whole-brain corrected; contrasting trials in which the Taker took no chips vs trials in which chips were taken). Brain regions that correlated with the degree of unfairness were the AIns, anterior cingulate cortex (ACC), DLPFC, precuneus and ventrolateral prefrontal cortex (VLPFC; Table 3; Fig. 6; $p < 0.05$, FWE whole-brain corrected; contrasting trials in which the Taker took chips vs trials in which no chips were taken, with activity correlating with the number of chips taken). When examining brain activity correlating with the amount of punishment or compensation administered, we found that activation in these areas also correlated positively with decision-making. That is, the higher the activation in this network, the more punishment and compensation was subsequently administered ($p < 0.05$, FWE whole-brain corrected; Table 4; contrasting trials in which participants punished/compensated vs trials in which participants did not punish/compensate, with activity correlating with the amount of punishment/compensation administered). These findings align well with previous neuroimaging work on fairness and unfairness, providing support for the validity of our novel task (Sanfey et al., 2003; Tabibnia et al., 2008; Harlé et al., 2012; Feng et al., 2015).

Neural correlates of punishment versus compensation. Comparing brain activity for punishment versus compensation showed higher activity in ventral striatum for punishment decisions (initial threshold set at $p = 0.001$ uncorrected, cluster corrected at $p < 0.05$ FWE, cluster of 57 voxels; Fig. 7*A*; Table 5; contrasting trials in which punishment is administered in second- and third-party punishment games vs trials in which compensation is administered).

Neural correlates of deciding not to punish. The decision to refrain from punishing in response to unfairness was associated with enhanced activity in the temporal parietal junction (TPJ; initial threshold set at $p = 0.001$ uncorrected, cluster corrected at $p < 0.05$ FWE, cluster of 60 vox-

