eNeuro

Research Article: New Research | Cognition and Behavior

Functional connectivity basis and underlying cognitive mechanisms for gender differences in guilt aversion

https://doi.org/10.1523/ENEURO.0226-21.2021

Cite as: eNeuro 2021; 10.1523/ENEURO.0226-21.2021

Received: 19 May 2021
Revised: 5 November 2021
Accepted: 18 November 2021

Manuscript Title: Functional connectivity basis and underlying cognitive mechanisms for gender differences in guilt aversion

Abbreviated title: Gender differences in guilt-based prosociality

List all Author: Tsuyoshi Nihonsugi[1,2], Shotaro Numano[2,3] & Masahiko Haruno[2,3]

[1]Faculty of Economics, Osaka University of Economics, 2-2-8, Osumi, Higashiyodogawa-ku, Osaka, 533-8533, Japan. Email: t.nihonsugi@gmail.com.

[2]Center for Information and Neural Networks, NICT, 1-4 Yamadaoka, Suita, Osaka, 565-0871, Japan. Email: shotaro.numano@gmail.com.

[3]Graduate School of Frontier Biosciences, Osaka University 1-3 Yamadaoka, Suita, Osaka 565-0871 Japan. Email: mharuno@nict.go.jp.

Author Contributions. T.N and M.H designed the study. T.N and M.H conducted the experiments. T.N., S. N. and M.H analyzed the data and wrote the paper.

Corresponding authors: Correspondence should be addressed to Masahiko Haruno (mharuno@nict.go.jp) or Tsuyoshi Nihonsugi (t.nihonsugi@gmail.com).

Acknowledgements. We are grateful to Satoshi Tada and Tomoki Haji for technical assistance, and Peter Karagiannis for editing an early version of the manuscript.

Conflict of interest disclosure. The authors declare no competing financial interests.

Funding sources. This work was supported by CREST, and COI to Osaka University, both by JST and KAKENNHI (17H06314 and 26242087).

Abstract

Prosocial behavior is pivotal to our society. Guilt aversion, which describes the tendency to reduce the discrepancy between a partner's expectation and his/her actual outcome, drives human prosocial behavior as does well-known inequity aversion. Although women are known to be more inequity averse than men, gender differences in guilt aversion remain unexplored. Here we conducted a functional magnetic resonance imaging (fMRI) study (n = 52) and a large-scale online behavioral study (n = 4723) of a trust game designed to investigate guilt and inequity aversions. The fMRI study demonstrated that men exhibited stronger guilt aversion and recruited right DLPFC-VMPFC connectivity more for guilt aversion than women, while VMPFC-DMPFC connectivity was commonly used in both genders. Furthermore, our regression analysis of the online behavioral data collected with Big Five and demographic factors replicated the gender differences and revealed that Big Five Conscientiousness (rule-based decision) correlated with guilt aversion only in men, but Agreeableness (empathetic consideration) correlated with guilt aversion in both genders. Thus, this study suggests that gender differences in prosocial behavior are heterogeneous depending on underlying motives in the brain and that the consideration of social norms plays a key role in the stronger guilt aversion in men.

Significance Statement

Although women are established to be more prosocial than men in terms of inequity aversion, gender differences in prosocial behavior based on motives prominent in guilt aversion are far less explored. Here we conducted a fMRI study and a large-scale online behavioral study to address gender differences in guilt aversion. We demonstrate that men are more sensitive to guilt aversion than women, and a prefrontal social-norm network is key to men's predominance in guilt-based prosocial behavior. These findings

revealed the heterogeneity of gender differences in prosocial behavior depending on underlying motives and underlying neural mechanisms.

Introduction

Prosocial behaviors are fundamental to human society. The most perceived motivation behind prosocial behaviors is inequity aversion (Fehr and Schmidt, 1999), which is defined as the propensity to avoid an imbalance between outcomes for the self and the other person. A great deal of behavioral research (Andreoni and Vesterlund, 2001; Bolton and Katok, 1995; Croson and Gneezy, 2009; Dickinson and Tiefenthaler, 2002; Eckel and Grossman, 1998; Grosch and Rau, 2017; Kamas and Preston, 2015) has accumulated evidence that women are more prosocial than men, since women are more inequity-averse.

However, economic research has shown that human prosocial behavior depends on not only preferred behavioral outcomes (e.g., fairness), but also on the belief of others (for a review, see Fehr and Schmidt, 2006). People tend to live up to the expectations of others, since they suffer from guilt if they disappoint others (Baumeister et al., 1994). In behavioral game theory, this psychological process is named “guilt aversion” (Battigalli and Dufwenberg, 2007, 2009; Charness and Dufwenberg, 2006), in which an individual dislikes disappointing another person relative to what the other person believes they should receive (see Materials and Methods for a more detailed definition). However, gender differences in guilt aversion have been far less explored.

Previous functional magnetic resonance imaging (fMRI) studies of guilt aversion have revealed involvement of the dorsolateral prefrontal cortex (DLPFC), dorsal medial prefrontal cortex (DMPFC), ventromedial prefrontal cortex (VMPFC), insula, supplementary motor area, and temporal parietal junction (Chang et al., 2011; Nihonsugi et al., 2015; van Baar et al., 2019). For instance, it was demonstrated that the

DLPFC is causally involved in the implementation of guilt aversion by integrating fMRI and transcranial direct current stimulation (tDCS) (Nihonsugi et al., 2015). Considering these contributions of prefrontal cortices in guilt aversion, we assumed that prefrontal network interactions among the DLPFC, DMPFC and VMPFC play a key role in producing the gender difference in guilt aversion, if any. In particular, the VMPFC may well be central to the gender difference in guilt aversion because several lesion studies (Sutterer et al., 2015; Tranel et al., 2005) demonstrated that the VMPFC is involved in the gender differences in social cognition.

Additionally, it is also possible that the gender difference in guilt aversion may reflect different cognitive strategies used by men and women. Guilt aversion requires the ability to assess another individual's expectations and directly relates to his or her disappointment (i.e., empathy or theory of mind; Hoffman, 1982). At the same time, guilt aversion is a normative behavior elicited by experience (i.e., rule-based decisions; Haidt, 2003). Therefore, we also hypothesized that if there is a gender difference in guilt aversion, these two potential cognitive strategies: empathetic consideration and rule-based decision-making may contribute to the difference.

Regarding inequity aversion, previous fMRI studies (Crockett et al., 2013; Gospic et al., 2011; Haruno and Frith, 2010; Haruno et al., 2014; Tanaka et al. 2017; Tricomi et al., 2010) revealed involvement of the ventral striatum and amygdala. An integration of pharmacological intervention and fMRI also demonstrated that activity in the ventral striatum is critical for gender differences in this aversion (Soutschek et al., 2017). Therefore, we hypothesized that women show stronger inequity aversion than men, with the striatum and amygdala playing a critical role.

To test these hypotheses from a neuro-cognitive point of view, we conducted a model-based fMRI study and a large-scale online behavioral study of the trust game task, which was designed to measure guilt aversion and inequity aversion. The fMRI study

investigated the neural and network mechanisms for the guilt and inequity aversions, with particular focus on gender differences. For the online behavioral data, a regression analysis of guilt aversion was conducted based on Big Five and social factors, such as age and socioeconomic status, by which we expected cognitive and societal aspects of guilt aversion would be revealed.

Materials and Methods

Intersection of fMRI and online studies

Trust game*.** Participants performed a trust game adapted from the task originally used by Charness and Dufwenberg (2006). In this task, two subjects are paired as players A and B (see Fig. 1A***). First, player A must choose between *In* and *Out* options and simultaneously reveal their belief about τ_A (from 0% to 100%), the probability that player B will choose *Cooperate*. In other words, τ_A is player A's level of trust in player B. If player A chooses *Out*, players A and B receive payments z_A and z_B, respectively. If player A chooses *In*, then knowing player A's belief probability, player B must choose *Cooperate* or *Defect*. If player B chooses *Defect*, player A receives y_A and player B receives y_B; if player B chooses *Cooperate*, then the two players receive x_A and x_B, respectively. In the example shown in Fig. 1***B***, the belief probability of player A was 80%. If player B defected, player A and player B would receive 220 yen and 910 yen, respectively; if player B cooperated, they would receive 780 yen and 650 yen, respectively.