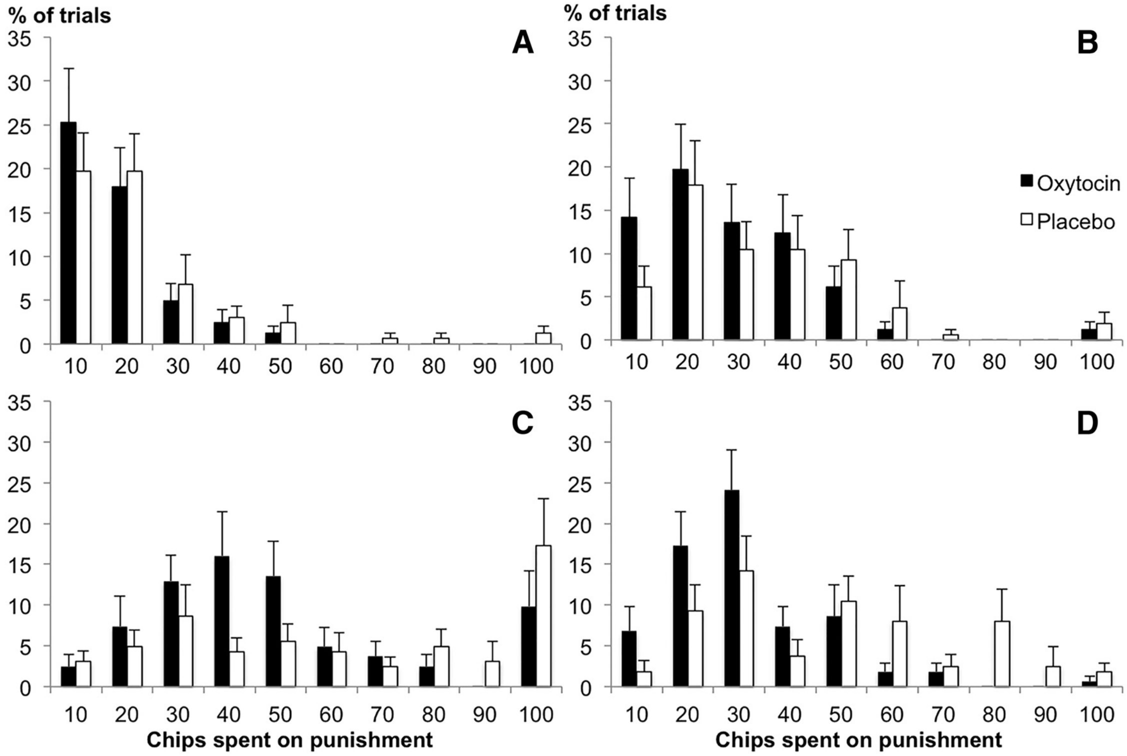

Figure 4. Frequency of punishment in third-party punishment game. *y*-axis, Percentage of trials; *x*-axis, number of chips spent per punishment. ***A***, Taker takes 25 chips; (***B***) Taker takes 50 chips; (***C***) Taker takes 75 chips; and (***D***) Taker takes 100 chips. Error bars indicate SEM. The more chips the Taker takes, the stronger the effect of oxytocin on the administration of smaller punishments. $n = 54$.

Table 3. Significant activation clusters associated with the perception of fairness and unfairness

Brain region	Hemisphere	*x*	*y*	*z*	No. of voxels	*Z*
Clusters correlating with fairness (no chips taken) vs unfairness (amount of chips taken >0)						
—VMPFC	R	10	53	−10	355	6.04
—Posterior insula	L	−53	−24	22	468	6.65
—Posterior insula	R	52	4	4	96	6.34
—Posterior cingulate	L	−11	−24	46	145	6.65
—Superior temporal sulcus	L	−60	−20	−10	96	5.99
Clusters correlating with unfairness (amount of chips taken on unfair trials) vs fair trials						
—ACC	R	3	39	25	576	7.77
—VLPFC	L	−46	42	8	249	7.8
—DLPFC	L	−36	25	42	(same cluster as L VLPFC)	5.94
—VLPFC	R	45	46	8	411	6.93
—AIns	R	48	22	−6	(same cluster as R VLPFC)	6.6
—Precuneus	R	6	−76	42	553	6.87
—Cerebellum	L	−39	−62	−34	2055	7.62
—Occipital cortex	L	−53	−66	−6	(same cluster as cerebellum)	7.52

$p < 0.05$, FWE whole-brain corrected.

els; Fig. 7*B*; Table 5; contrasting decisions to not punish vs decisions to punish in second- and third-party punishment games).

Results for computational modeling of second- and third-party punishment

Parameter θ: willingness to punish. Repeated-measures ANOVA with game type as a within-subject factor and treatment as a between-subject factor, demonstrated that estimates of θ did not differ between second- and third-party punishment games ($F_{(1,51)} = 0.91$, $p = 0.345$, partial $\eta^2 = 0.02$; second-party punishment: range = 0.45–0.9, mean = 0.65, SD = 0.13; third-party punishment: range = 0.45–0.9, mean = 0.64, SD = 0.14). However, there was an interaction between game type and treatment, with estimates of θ significantly higher in second-party games for participants receiving oxytocin compared with placebo ($F_{(1,51)} = 4.5$, $p = 0.034$, partial $\eta^2 = 0.09$; Fig. 8*A*). Oxytocin administration did not affect estimates of θ in third-party games, suggesting that oxytocin might impact the decision to punish only when one is hurt directly. In third-party game types the correlation between α and θ was −0.63 ($p < 0.001$). As expected, this correlation is negative: when people punished frequently they were less likely to punish severely.