There are two important conditions regarding the payments in Fig. 1***A*** (see also the definitions of guilt and inequity aversion below): if (1) $y_A < z_A < x_A$, then player A signals trust (cooperation) to player B when player A chooses *In*; if (2) $z_B < x_B < y_B$, then player B feels guilt upon disappointing player A relative to player A's belief in

what player A will receive. This trust game was originally designed and used in Nihonsugi et al. (2015).

Guilt aversion and inequity aversion. Guilt aversion (Battigalli and Dufwenberg, 2007, 2009; Charness and Dufwenberg, 2006) assumes that an individual dislikes not meeting another's belief. Note that guilt sensitivity elicited in the trust game by guilt aversion theory is fundamentally related to the Test of Self-Conscious Affect-3 (TOSCA-3) and the Guilt and Shame Proneness Scale (GASP), which is a common measure of guilt sensitivity in psychology, but is unrelated to shame (Bellemare et al., 2019; Bracht and Regner, 2013).

This model includes social pressure on player B if the profile (*In*, *Defect*) is played (see Fig. 1***A***). Player B is assumed to believe that if player A chooses *In*, then player A believes that he will get a return of $\tau_A \cdot x_A + (1-\tau_A) \cdot y_A$, because the setting of player A's payoff is $y_A < z_A < x_A$. The difference, $\{\tau_A \cdot x_A + (1-\tau_A) \cdot y_A\} - y_A = \tau_A(x_A - y_A)$, which is non-negative in our settings, can measure how much player B believes that he/she has disappointed player A relative to player A's belief had player B chosen *Defect*. In other words, the difference $\tau_A(x_A - y_A)$ is the amount of guilt that player B experiences. Let us assume that γ_B is the parameter that measures player B's sensitivity to guilt. A player is guilt-averse and will *Cooperate* if $y_B - \gamma_B \cdot \tau_A(x_A - y_A) < x_B$. In the example trial in Fig. 1***B***, if $910 - \gamma_B \cdot 0.8 \cdot (780 - 220) < 650$, player B will choose *Cooperate*. Since γ_B does not directly measure guilt experiences or emotional traits, we can only infer that "γ_B expresses sensitivity of guilt". As mentioned in the Results section, however, our interpretation that γ_B expresses a guilty experience is consistent with the results of the post-experiment questionnaire.

By contrast, inequity aversion assumes a social preference for equitable payoffs (Fehr and Schmidt, 1999). An individual is inequity-averse if, in addition to their monetary self-interest, their utility decreases when the allocation of monetary payoffs is

different. If an inequity-averse player suffers from inequity, they will choose an option that results in a smaller difference between their own and the other's monetary payoffs. Notably, the advantageous-inequity (receiving a larger reward than others) in Fehr and Schmidt's inequity-aversion model is also referred to as "guilt". However, it is important to note that this outcome-based "guilt" and the intension-based "guilt" we treat in guilt-aversion are completely different.

As mentioned below, based on the results of the model selection using both the cross-validation analysis (predictive likelihood) and the Bayesian information criterion (BIC) (Fig. 2***B*** and 2***C***; see also Model validation and comparison in Materials and Methods), the absolute difference for inequity was found superior than the standard inequity aversion model, which splits the inequity into positive and negative terms, in the present study.

We integrated guilt aversion and inequity aversion into a utility function (u_B) for player B as follows:

$$u_B = \begin{cases} x_B - \alpha_B |x_A - x_B| \; if \; the \; profile \; (In, Cooperate) \\ y_B - \gamma_B \cdot \tau_A \cdot (x_A - y_A) - \alpha_B |y_A - y_B| \; if \; the \; profile \; (In, Defect), \end{cases}$$

where α_B is a constant that measures player B's sensitivity to inequity. A narrowly self-interested agent is given the special case $\gamma_B = \alpha_B = 0$. In our game, players choose between binary actions that yield two different monetary payoff allocations, $X = (x_A,\ x_B)$ and $Y = (y_A,\ y_B)$. The utilities of these allocations are given by the formula above, yielding $u_B(X)$ and $u_B(Y)$.

Statistical analysis of behavioral data. We estimated three separate components—monetary self-interest, guilt, and inequity—for each participant based on the logistic model of stochastic choice. The probability that player B chooses *Cooperate* can be expressed as $P_{B,Cooperate} = 1/1 + e^{-\{u_B(X) - u_B(Y)\}}$. Although our model does not include an inverse temperature parameter explicitly, this does not imply the model

does not consider decision noise. In fact, our model implicitly assumed the inverse temperature parameter to be 1. Such an implementation of the softmax function with the inverse temperature parameter = 1 is often seen in the behavioral analysis of the economic decision-making (e.g., Boorman et al., 2009; Cai, & Padoa-Schioppa, 2014; Suzuki et al., 2015) because the inverse temperature is relatively difficult to estimate. Based on this logistic model, we used a logistic regression as follows:

$$logit(P_{B,Cooperate}) = \beta_0 + \beta_1 Reward_t + \beta_2 Guilt_t + \beta_3 Inequity_t,$$

where $Reward_t$ is the size of the reward and calculated as $x_B - y_B$ at time *t*, $Guilt_t$ is the size of guilt and calculated as $-\{0 - \tau_A \cdot (x_A - y_A)\}$, and $Inequity_t$ is the size of inequity and calculated as $-(|x_A - x_B| - |y_A - y_B|)$. For convenience, β_1, β_2, and β_3 are denoted as $\beta(Reward)$, $\beta(Guilt)$, and $\beta(Inequity)$, respectively. In order to orthogonalize the three explanatory variables, the actual τ_A used in the experiments was also set by the experimenter. Player B was asked to make decisions assuming that player A chose the *In* option. We therefore set τ_A to 60% or higher (player A is expected to choose the *Out* option when τ is small). More specifically, τ_A was 60% 7 times, 70% 5 times, 80% 13 times, 90% 11 times, and 100% 9 times. We display the actual values of x, y, z, and τ_A in Extended Data Figure 1-1. The correlation coefficients among the three explanatory variables were less than 0.30 and insignificant ($P > 0.05$); the values of guilt and inequity were designed to be orthogonal (the correlation coefficient of these two variables was -0.138 and nonsignificant ($P = 0.367$)) to dissociate the computational processes for guilt aversion and inequity aversion.

This logistic regression was computed using the R statistical package (R Development Core Team, 2008). We used the brglm package to conduct our maximum likelihood estimation with the bias-reduction method (Kosmidis, 2019).

Model validation and comparison. Our utility model comprises three separate

components: Reward, Guilt, and Inequity, as defined above. With regard to Inequity, we adopted the absolute difference for Inequity. However, participants may alternatively use Fehr and Schmidt's model, which splits the inequity into positive (called Inequity-positive hereafter) and negative (called Inequity-negative hereafter) terms. Therefore, we need to verify which model (component) better explains the data for the current experiments.

To address this issue, we first compared 10 possible models (for details of the 10 models, see Fig. 2***B***) based on the predictive negative log likelihoods using a cross-validation. This cross-validation approach for value-based decision-making allows us to avoid overfitting the data and to compare models with different numbers of parameters robustly. It has also been adopted in many recent studies (Daw, 2011; Linderman and Gershman, 2017; Park, et al., 2019; Smith, et al., 2014; Fig. 2***B***). We also compared more familiar BIC values for the models and exemplified the ones with the first and second minimum BIC to confirm the results (Fig. 2***C***).