Parameter α: severity of punishment. Repeated-measures ANOVA for α yielded a main effect of game type, indicating that participants valued harsher punishments more in second-party games ($F_{(1,51)} = 6.4$, $p = 0.015$, partial $\eta^2 = 0.11$; second-party punishment: range = 0.10–1.9, mean = 0.75, SD = 0.41; third-party punishment: range = 0.10–1.1, mean = 0.60, SD = 0.31). Additionally, there was a main effect of treatment on α, with estimates for both second- and third-party games significantly lower for participants in the oxytocin versus placebo group ($F_{(1,51)} = 8.4$, $p = 0.006$, partial $\eta^2 = 0.14$; Fig. 8*B*). This main effect echoes the earlier reported behavioral finding showing that oxytocin administration impacted the administration for low punishments in particular, an effect that was present for both game types. In second-party game types the correlation between α and θ was −0.27 ($p = 0.049$).

Neuroimaging results

Our second aim was to gain greater understanding of the subcomponent processes underlying punishment in second- and third-party contexts. To this end, we examined whether brain activity in the AIns, DLPFC, and amygdala was associated with individual parameter estimates obtained from our computational punishment model.

Second-party punishment. For second-party punishment games, brain maps used contrasted unfairness (amount of chips taken) versus fairness in second-party punishment games. Results showed that individual estimates of parameter θ correlated with a significant increase in activity in bilateral AIns (right AIns: $t = 3.75$, $p < 0.001$, contrast value = 5.59; left AIns: $t = 1.87$, $p = 0.034$, contrast value = 2.94), but not with activity in

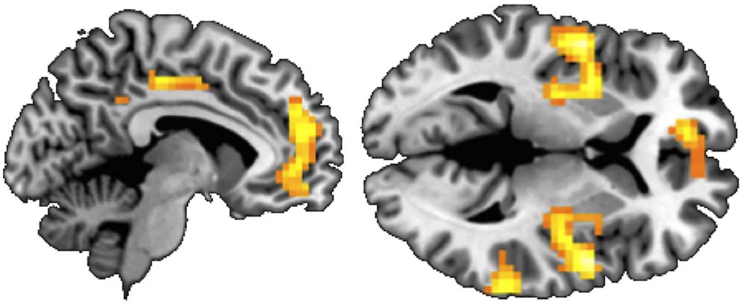

Figure 5. Neural correlates of fairness: trials in which Taker took no chips versus trials in which chips were taken. MNI slice: left, $x = -8$; right, $z = 4$. Displayed at $p < 0.05$ FWE whole-brain corrected. $n = 53$.

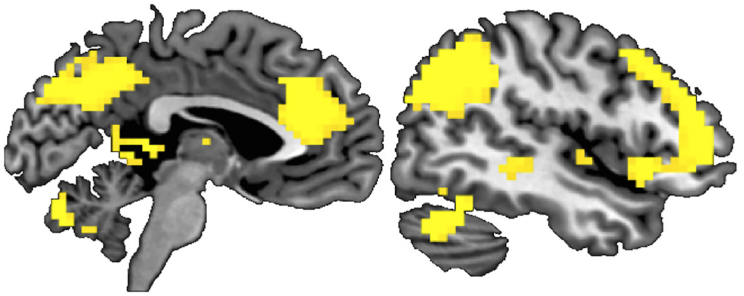

Figure 6. Neural correlates of unfairness: trials in which Taker took chips versus trials in which no chips were taken, with activity correlating with the number of chips taken. MNI slice: left, $x = 46$; right, $x = -4$. Displayed at $p < 0.001$ corrected. $n = 53$.

Table 4. Significant activation clusters correlating with amount punished/compensated

Brain region	Hemisphere	x	y	z	No. of voxels	Z
Clusters correlating with amount of punishment/compensation						
ACC	R	3	42	25	93	5.61
VLPFC	L	−46	42	11	46	5.83
VLPFC	R	45	46	11	43	5.76
Precuneus	R	6	−70	50	56	5.42
Superior parietal cortex	R	52	−42	50	43	6.43
Superior parietal cortex	L	−53	−42	46	205	6.04
Parieto-occipital sulcus	R	52	−66	22	58	5.24
Cerebellum	L	−36	−70	−31	122	5.95
Cerebellum	R	34	−70	−34	122	5.41

$p < 0.05$, FWE whole-brain corrected.

the amygdala or DLPFC. Oxytocin administration increased activity in the right AIns ($t = 1.69$, $p = 0.049$, contrast value $= 2.69$). No neural activity correlated with individual estimates of parameter α, nor with interactions between treatment and estimates of θ or α.

Third-party punishment. For third-party punishment games, brain maps used contrasted unfairness (amount of chips taken) versus fairness in third-party punishment games. Results showed that individual estimates of parameter θ activity correlated with an increase in activity in the right DLPFC ($t = 2.12$, $p = 0.019$, contrast value $= 3.61$) and right AIns (right: $t = 1.65$, $p = 0.05$, contrast value $= 4.18$). Individual estimates of parameter α correlated with an increase of activity in the left amygdala ($t = 1.80$, $p = 0.039$, contrast value $= 2.19$). Oxytocin administration did not significantly increase activity in the bilateral AIns in third-party punishment games (left: $t = 1.60$, $p = 0.058$, contrast value $= 4.69$; right: $t = 1.55$, $p = 0.063$, contrast value $= 5.55$), and there no neural activity correlated with interactions between treatment and estimates of θ or α.

Correlation comparisons showed that the correlation between estimates of θ (willingness to punish) and right AIns, but not left AIns, was significantly stronger in second-party than third-party game types (right AIns: $Z = 1.91$, $p = 0.028$; $r_{\text{right AIns-}\theta\text{ second-party}} = 0.465$, $r_{\text{right AIns-}\theta\text{ third-party}} = 0.171$; left AIns: $Z = 0.842$, $p = 0.199$, $r_{\text{left AIns second-party}} = 0.259$, $r_{\text{left AIns-}\theta\text{ third-party}} = 0.121$). Correlations between estimates of θ and left DLPFC were not significantly stronger in third-party compared with second-party game types ($Z = 0.024$, $p = 0.510$, $r_{\text{left DLPFC-}\theta\text{ third-party}} = 0.152$, $r_{\text{left DLPFC-}\theta\text{ second-party}} = 0.156$). However, the correlation between the amygdala and estimates of α (severity of punishment) were significantly greater for third-party compared with second-party punishment games ($Z = -1.83$, $p = 0.033$, $r_{\text{left amygdala-}\alpha\text{ third-party}} = 0.184$, $r_{\text{left amygdala-}\alpha\text{ second-party}} = -0.134$).

Discussion

Our findings provide new insights into the neural and psychological processes of punishment and compensation in response to social norm violations. We developed a novel experimental decision-making task that elicited clear behavioral responses to violations of fairness norms, and that additionally allowed for an investigation of the neural processes underlying decisions about punishing and compensating. Our findings align clearly with previous reports on the neural correlates of perceived fairness and unfairness (Sanfey et al., 2003; Tabibnia et al., 2008), and add several new contributions to the burgeoning neuroimaging literature on social norm violations.

Punishment preferred to compensation

Our first aim was to understand better the neural mechanisms underlying decisions to punish transgressors versus decisions to compensate victims of a norm violation. We found that enhanced ventral striatal activity was associated more strongly with deciding to punish a player who behaved unfairly than to compensate a player who had been disadvantaged. Although inferring psychological processes from brain activity requires caution (Poldrack, 2011), the link between the ventral striatum and reward processing has been well established in a recent meta-analysis of >200 neuroimaging studies (Bartra et al., 2013). Therefore, this finding suggests that, in the present paradigm, punishment may have been preferred over compensation, because punishment was experienced as more rewarding. Indeed, the behavioral data showed that participants preferred punishment above compensation.

Previous work on respective preference for compensation and punishment has yielded mixed results, with participants sometimes preferring punishment over compensation, whereas other studies have demonstrated the reverse (Leliveld et al., 2012; Chavez and Bicchieri, 2013; Hu et al., 2016). One potential explanation of why participants largely preferred punishment here is that these other studies have typically used variants of the Dictator or Ultimatum Game in which the fairness norm is violated by sharing resources unequally. However, in the Justice Game the fairness norm is violated by someone directly taking resources from you or someone else. This may be experienced as a stronger norm violation than unequal sharing (Bosman and Van Winden, 2002; Ben-Shakhar et al., 2007; Korenok et al., 2014), resulting in a stronger preference for punishment over compensation, as observed here.