More specifically, to compute the minimum predictive negative log-likelihood, we repeated bootstrap (500 iterations) three-fold cross-validations for the model validation and comparison. For each model, we randomly divided 45 trials for each participant into three groups of equal size (i.e., 15). We fitted the model to 30 trials and predicted the behavior in the held-out 15 trials and repeated this process three times. We repeated this three-fold cross-validation procedure 500 times and selected the model with the minimum predictive negative log-likelihood for held-out trials.

fMRI study

Participants. 52 participants (mean age 21.2 years; SD = 1.4 years; 26 females) participated in the fMRI experiments. They were scanned on a Siemens 3T Trio scanner at the Center for Information and Neural Networks (CiNet) of the National Institute of

Information and Communications Technology (NICT). The ethical committees of the NICT approved this study, and all participants gave informed consent. Participants received money proportional to the number of payoffs earned during the experiment (equivalent to 45 to 60 US dollars). Although our task was the same as the one in Nihonsugi et al. (2015), we collected completely different participants in this study for two main reasons. First, we had access to a 64-channel MRI coil to analyze the DMPFC and VMPFC. The 64-channel brain coil provides a 1.3-fold higher signal-to-noise ratio in the brain cortex than the 32-channel array (Keil et al., 2013). Second, the number of participants (n = 42) in Nihonsugi et al. (2015) was not enough for re-analysis; Yarkoni (2009) suggested that a sample size of more than 50 is necessary for identifying a moderate correlation at relatively conservative thresholds. Additionally, we also wished to test whether we could replicate our previous results.

Experimental design and procedure. We conducted two experiments in which participants played a trust game in different roles (see Fig. 1***C***; see also the instructions in Extended Data Figure 1-2). In the first (behavioral) experiment, more than ten participants per experiment were invited into a room and read instructions of the rules and procedure of the trust game. Every participant played the trust game as player A (i.e., choose *In* or *Out* and reveal belief probability τ_A) and experienced one trial. The participants were informed that these choices would be used when player B made their choice in the second (fMRI) experiment. However, player A was not informed of player B's identity. Participants were told that earnings for player A will be determined according to the actual outcome made by both players' choices if A's choice is used in the second experiment.

The second experiment was conducted on average six days (range = 1–10 days) after the first experiment. All participants played the game as player B (i.e., choose *Cooperate* or *Defect* with knowledge of player A's belief probability) for 45 trials.

Participants were instructed to assume that player A chose *In* in this experiment (the *Out* option is illustrated as a dashed line in Fig. 1***B***). The sequence of the trials was randomized across subjects. Participants were told that the other participant (player A) differed for each trial and that the pairings were anonymous. We did not provide any feedback to the participants during the experiment. Participants were also informed that earnings for player B will be the sum of the show-up fee and the actual outcome obtained from both players' choices in the 45 trials.

Because there was the risk that Player B felt that the other player was hypothetical rather than real, we invited more than 10 participants at a time into a room in the first experiment to make them realize the other's presence and impress on them that they would have a real partner in the second experiment. In addition, when giving instructions for the second experiment, we repeatedly explained that we had conducted similar first experiments many times and that there were many player As and the partner in the second experiment was one of them. In other words, on the day of the second experiment, the participants were likely to think about other participants in the first experiment. Thus, although the experiment was hypothetical, we assume that the participants were engaged in the tasks as if they were in a real interaction. Indeed, no participant reported or even referred to the absence of their partner in a post-experiment interview.

After reading the instructions for the task and procedure, the participants were briefed about the rules of the game by the experimenter and tested to confirm that they understood the rules. They were then individually invited into the scanning room and practiced the game using the response buttons in the scanner.

Functional images were acquired as participants played the game. The timeline of a trial is shown in Fig. 1***B***. Each trial began with a 2-5 s preparation interval during which time a green fixation cross was presented for the first 1 s and then a yellow

fixation cross (cue phase) was presented for the remainder of the time. The participants were then presented with the trust game, including the allocation of monetary payoffs for each choice and player A's belief, and selected *Cooperate* or *Defect* by pressing the corresponding button within 5 s (choice phase). In each trial, participants made their choice on the assumption that player A chose *In*. This was followed by the presentation of a fixation cross for a variable time period of 6-15 s (rest phase).

After scanning, all participants answered the questionnaire. For guilt aversion behavior, participants were asked to answer the following three questions on a 5-point scale (1: I don't think so,…, 5: I think so).

a. Did you think that the reason why Player A chose *In* was because they expected (and aimed) to gain x_A yen (i.e., the result of Player B choosing *Cooperate*)?

b. Did you think that choosing *Defect* would reduce the payoff (x_A yen) expected by Player A?

c. Did you feel guilt that your choice of *Defect* would reduce the payoff (x_A yen) expected by Player A?

Question *a* examined whether the respondent understood the partner's intention of choosing *In* (the meaning behind the expectation); Question *b* examined whether the respondent was aware that their choice of *Defect* reduces their partner's expected payoff; and Question *c* asked whether the respondent felt guilty when he/she reduced their partner's expected payoff.

***fMRI image acquisition*.** Scanning was performed on a Siemens 3T Trio scanner with a 64-channel coil at CiNet using an echo planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 3000 ms, echo time (TE) = 25 ms, flip angle = 90°, matrix = 64 × 64, field of view (FOV) = 192 mm, slice thickness = 3 mm,

gap = 0 mm, and ascending interleaved slice acquisition of 51 axial slices. High-resolution T1-weighted anatomical scans were acquired using an MPRAGE pulse sequence (TR = 2000 ms, TE = 1.98 ms, FOV = 256 mm, image matrix 256 × 256, slice thickness = 1 mm). We discarded the first two EPI images before data processing to compensate for T1 saturation effects.

fMRI data preprocessing. SPM12 (http://www.fil.ion.ucl.ac.uk/spm) was used for the MRI data preprocessing and analysis. Preprocessing included motion correction, co-registration to the participant's anatomical image, and spatial normalization to the standard Montreal Neurological Institute (MNI) T2 template with a resampled voxel size of 2 mm. Co-registered EPI data were normalized using an anatomical normalization parameter. Spatial smoothing was performed using an 8 mm Gaussian kernel.

General analysis methods. To explore the neural basis of guilt, inequity and value difference, we performed a general linear model (GLM) analysis of the functional data. We constructed two GLM models.

GLM 1. To model the blood-oxygen-level dependent (BOLD) signal driven by Guilt and Inequity, the two variables were convolved with a hemodynamic response function (HRF) (spm_hrf function with TR equal to 3.0 s). For first level GLM analysis, the onset and duration were the onset timing of "Choice phase" and 0 s, respectively. In addition to a response-period constant regressor, we introduced (1) an HRF for Guilt and (2) an HRF for Inequity. Additional regressors modeling head motion, as derived from the realignment procedure, were included in the model. Serial autocorrelation was modeled as a first-order regressor, and data were high-pass filtered at a cutoff of 128 s.

We calculated second-level group contrasts using one-sample *t*-tests to reveal the main effect of each parametric regressor within participants using the individual

contrast images. To correct for multiple comparisons, we used for Guilt contrast the familywise error (FWE) correction across the whole brain at $P < 0.05$ based on Gaussian random field theory as implemented in SPM12 (minimum cluster extent (k) > 20 voxels, see also Extended Data Figure 3-1 for the actual cluster size). Since the analysis of Inequity targets small regions, such as the striatum and amygdala, we set the minimum cluster extent to 20 voxels in order to keep the extent size the same throughout the analysis of Guilt and Inequity. When analyzing Inequity, we used for the whole brain analysis a threshold of $P < 0.001$ uncorrected.

GLM 1.1. After calculating GLM1, a two-sample *t*-test was used to compare Guilt contrast between men and women. For the whole-brain analysis, a threshold of $P < 0.001$ uncorrected at peak voxel level with an extent threshold of $k = 20$ was adopted.