In addition to an increase in ventral striatal activity for punishment versus compensation decisions, we found that the decision to refrain from punishment was associated with increased activity in the TPJ, an area associated with perspective-taking

(Carter and Huettel, 2013). Interestingly, whereas in our study this TPJ activity was associated with decisions to refrain from punishment, other studies have found the TPJ to be involved in a mentalization process supporting punishment decisions (Buckholtz et al., 2008; Krueger et al., 2014; Treadway et al., 2014; Ginther et al., 2016; Hu et al., 2016). These contradictory findings are likely due to differences in experimental design, with earlier work focusing on hypothetical criminal scenarios in which the perceived intent of the wrongdoer was manipulated.

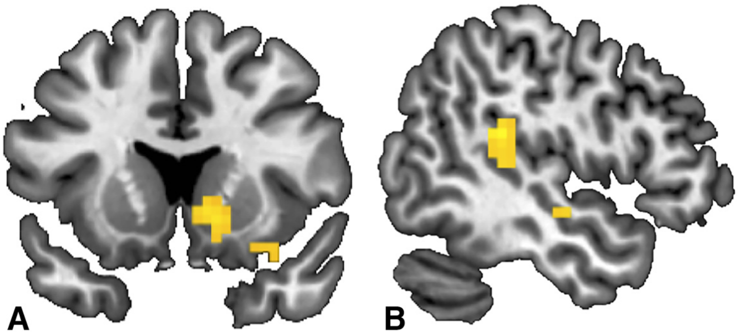

Figure 7. Brain contrast maps displayed at $p < 0.001$ uncorrected. ***A***, Punishment versus compensation: trials in which participants invested in second- and third-party punishment contrasted with trials in which participants invested in compensation; $y = 14$ (MNI). $n = 42$. ***B***, Not punishing in response to unfair treatment: trials in which the Taker took chips and participants chose to not punish versus trials in which the Taker took chips and participants punished; $x = -49$ (MNI). $n = 20$.

Dissociable neural correlates for punishment decision and punishment severity

Our second aim was to compare the psychological motives and associated brain activity underlying the decision of whether to punish, and then of how severely to punish, across both second- and third-party punishment contexts. First, and importantly, results from our computational model support the notion of dissociable processes involved in punishment, comprising an initial decision to punish, followed by a subsequent choice of how severely to punish. Specifically, examining the neural correlates of the model-derived parameters demonstrates that the AIns was related to the decision to punish, with this correlation stronger in second-party compared with third-party punishment scenarios. Given the role of this brain area in the signaling of social norm violations (Güroğlu et al., 2010; Corradi-Dell'Acqua et al., 2013; Zhong et al., 2016), a reasonable interpretation is that individuals' willingness to punish depends on their fairness considerations, with individuals' fairness preferences being stronger when one was directly involved than when another person was the victim. In third-party punishment cases, when participants were not themselves direct victims of injustice, results showed that the decision to punish was also associated with enhanced activity in the DLPFC. However, this correlation was only present in the regression analysis when the correlation between estimates of θ and DLPFC activity were controlled for the effects of other variables, including the effect of treatment and estimates of parameter α. A comparison of correlations across games between θ and DLPFC activity directly showed that this relationship was not significantly stronger in third-party as opposed to second-party punishment games. Future research using larger sample sizes could productively investigate the involvement of the DLPFC in third-party contexts in more detail. The DLPFC is known for its role in integrating information and response selection (Ridderinkhof et al., 2004), and previous work on third-party punishment in hypothetical criminal contexts has found this area involved in integrating information about the responsibility of the wrongdoer and the amount of harm done (Buckholtz et al., 2015). Possibly therefore, pure fairness considerations (as represented here by model-derived AIns activation) play a decisive role in second-party punishment situations, whereas additional contextual information is integrated in punishment decisions in third-party contexts, when others are harmed (possibly involving the DLPFC). In third-party scenarios, we additionally found the amygdala associated with the willingness to punish severely, a finding that resonates with previous work showing that amygdala activity correlated with punishment severity in hypothetical crime scenarios, supporting the hypothesis that the amygdala encodes affective arousal associated with harm done to someone else (Buckholtz and Marois, 2012; Krueger and Hoffman, 2016).