GLM 1.2. After calculating GLM1, a two-sample *t*-test was used to compare Inequity contrast between men and women. For the whole-brain analysis, a threshold of $P < 0.001$ uncorrected at peak voxel level with an extent threshold of $k = 20$ was adopted.

GLM 2. We modeled brain activity related to utility. For the first-level analysis, we entered the value difference between choice options (larger utility-smaller utility) as a parametric modulator of a regressor. The onset and duration were the onset timing of the "Choice phase" and 0 s, respectively. Additional regressors modeling head motion, as derived from the realignment procedure, were included in the model. Serial autocorrelation was modeled as a first-order regressor, and data were high-pass filtered at a cutoff of 128 s.

We calculated second-level group contrasts using one-sample *t*-tests to reveal the main effect of each parametric regressor within participants using the individual contrast images. Additional regressor modeling of a gender-indicating variable was

included in the model. We used for the whole brain analysis a threshold of $P < 0.001$ uncorrected.

Region of interest analysis. For the Guilt contrast in GLM1, due to the lack of adequate previous neuroimaging studies and consistent imaging results for guilt aversion, we had no specific priori hypothesis and performed no region of interest (ROI) analysis. However, for GLM1.1 (gender difference in guilt), we did have a priori hypothesis from previous lesion studies that showed the VMPFC is involved in gender differences in social cognition (Sutterer et al., 2015; Tranel et al., 2005). Therefore, we performed a ROI analysis on whether this region survived a small volume correction at $P < 0.05$ with an FWE correction. For Inequity contrast in GLM1 and GLM1.2, because we had a priori hypothesis from previous research that found the amygdala and striatum are involved in inequity (Crockett et al., 2013; Gospic et al., 2011; Haruno and Frith, 2010; Haruno et al., 2014; Tanaka et al. 2017; Tricomi et al., 2010) and there exists a gender difference in inequity (Soutschek et al., 2017), we employed a ROI analysis with a small volume correction ($P < 0.05$; small volume FWE corrected). With regard to value difference in GLM2, we again had a priori hypothesis from previous research that found the VMPFC is involved in value difference (e.g., Hunt et al., 2012; Nicolle et al., 2012). Therefore, we employed a ROI analysis with a small volume correction ($P < 0.05$; small volume FWE corrected).

The small volume of the VMPFC and DMPFC was based on a 15-mm sphere around the coordinates ($x = 2$, $y = 41$, $z = -6$) and ($x = -3$, $y = 48$, $z = 30$), because these coordinates were used in a neuroimaging study (Baumgartner et al., 2011) of social preferences similar to ours. In that study, the VMPFC coordinates were determined by averaging the peak coordinates across five neuroimaging studies (value and economic decision-making), and the DMPFC coordinates were based on a meta-analysis study on social cognition (van Overwalle, 2009). Furthermore, the VMPFC coordinates

(subjective value: $x = 2$, $y = 46$, $z = -8$; decision stage: $x = 2$, $y = 40$, $z = -8$) in a previous meta-analysis (Bartra et al., 2013) are very close to the coordinates we used. The small volumes for the amygdala and striatum were defined using the WFU PickAtlas toolbox (Maldjian et al, 2003).

Psycho-physiological interaction (PPI) analysis. We performed two PPI analyses using the function of SPM12.

PPI1. Having confirmed that the VMPFC was involved in value difference by the GLM2 analysis, we next conducted a hypothesis-based PPI analysis to examine whether this VMPFC activity truly integrates the value components of Guilt and Inequity. More specifically, we used VMPFC (shown in Fig. 3***C***) as a seed region and examined whether brain areas associated with VMPFC × Guilt overlapped with the Guilt-correlated areas (i.e., DLPFC and DMPFC in Fig. 3***A***) and whether brain areas associated with VMPFC × Inequity overlapped with the Inequity-correlated area (i.e., striatum in Fig. 3***B***). For each subject, we extracted the time course of activity from a 5 mm-radius volume of interest (VOI) around the peak voxel in the VMPFC (shown in Fig. 3***C***). Based on the procedure by Gitelman et al. (2003), the time series of the VOI was extracted and then deconvolved, multiplied with the psychological variable (size of Guilt or Inequity), and reconvolved with the HRF set up as the PPI regressor. The three regressors (i.e., PPI regressor, VOI time series, and psychological variable) were then convolved with the canonical HRF and entered into the regression model along with six head motion parameters. The individual parameter estimate image for the PPI regressor was subsequently subjected to a one-sample *t*-test. Finally, we also included a gender-indicating variable and performed a group analysis to identify brain regions showing increased functional connectivity with the seed VOI during the Choice phase. For the whole-brain analysis, we used a threshold of $P < 0.001$ uncorrected.

PPI2. The goal of this analysis was to examine whether different brain networks are involved in the computation of guilt and inequity between men and women. More specifically, this analysis aimed to find differences between men and women in brain regions that correlate more strongly with VMPFC or striatum activity as guilt or inequity increases. For each subject, we extracted the time course of activity from VOIs with a 5 mm radius around the peak voxel in the VMPFC, as shown in Fig. 4***A***, and the ventral striatum, as shown in Fig. 5***A***. For this analysis, the PPI terms were defined as VMPFC × guilt and ventral striatum × inequity. We entered six variables (i.e. PPI regressor, VOI time series and psychological variable for guilt and inequity, respectively) and movement regressors into a GLM. The individual parameter estimate image for the PPI regressor was subsequently subjected to a one-sample *t*-test. Finally, group analysis was performed to identify brain regions showing increased functional connectivity with the seed VOIs. A two-sample *t*-test was performed to further assess different connectivity patterns between men and women. For the whole-brain analysis, a threshold of $P < 0.001$ uncorrected at the peak voxel level with an extent threshold of $k = 20$ was adopted.

Mediation analysis*.** We performed a mediation analysis to test whether the interaction between gender and guilt-based prosocial behavior was mediated by a brain function using a mediation toolbox (https://github.com/canlab/MediationToolbox) (Wager et al., 2008). Briefly, this analysis was based on a standard three-variable path model, as shown in Fig. 4D***. This analysis quantifies the degree to which a relationship between two variables, X and Y, can be explained by another variable, M.

For the guilt-aversion behavioral analysis, we defined X as the gender-indicating variable (1 = men), Y as the behavioral variable, $\beta(Guilt)$, and M as the brain variable functional connectivity between the right DLPFC and VMPFC (Fig. 4***A***). Following convention, we required that three tests reach statistical significance in the mediation

analysis. First, path a measured the association between the gender-indicating variable and the functional connectivity. Second, path b measured the association between the functional connectivity and $\beta(Guilt)$ after controlling for the gender-indicating variable. Third, the mediation effect, defined as the product of the indirect paths ($a \times b$), must be significant. We refer to the overall predictor-outcome relationship as effect c and the direct effect controlling for the mediator c'. Thus, the $a \times b$ effect tests the significance of $c - c'$. We conducted bootstrap tests (10,000 iterations) for statistical significance of the mediators.

For inequity-aversion behavioral analysis, we defined X as the gender-indicating variable (1 = women), Y as the behavioral variable $\beta(Inequity)$, and M as the brain variables (striatum shown in Fig. 5***A***).

Online study

Participants. We analyzed data from 4723 participants (mean age 37.9 years, SD = 15.4 years, 2737 females; for more detailed descriptive statistics, see Table 1) who followed the task instructions correctly and spent longer than 1 hour to complete 7 different personality trait tests such as Big Five Inventory, anxiety (STAI) and depression (BDI) and the trust game task. These data were collected using our in-house online experiment system. The study protocol was approved by the ethical committees of the NICT, and all participants gave informed consent. For their participation, participants were paid in cashable points proportional to the number of payoffs earned during the experiment (equivalent to 3 to 5 US dollars).

Experimental design and procedure. Participants performed a trust game on our in-house online experiment system in a similar way to the fMRI study (see Fig. 1***C***). We conducted two consecutive experiments in which participants played a trust game in a different role. Before the first experiment, online participants read the rules of the trust

448 game and the procedure. In the first experiment, every participant played the trust game
449 as player A (i.e., choose *In* or *Out* and reveal belief probability τ_A) and experienced one
450 trial. Participants knew these choices would be used, and the pairings were anonymous
451 when player B made their choice in the second behavioral experiment.

452 In the second experiment, all participants played the game as player B (i.e.,
453 choose *Cooperate* or *Defect* with knowledge of player A's belief probability).
454 Participants (player B) were instructed to assume that player A chose *In*. Every
455 participant experienced 45 trials. The sequence of the trials was randomized across
456 subjects. Participants were told that the other participant (player A) differed for each trial
457 and that the pairings were anonymous. We did not provide any feedback to the
458 participants during the experiment. All participants answered seven different personality
459 trait tests including the Big Five Inventory. The final earnings were calculated following
460 the same pattern as the fMRI study.

461 ***Evaluation of cognitive mechanisms using Big Five Inventory.*** We first examined the
462 relationship between guilt-aversion ($\beta(Guilt)$) and gender. Specifically, we estimated
463 $\beta(Guilt)$ for participants by the same logistic regression as the fMRI study and
464 compared $\beta(Guilt)s$ between men and women. To investigate two different cognitive
465 processes (i.e. agreeableness and conscientiousness) potentially underlying gender
466 difference in guilt aversion and to control for the confounding effects of the participant's
467 socioeconomic status, we conducted a multiple linear regression analysis based on the
468 following equation:

$$\beta(Guilt)_i = \beta_1 Neuroticism_i + \beta_2 Extraversion_i + \beta_3 Openness_i + \beta_4 Agreeableness_i + \beta_5 Conscientiousness_i + \beta_6 Age_i + \beta_7 SelfEduHistory_i + \beta_8 ParentsEduHistory_i + \beta_9 Income + \beta_{10} Occupation + \beta_{11} SubjectiveSES_i + \beta_{12} Sex_i \times Neuroticism_i + \beta_{13} Sex_i \times Extraversion_i + \beta_{14} Sex_i \times Openness_i + \beta_{15} Sex_i \times Agreeableness_i + \beta_{16} Sex_i \times Conscientiousness_i + \beta_{17} Sex_i \times Age_i + \beta_{18} Sex_i \times SelfEduHistory_i + \beta_{19} Sex_i \times ParentsEduHistory_i + \beta_{20} Sex_i \times Income + \beta_{21} Sex_i \times Occupation + \beta_{22} Sex_i \times SubjectiveSES_i + \varepsilon_i,$$

where $Neuroticism_i$, $Extraversion_i$, $Openness_i$, $Agreeableness_i$ and $Conscientiousness_i$ are the individual's Big Five score (Murakami and Murakami, 1999), Age_i is the individual's age, $SelfEduHistory_i$ and $ParentsEduHistory_i$ are the individual's scores of educational history and his/her parents' score of educational history, respectively (Okada et al., 2014), $Income_i$ and $Occupation_i$ are the individual's income and occupation, respectively (Ganzeboom et al., 1992), and $SubjectiveSES_i$ is the individual's subjective socioeconomic status (Adler et al., 2000). Sex_i is the binary variable representing individual (i)'s sex (men = 1) and used to represent interactive effects with Big Five scores and socioeconomic status variables. The multiple linear regressions were conducted using the glm package based on the R statistical package (R Development Core Team, 2008).

Results

fMRI study

Behavioral results of the fMRI study. We first performed a logistic regression analysis to determine whether reward, guilt, and inequity had an effect on participant behavior (Cooperate or Defect). Behavioral data from the fMRI experiment (n = 52) were analyzed using the utility function, which comprises a linearly weighted sum of reward, guilt, and (absolute) inequity (for details, see Materials and Methods). The β values of the three predictors, Reward, Guilt and Inequity, were positive and significant ($P < 0.001$, Table 2), indicating that they all played critical roles in the current task.

Having confirmed that these three factors play crucial roles in the current task, we then compared β values between men and women. This analysis showed that the β value of Guilt (called $\beta(Guilt)$ hereafter) of men was significantly higher ($t(41.6) = 2.05$, $P = 0.046$; Fig. 2***A***) than that of women, whereas the β value of Inequity (called $\beta(Inequity)$ hereafter) of women was significantly higher ($t(48.7) = 2.11$, $P = 0.039$; Fig. 2***A***) than that of men. These findings show that gender differences in prosocial behavior are heterogeneous depending on the underlying motives.

For the model validation and selection, 10 possible models were compared based on the predictive negative log likelihoods by a cross-validation. This cross-validation approach for value-based decision-making allows us to avoid overfitting the data and to compare models with different numbers of parameters robustly; it has also been adopted in many recent studies (Daw, 2011; Linderman and Gershman, 2017; Park, et al., 2019; Smith, et al., 2014; Fig. 2***B***; see also Model validation and comparison in Materials and Methods). More specifically, we introduced a bootstrap sampling (500 iterations) and compared the model predictions to the held-out data across all folds based on the negative log-likelihood of the estimated model for each participant. We then selected the model with the minimum negative log-likelihood and found that the best-fit model contained three predictors: Reward, Guilt and Inequity. In addition, we compared the BIC and found not only that the best model was the same with the smaller mean BIC than

the second best model of Fehr and Schmidt (1999) (39.44 vs 40.61), but also that for 40 of the 52 participants (76.9%), the smallest BIC model was the best individual model (see Fig. 2***C***).

Finally, we examined whether the guilt aversion parameter $\beta(Guilt)$ reflects the guilt experience of the participants in the current experiment. Note that $\beta(Guilt)$ captures a decision strategy to avoid future guilt but does not directly measure guilt. To address this issue, we analyzed the relationship between $\beta(Guilt)$ and the score of the post-experiment questionnaire (see Materials and Methods for the questionnaire). Question *a* asked whether participants understood the intentions behind player A's action, question *b* asked whether participants understood that they reduced player A's payoff if they chose *Defect*, and question *c* asked whether participants felt guilty when they reduced player A's expected payoff. We found significant or marginal positive correlation between $\beta(Guilt)$ and scores for the questions (Fig. 2***D***; question *a*, P = 0.0557; question *b*, P = 0.0491; question *c*, P = 0.0451). These results indicates that the guilt aversion parameter reflects the guilt experience in the current study.

Imaging results of guilt, inequity and utility. For the imaging, we first examined the brain regions activated commonly in both genders. Similar to the logistic regression, a general linear model analysis was conducted (SPM 12) to identify brain regions whose activity was correlated with the difference in guilt and inequity between the two choice options (hereafter, we call these differences guilt and inequity, respectively, for simplicity. See GLM1 in Materials and Methods). We included guilt and inequity as additional regressors attached to the task presentation event. We found a significant correlation between the amount of guilt and activity in the bilateral DLPFC and DMPFC (right DLPFC, $P < 0.001$; left DLPFC, $P < 0.001$; DMPFC, $P < 0.001$; family-wise error (FWE) corrected; Fig. 3***A*** and Extended Data Figure 3-1). By contrast, the amount of inequity was correlated with activity in the bilateral ventral

striatum (right ventral striatum, P = 0.035; left ventral striatum, P = 0.042; small volume FWE corrected; Fig. 3***B*** and Extended Data Figure 3-2). Additionally, we confirmed that the same results were obtained even when the two parameters (Guilt and Inequity) of GLM1 were analyzed as separate GLMs.