Table 5. Significant activation clusters associated with punishment versus compensation decisions and the decision to not punish

Brain region	Hemisphere	x	y	z	No. of voxels	Z
Clusters correlating with the decision to punish (second- and third-party punishment trials) versus the decision to compensate						
—Ventral striatum	R	14	18	−6	57	3.94
—Occipital cortex	R	17	−70	−3	125	4.32
Clusters correlating with the decision to not punish (second- and third-party punishment trials) versus the decision to punish						
—Temporal Parietal Junction	L	−42	−42	18	60	3.86

Initial whole-brain threshold at $p < 0.001$ uncorrected and cluster corrected at $p < 0.05$ FWE.

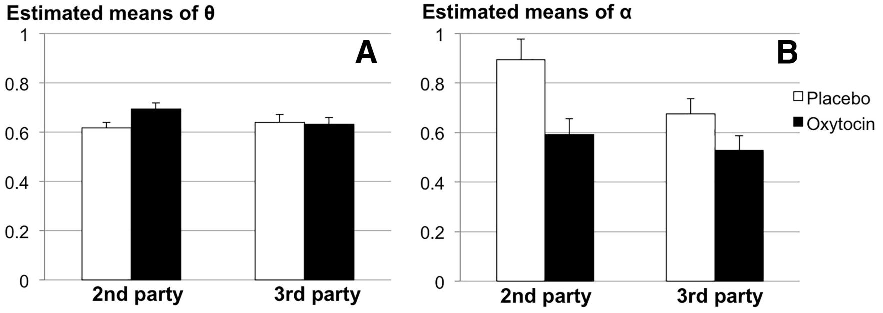

Figure 8. Mean parameter estimates per game type. ***A***, Means of parameter θ reflecting the decision to punish. ***B***, Means of parameter α reflecting the decision to punish hard. Error bars indicate SEM. $n = 53$.

Oxytocin enhances a "corrective" punishment

The third aim was to explore the role of the neuropeptide oxytocin in compensation and punishment decisions. Decisions to compensate victims, along with their neural correlates, were unaffected by oxytocin. These results therefore do not support the notion of oxytocin as a general enhancer of empathy and prosocial, altruistic, decisions (Bartz et al., 2011). As we had a relatively little amount of compensation behavior, lack of oxytocin effects on compensation may be due to restriction of range. However, with regard to punishment decisions, oxytocin had a systematic effect at both the behavioral and neural levels.

Behaviorally, oxytocin enhanced the frequency of small punishments by lowering individuals' willingness to punish harshly, suggesting that oxytocin may underlie a corrective response to norm violations; that is, enhance the decision to give a "slap on the wrist" to an offender when they behave (moderately) badly. We speculate that oxytocin could therefore impact more "cognitively-based" vigilant punishment behavior, as opposed to operating via an affective response to a transgression, for example one based on anger or fear. Results from the computational model revealed that oxytocin administration decreased participants' willingness to punish harshly in both second- and third-party punishment games. In second-party punishment games, oxytocin also impacted participants' decision to punish, suggesting there may not be a one-to-one mapping between parameter θ and sensitivity to unfairness, but that the final decision to punish may also depend on other factors, such as whether one is hurt directly. Neurally, oxytocin enhanced activity in the AIns in second- and third-party punishment games, though only significantly in the former. Given the role of the AIns in signaling fairness violations (Güroğlu et al., 2010; Corradi-Dell'Acqua et al., 2013; Zhong et al., 2016), these results suggest that one potential mechanism for oxytocin to impact punishment is that via subtly altering fairness preferences. Indeed, results of self-report measures suggest that oxytocin influenced participants' fairness expectations in consistent ways. However, an interaction between θ and game type, as was observed at the behavioral level, was not present in the AIns. To find significant interactions between treatment and parameter correlates at the neural level larger sample sizes may be required.