To identify the neural substrates that integrate different types of values, such as guilt and inequity, we searched for the neural correlates of the value difference between the choice options (larger utility-smaller utility; see GLM2 in Materials and Methods). We found a significant correlation between the value difference and activity in the VMPFC (P = 0.040; small volume FWE corrected; Fig. 3***C*** and Extended Data Figure 3-3), which is consistent with previous neuroimaging studies of value-based decision-making (Hunt et al., 2012; Nicolle et al., 2012).

We next performed a PPI analysis (Friston et al., 1997) to confirm the value signals in the VMPFC reflect the value components of both Guilt and Inequity. In our behavioral hypothesis, because participants make decisions depending on both the guilt and inequity components, the VMPFC should link with both the guilt-correlated area (DLPFC and DMPFC shown in Fig. 3***A***) and inequality-correlated area (striatum shown in Fig. 3***B***). To validate this hypothesis, we estimated a PPI in which signals in the VMPFC were modulated by the Guilt or Inequity values separately for each condition (see PPI1 in Materials and Methods). More specifically, we used the VMPFC (shown in Fig. 3***C***) as the seed region to determine which other brain regions correlated with VMPFC × Guilt and VMPFC × Inequity, respectively. For the PPI of VMPFC × Guilt, this analysis revealed positive coupling between the VMPFC and the DMPFC (P < 0.001, uncorrected; Fig. 3***C*** and Extended Data Figure 3-4). Notably, the VMPFC × Guilt contrast overlapped the guilt-correlated region in Fig. 3***A*** (see Fig. 3***C***). On the other hand, for VMPFC × Inequity, we found positive coupling between the VMPFC and the striatum (P < 0.001, uncorrected; Fig. 3***C*** and Extended Data Figure 3-5). The

VMPFC × Inequity contrast overlaps the inequity-correlated region in Fig. 3***B*** at the relaxed threshold (see Fig. 3***C***; VMPFC × Inequity, uncorrected $P < 0.05$). These results suggest that the guilt difference and inequity difference between the two options computed in the DMPFC and striatum contribute to the value difference in the VMPFC for both men and women.

Imaging results of gender differences for guilt. Next, we explored the different neural substrates for guilt aversion between men and women (see GLM1.1 in Materials and Methods). Men showed higher correlation with guilt in the VMPFC ($P = 0.029$; small volume FWE corrected; Fig. 4***A*** and Extended Data Figure 4-1) compared with women, whereas there was no significant brain activity in the opposite contrast even at moderate threshold (uncorrected $P < 0.005$). Fig. 4***A*** illustrates a box plot of the contrast estimate from the VMPFC, confirming that men showed increased VMPFC activity ($t(49.9) = 3.68$, $P < 0.001$) when responding to guilt. Furthermore, this activation of the VMPFC overlapped with the activity correlated with guilt (see Fig. 4***B***), indicating that the VMPFC is sensitive to guilt aversion overall and more so in men than in women. Importantly, the VMPFC activity correlating with the value difference was spatially close but did not overlap with the VMPFC activity correlating with the gender difference (Fig. 4***C***). This observation suggests that the two VMPFC areas are involved in related but distinct computations.

Having revealed gender differences in brain activity for guilt, we next performed a PPI analysis to examine whether different neural links work for guilt aversion in men and women. More specifically, we used the VMPFC (shown in Fig. 4***A***) as a seed region to search which other cortical regions correlated with the VMPFC × guilt and then conducted two-sample *t*-tests to compare this contrast between men and women (see PPI2 in Materials and Methods). In other words, the aim of this analysis was to find differences between men and women in brain regions whose activity correlate more

strongly with VMPFC activity in accordance with the increase of guilt. This analysis revealed that connectivity between the VMPFC and the right DLPFC is significantly stronger in men than in women ($P < 0.001$, uncorrected; Fig. 4***A*** and Extended Data Figure 4-2). The active right DLPFC area overlapped with the common activity correlated with guilt for men and women (see Fig. 4***D***), suggesting that men recruit DLPFC-VMPFC connectivity more for guilt aversion, although the DLPFC works with the DMPFC to compute guilt in both genders.

The results so far suggest the possibility that the relationship of gender and guilt aversion is mediated by DLPFC-VMPFC connectivity. We therefore performed a mediation analysis to examine this hypothesis (Fig. 4***E***; see Materials and Methods). Fig. 4***E*** shows the results of this analysis and suggests that DLPFC-VMPFC connectivity is a complete mediator of the interaction between gender and guilt-aversion behavior.

In summary, according to our PPI and mediation analyses, the DMPFC works with the DLPFC to compute guilt for both genders, and the VMPFC encodes not only the value difference in collaboration with the DMPFC in both genders, but also the amount of guilt (difference) in collaboration with the DLPFC predominantly in men (Fig. 4***F***).

Imaging results of gender difference for inequity*.** We also searched for gender-related neural substrates for inequity aversion (see GLM1.2 in Materials and Methods). We found that the ventral striatum was significantly more active in women than in men ($P = 0.008$; small volume FWE corrected; Fig. 5A*** and Extended Data Figure 5-1), but there was no significant brain activity in the opposite contrast even at moderate threshold (uncorrected $P < 0.005$). The box plot of the contrast estimates in the ventral striatum (Fig. 5***A***) demonstrates that activity in this region was correlated with the increased inequity in women ($t(50.0) = 4.26$, $P < 0.001$). When we computed PPI for functional

connectivity between the ventral striatum (Fig. 5***A*** as the seed region) and other brain areas in correlation with ventral striatum × inequity (see PPI2 in Materials and Methods), no differential link was identified between men and women (at uncorrected $P < 0.001$), indicating the important role of the ventral striatum in inequity aversion. Indeed, we performed a mediation analysis for our hypothesis that the relationship of gender and inequity-aversion behavior is mediated by the ventral striatum (Fig. 5***B***; see also mediation analysis in Materials and Methods) and found that the mediation effect of the striatum is significant (a*b, $P < 0.001$).

Online study

The behavioral data of our fMRI study ($n = 52$) showed that men display greater guilt aversion than women. However, this analysis provided only weak evidence because it was based on a relatively small dataset. In addition, our fMRI results did not specify the cognitive processes underlying the gender differences in guilt aversion, although the DLPFC-VMPFC connectivity result suggested a possibility that social norms play a key role, as discussed below. To clarify these issues and make the results more robust, we conducted a large-scale online behavioral study that also considered Big Five Inventory scores (Costa et al., 1992) and socioeconomic status.

The differential use of prefrontal networks during guilt aversion may reflect different cognitive strategies used by men and women. Guilt aversion requires the ability to assess another individual's expectations and directly relates to his or her disappointment (i.e., empathy or theory of mind; Hoffman, 1982). On the other hand, guilt aversion is also a normative behavior elicited by experience (Haidt, 2003) and therefore may be executed by self-discipline without requiring empathy or inference about another's mind (i.e., rule-based decisions or systemizing). Thus, we can think of two potential cognitive underpinnings of guilt-based prosocial behavior: empathy with the disappointment of others and rule-based decisions by self-discipline. Related to this,

previous studies have reported that the link between the DMPFC and VMPFC and the one between the DLPFC and VMPFC are involved in the theory of mind (e.g., De Martino et al., 2013) and in social norms (e.g., Baumgartner et al., 2011; Hackel et al., 2020; Pornpattananangkul et al., 2018) and self-control (e.g., Hare et al., 2009; Steinbeis et al., 2016), respectively. However, it is also important to be careful of reverse inference.

Because evidence connecting these prefrontal networks and gender differences in guilt aversion remain elusive, we further investigated this issue using the Big Five Inventory (Costa et al., 1992), which defines five fundamental dimensions of personality (i.e., neuroticism, extraversion, openness, agreeableness, and conscientiousness). Because agreeableness is characterized by the understanding of others' emotions, intentions and mental states, and conscientiousness is characterized by rule-based regulation and self-discipline (DeYoung et al., 2010), we hypothesized that guilt aversion correlates with agreeableness and conscientiousness and may also explain gender differences.