The finding that oxytocin impacted punishment in both second- and third-party games is difficult to reconcile with the idea of oxytocin motivating direct retaliation, defined as preferentially punishing when you yourself are harmed. Rather, our findings better fit the hypothesis that oxytocin upregulates a motivation to create and maintain fairness vis-à-vis direct interaction partners (Daughters et al., 2017). Seen as such, our results provide evidence for the hypothesis that oxytocin shifts the focus from self-interest to overarching, normative, group interests (De Dreu and Kret, 2016).

In sum, we developed a novel behavioral task and a new computational model of punishment that allowed us to study both the psychological and neural mechanisms underlying punishment and compensation in response to norm violations. The present findings not only provide more insight into the fundamental brain mechanisms underlying punishment and compensation, but also demonstrate how combining a computational approach with neuroimaging methods can help disentangling the psychological motives underlying complex decision-making processes.

References

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412–427. CrossRef Medline

Bartz JA, Zaki J, Bolger N, Ochsner KN (2011) Social effects of oxytocin in humans: context and person matter. Trends Cogn Sci 15:301–309. CrossRef Medline

Baumgartner T, Heinrichs M, Vonlanthen A, Fischbacher U, Fehr E (2008) Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron 58:639–650. CrossRef Medline

Ben-Shakhar G, Bornstein G, Hopfensitz A, van Winden F (2007) Reciprocity and emotions in bargaining using physiological and self-report measures. J Econ Psychol 28:314–323. CrossRef

Bosman R, Van Winden F (2002) Emotional hazard in a power-to-take experiment. Econ J 112:147–169. CrossRef

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

Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, Marois R (2008) The neural correlates of third-party punishment. Neuron 60: 930–940. CrossRef Medline

Buckholtz JW, Martin JW, Treadway MT, Jan K, Zald DH, Jones O, Marois R (2015) From blame to punishment: disrupting prefrontal article from blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron 87:1369–1380. CrossRef Medline

Carter RM, Huettel SA (2013) A nexus model of the temporal-parietal junction. Trends Cogn Sci 17:328–336. CrossRef Medline

Chavez AK, Bicchieri C (2013) Third-party sanctioning and compensation behavior: findings from the ultimatum game. J Econ Psychol 39:268–277. CrossRef

Corradi-Dell'Acqua C, Civai C, Rumiati RI, Fink GR (2013) Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc Cogn Affect Neurosci 8:424–431. CrossRef Medline

Daughters K, Manstead ASR, Ten Velden FS, Dreu CKW De (2017) Oxytocin modulates third-party sanctioning of selfish and generous behavior within and between groups. Psychoneuroendocrinology 77:18–24. CrossRef Medline

De Dreu CKW, Greer LL, Handgraaf MJJ, Shalvi S, Van Kleef GA, Baas M, Ten Velden FS, Van Dijk E, Feith SWW (2010) The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328:1408–1411. CrossRef Medline

De Dreu CKW, Greer LL, Van Kleef GA, Shalvi S, Handgraaf MJJ (2011) Oxytocin promotes human ethnocentrism. Proc Natl Acad Sci U S A 108:1262–1266. CrossRef Medline

De Dreu CK, Kret ME (2016) Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense. Biol Psychiatry 79:165–173. CrossRef Medline

Diedenhofen B, Musch J (2015) cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10:e0121945. CrossRef Medline

Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. Games Econ Behav 47:268–298. CrossRef