Behavioral results of online study. We first conducted a model selection using the same cross-validation analysis as the fMRI study and found that as in our fMRI study the same model containing three predictors: Reward, Guilt and Inequity (Fig. 2***B***; see also Materials and Methods) was selected as the best model. The same result was also found by the BIC analysis (see Fig. 2***C***). We then performed the logistic regression comprised of reward, guilt, and inequity and found that the β values of Reward, Guilt and Inequity were positive and significant ($P < 0.001$, Table 2), indicating that they all played a critical role in the online experiment.

To identify the relationship between guilt aversion ($\beta(Guilt)$) and gender, we first performed a general linear model analysis based on the explanatory variables including the gender term (Sex; men = 1), Big Five and socioeconomic status scores

(target variable: $\beta(Guilt)$) for all participants. The coefficients of Sex was positive and significant ($P < 0.001$; for other significant coefficients, Agreeableness and Income, $P < 0.001$), demonstrating that men displayed greater guilt aversion than women, validating the behavioral result in the fMRI study with even larger data.

Next, to identify the cognitive mechanisms specific to either gender, we performed the second general linear model analysis that included interaction terms between the gender variable sex and Big Five and socioeconomic status scores (for more details, see Materials and Methods). We found that the coefficients of Agreeableness and Sex × Conscientiousness were positive and significant (Agreeableness, $P = 0.00802$; Sex × Conscientiousness, $P = 0.00712$; see Table 3). These findings support our hypothesis that for guilt aversion, both men and women utilize the empathic strategy, while men also recruit the rule-based strategy (i.e., social norms).

Discussion

In this study, in correspondence with stronger guilt aversion in men than women, we demonstrated that men recruit DLPFC-VMPFC connectivity more in the processing of guilt than women do. We also found that the DMPFC is involved in the processing of guilt and the value difference between the choice options for both men and women. The analysis of the online behavioral data of 4723 participants not only replicated the gender difference in guilt aversion, but also suggested that the stronger guilt aversion in men than women is attributable to the use of rule-based (social norm-based) strategies more, while both genders commonly utilize empathetic consideration. Previous behavioral economics studies have closely examined guilt aversion in social interactions (Bellemare et al., 2017, 2018; Charness and Dufwenberg, 2006; Khalmetski, 2016), but to our knowledge, this is the first study reporting the evidence of gender differences in guilt aversion. Additionally, we also replicated a previously reported result (Soutschek et al., 2017) that

689 women show greater activity of the ventral striatum than men for stronger inequity
690 aversion.

691 For inequity-based prosocial behaviors, previous behavioral studies have
692 reported that men choose efficient allocations, while women are more inequality-averse
693 (Croson and Gneezy, 2009; Kamas and Preston, 2015). In the ultimatum game, women
694 are significantly more likely to propose an equal split than men (Güth et al., 2007) and
695 more likely to reject lower offers than men (Solnick et al., 2001). Furthermore, in the
696 dictator game and social value orientation tasks, women are more inequality-averse in
697 their dictator-giving (Andreoni and Vesterlund, 2001; Bolton and Katok, 1995;
698 Dickinson and Tiefenthaler, 2002; Eckel and Grossman, 1998; Grosch and Rau, 2017).
699 With regard to brain function, previous studies have reported a key role of the ventral
700 striatum in resource allocation and inequity aversion. Not only is ventral striatum
701 activity positively correlated with the ratio of the payoff (i.e., the self's payoff vs. the
702 other's payoff) (Fliessbach et al, 2007), it is also activated when inequity between the
703 self and the other is reduced (Tricomi et al., 2010) and when making a decision to
704 punish someone for acting unfairly (Crockett et al., 2013). A recent study suggested that
705 activation patterns of the ventral striatum are gender-specific, being more sensitive to
706 sharing money with others in women (Soutschek et al., 2017). The present study is
707 consistent with these previous studies in the sense that women show stronger inequity
708 aversion than men, with the ventral striatum playing a critical role.

709 At the neural level, previous studies have reported that DLPFC and DMPFC
710 activity varies with guilt (Chang et al., 2011; Nihonsugi et al., 2015; van Baar et al.,
711 2019). With regard to gender differences, the current study showed that the VMPFC
712 plays a critical role in computing guilt in men. The VMPFC has been implicated in social
713 cognition (Blakemore, 2008). For instance, the VMPFC was implicated in affective
714 regulation and depression (Ressler and Mayberg, 2007), the evaluation of moral

715 dilemmas (Crockett at al. 2017), and social value decision-making (Baumgartner et al,
716 2011; Hare et al., 2010). In line with these studies, some studies showed that men with
717 right VMPFC lesions have deficits in social emotion and decision-making compared to
718 men with left VMPFC lesions, but no such difference was seen in women (Sutterer et al.,
719 2015; Tranel et al., 2005). In addition, men with right VMPFC lesions tended to show a
720 significant elevation in paranoia and introversion according to the Minnesota Multiphasic
721 Personality Inventory-2 Scale, a widely-used measure of personality and
722 psychopathology (Tranel et al., 2005). These results suggest that the right VMPFC plays
723 an important role in social decision-making in men, consistent with the present study
724 reporting that men use the right VMPFC (coordinates 10, 42, -16) more than women to
725 implement guilt aversion.

726 A previous study showed that connectivity between the VMPFC and DLPFC in
727 men is associated with normative decisions in the ultimatum game (Baumgartner et al.,
728 2011). The study recruited male subjects (n = 32) and demonstrated that repetitive
729 transcranial magnetic stimulation applied to the right DLPFC of responders in the
730 ultimatum game subsequently reduced their rejection rate (i.e., normative decision) and
731 also diminished activity in the DLPFC and VMPFC. This result is consistent with our
732 view that connectivity between the VMPFC and DLPFC plays a key role in guilt-based
733 prosocial behavior in men.

734 For the cognitive mechanisms underlying gender differences in guilt aversion, our
735 online study showed that guilt aversion in men correlates with conscientiousness. The
736 Empathizing-Systemizing theory is widely known as a measure of individual differences
737 in cognition (Baron-Cohen, 2003). Empathizing is the drive to identify another's mental
738 state and to respond with an appropriate emotion, and has a positive correlation with
739 agreeableness (Nettle, 2007; Wakabayashi and Kawashima 2015). On the other hand,
740 systemizing is defined as the drive to analyze, understand, predict, control and construct

741 rule-based systems (e.g., map-reading, physics, and mathematics), and has a positive
742 correlation with conscientiousness, which has a desire for order as one of its components
743 (Nettle, 2007; Wakabayashi and Kawashima 2015). Interestingly, several previous
744 studies showed that men are more interested in systemizing than women (Baron-Cohen,
745 2004; Greenberg et al., 2018). These behavioral backgrounds are consistent with our
746 functional connectivity result of the DLPFC-VMPFC, as this link has been associated
747 with social norms (Baumgartner et al., 2011; Hackel et al., 2020; Pornpattananangkul et
748 al., 2018). It may also be worth noting that our results suggest that guilt aversion contains
749 both empathizing (empathy or theory of mind) and systemizing (social norms)
750 components. By conducting a large-scale online behavioral study, we strengthened the
751 neuroscientific hypothesis that the DLPFC-VMPFC connectivity predominantly seen in
752 men contributes to their stronger guilt aversion by the influence of social norms.
753 Because the recruitment of an equally large sample for fMRI experiments is very
754 difficult, we believe that integrating fMRI and large-scale online experiments provides a
755 powerful tool to obtain broader and more reliable insights into human cognitions.