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

Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114:817–868. CrossRef

Feng C, Luo YJ, Krueger F (2015) Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. Hum Brain Mapp 36:591–602. CrossRef Medline

Ginther MR, Bonnie RJ, Hoffman MB, Shen FX, Simons KW, Jones OD, Marois R (2016) Parsing the behavioral and brain mechanisms of third-party punishment. J Neurosci 36:9420–9434. CrossRef Medline

Gospic K, Mohlin E, Fransson P, Petrovic P, Johannesson M, Ingvar M (2011) Limbic justice-amygdala involvement in immediate rejection in the ultimatum game. PLoS Biol 9:e1001054. CrossRef Medline

Güroğlu B, van den Bos W, Rombouts SA, Crone EA (2010) Unfair? It depends: neural correlates of fairness in social context. Soc Cogn Affect Neurosci 5:414–423. CrossRef Medline

Harlé KM, Chang LJ, van 't Wout M, Sanfey AG (2012) The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. Neuroimage 61:32–40. CrossRef Medline

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

Hu Y, Scheele D, Becker B, Voos G, David B, Hurlemann R, Weber B (2016) The effect of oxytocin on third-party altruistic decisions in unfair situations: an fMRI study. Sci Rep 6:20236. CrossRef Medline

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. CrossRef Medline

Korenok O, Millner EL, Razzolini L (2014) Taking, giving, and impure altruism in dictator games. Exp Econ 17:488–500. CrossRef

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

Krueger F, Parasuraman R, Moody L, Twieg P, de Visser E, McCabe K, O'Hara M, Lee MR (2013) Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. Soc Cogn Affect Neurosci 8:494–498. CrossRef Medline

Krueger F, Hoffman M, Walter H, Grafman J (2014) An fMRI investigation of the effects of belief in free will on third-party punishment. Soc Cogn Affect Neurosci 9:1143–1149. CrossRef Medline

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

Ma Y, Shamay-Tsoory S, Han S, Zink CF (2016) Oxytocin and social adaptation: insights from neuroimaging studies of healthy and clinical populations. Trends Cogn Sci 20:133–145. CrossRef Medline

Nash J (1951) Non-cooperative games. Ann Math 54:286–295. CrossRef

Pedersen EJ, Kurzban R, McCullough ME, Mccullough ME (2013) Do humans really punish altruistically? A closer look. Proc Biol Sci 280:20122723. CrossRef Medline

Poldrack RA (2011) Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72:692–697. CrossRef Medline

Rabin M (1993) Incorporating fairness into game theory and economics. Am Econ Rev 83:1281–1302. CrossRef

Ridderinkhof KR, van Den Wildenberg WPM, Segalowitz SJ, Carter CS (2004) Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. Brain Cogn 56:129–140. CrossRef Medline

Rilling JK, DeMarco AC, Hackett PD, Thompson R, Ditzen B, Patel R, Pagnoni G (2012) Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. Psychoneuroendocrinology 37:447–461. CrossRef Medline

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. Science 300:1755–1758. CrossRef Medline

Silver NC, Hittner JB, May K (2004) Testing dependent correlations with nonoverlapping variables: a Monte Carlo simulation. J Exp Educ 73:53–69. CrossRef

Stallen M, De Dreu CKW, Shalvi S, Smidts A, Sanfey AG (2012) The Herding Hormone: Oxytocin Stimulates In-Group Conformity. Psychol Sci 23: 1288–1292. CrossRef Medline

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. CrossRef Medline

Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol Sci 19:339–347. CrossRef Medline

Treadway MT, Buckholtz JW, Martin JW, Jan K, Asplund CL, Ginther MR, Jones OD, Marois R (2014) Corticolimbic gating of emotion-driven punishment. Nat Neurosci 17:1270–1275. CrossRef Medline

Zhong S, Chark R, Hsu M, Chew SH (2016) Computational substrates of social norm enforcement by unaffected third parties. Neuroimage 129: 95–104. CrossRef Medline