756 There is the possibility that a small τ may elicit an emotion other than guilt, such
757 as distrust, because we did not directly measure emotions to belief (τ_A). Distrustful
758 behavior can be perceived as hostile acts and reduce cooperation (Fehr, E., &
759 Rockenbach, B., 2003). Related to this, a previous behavioral study (Balafoutas &
760 Fornwagner, 2017) showed that there is an inverted-U shape relationship between belief
761 and guilt aversion using a simple dictator game. This relationship suggests that there is a
762 threshold beyond which guilt aversion no longer applies and higher perceived
763 expectations lead to less kind behavior on the part of the decision makers. However, this
764 phenomenon may only occur in the dictator game, because the dictator is less likely to
765 feel guilt due to the lack of a rational reason to live up to the recipient's expectationa. In
766 any case, the fact that some previous research findings did not show a linear relationship
767 between belief (τ) and cooperation is likely to reflect psychological differences in

response to the size of belief. The present study did not allow us to address these issues, because it only considered reasonably high belief (τ_A). Future research should assess emotions to beliefs more precisely.

Finally, our findings do not preclude the possibility that social environments largely contribute to the gender differences in guilt aversion instead of biological reasons. At the same time, our behavioral data (n = 4723) suggested that gender differences in guilt aversion are independent of age (see Table 3), indicating that gender differences are only weakly dependent on contemporaneous social environmental factors and more affected by long-lasting determinants such as social systems and biological factors. Therefore, further investigation is necessary to address what causes the gender differences in guilt aversion. For instance, we need to examine the behavioral and neural gender differences in guilt aversion in different cultures (i.e., South-East Asian and European countries). Such studies would provide more biological and societal insights into our understanding in the diversity of human prosocial behaviors.

Extended Data

Extended Data Figure 1-1. The actual assignment of x, y, z and τ_A for the 45 trials.

Extended Data Figure 1-2. Instructions for the first and second experiments in the fMRI study.

Extended Data Figure 3-1. Activities related with Guilt in both genders.

Extended Data Figure 3-2. Activities related with Inequity in both genders.

Extended Data Figure 3-3. Activities related with value differences in both genders.

Extended Data Figure 3-4. Results of the PPI analysis for VMPFC × Guilt in both genders.

Extended Data Figure 3-5. Results of the PPI analysis for VMPFC × Inequity in both genders.

Extended Data Figure 4-1. Differences of activities related to guilt between men and women.

Extended Data Figure 4-2. Results of the PPI analysis for guilt when testing for gender differences.

Extended Data Figure 5-1. Differences of activities related to inequity between men and women.

References

Adler NE, Epel ES, Castellazzo G, Ickovics JR (2000) Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy white women. Health Psychol 19:586-592.

Andreoni J, Vesterlund L (2001) Which is the fair sex? Gender differences in altruism. Q J Econ 116:293-312.

Balafoutas L, Fornwagner, H (2017). The limits of guilt. Journal of the Economic Science Association, 3, 137-148.

Baron-Cohen S (2004) The essential difference: Men, women and the extreme male brain. London: Penguin.

812 Baron-Cohen S, Richler J, Bisarya D, Gurunathan N, Wheelwright S (2003) The
813 systemizing quotient: an investigation of adults with Asperger syndrome or high–
814 functioning autism, and normal sex differences. Philos Trans R Soc Lond B Biol
815 Sci 358:361-374.

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

819 Battigalli P, Dufwenberg M (2007) Guilt in games. Am Econ Rev 97:170-176.

820 Battigalli P, Dufwenberg M (2009) Dynamic psychological games. J Econ Theory 144:
821 1-35.

822 Baumeister RF, Stillwell AM, Heatherton TF (1994) Guilt: an interpersonal
823 approach. Psychol Bull 115:243.

824 Baumgartner T, Knoch D, Hotz P, Eisenegger C, Fehr E (2011) Dorsolateral and
825 ventromedial prefrontal cortex orchestrate normative choice. Nat
826 Neurosci 14:1468-1474.

827 Bellemare C, Sebald A, Suetens S (2017) A note on testing guilt aversion. Game Econ
828 Behav 102: 233-239.

829 Bellemare C, Sebald A, Suetens S (2018) Heterogeneous guilt sensitivities and incentive
830 effects. Exp Econ 21: 316-336.

831 Bellemare C, Sebald A, Suetens S (2019) Guilt aversion in economics and psychology. J
832 Econ Psychol 73:52-59.

833 Blakemore SJ (2008) The social brain in adolescence. Nat Rev Neurosci 9:267-277.

834 Boorman, ED, Behrens, TE, Woolrich, MW, Rushworth, MF (2009) How green is the
835 grass on the other side? Frontopolar cortex and the evidence in favor of alternative
836 courses of action. Neuron, 62: 733-743.

837 Bolton GE, Katok E (1995) An experimental test for gender differences in beneficent
838 behavior. Econ Lett 48:287-292.

839 Bracht J, Regner T (2013) Moral emotions and partnership. J Econ Psychol 39:313-326.

840 Cai, X, Padoa-Schioppa, C (2014) Contributions of orbitofrontal and lateral prefrontal
841 cortices to economic choice and the good-to-action transformation. Neuron 81:
842 1140-1151.

843 Chang LJ, Smith A, Dufwenberg M, Sanfey AG (2011) Triangulating the neural,
844 psychological, and economic bases of guilt aversion. Neuron 70:560-572.

845 Charness G, Dufwenberg M (2006) Promises and
846 partnership. Econometrica 74:1579-1601.

847 Costa PT, McCrae RR (1992) Revised NEO Personality Inventory (NEO-PI-R) and
848 NEO Five-Factor Inventory (NEO-FFI): Professional Manual. Odessa, FL:
849 Psychological Assessment Resources.

850 Crockett MJ, Apergis-Schoute A, Herrmann B, Lieberman MD, Müller U, Robbins TW,
851 Clark L. (2013) Serotonin modulates striatal responses to fairness and retaliation in
852 humans. J Neurosci 33:3505-3513.

853 Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ (2017) Moral transgressions
854 corrupt neural representations of value. Nat Neurosci 20:879-885.

855 Croson R, Gneezy U (2009) Gender differences in preferences. J Econ Lit 47:448-474.

856 Daw ND (2011) Trial-by-trial data analysis using computational models. In: Decision
857 Making, Affect, and Learning: Attention and Performance XXIII (Delgado MR,
858 Phelps EA, Robbins TW, eds.), pp 22-32, Oxford: Oxford UP.

859 De Martino B, O'Doherty JP, Ray D, Bossaerts P, Camerer C (2013) In the mind of the
860 market: theory of mind biases value computation during financial
861 bubbles. Neuron 79:1222-1231.

862 DeYoung CG, Hirsh JB, Shane MS, Papademetris X, Rajeevan N, Gray JR (2010)
863 Testing predictions from personality neuroscience: Brain structure and the big
864 five. Psychol Sci 21:820-828.

865 Dickinson DL, Tiefenthaler J (2002) What is fair? Experimental evidence. South Econ J
866 69:414-428.

867 Eckel CC, Grossman PJ (1998) Are women less selfish than men?: Evidence from
868 dictator experiments. Econ J (London) 108:726-735.

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

871 Fehr E, Schmidt KM (2006) The economics of fairness, reciprocity and altruism–
872 experimental evidence and new theories. In: Handbook of the economics of giving,
873 altruism and reciprocity volume 1 (Kolm SC, Ythier JM, eds.), pp615-691,
874 Amsterdam: Elsevier.

875 Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A (2007) Social
876 comparison affects reward-related brain activity in the human ventral
877 striatum. Science 318:1305-1308.

878 Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological
879 and modulatory interactions in neuroimaging. Neuroimage 6:218-229.

880 Ganzeboom HB, De Graaf PM, Treiman DJ (1992) A standard international
881 socio-economic index of occupational status. Soc Sci Res 21:1-56.

882 Gao X, Yu H, Sáez I, Blue PR, Zhu L, Hsu M, Zhou X (2018) Distinguishing neural
883 correlates of context-dependent advantageous-and disadvantageous-inequity
884 aversion. Proc Natl Acad Sci U S A 115:E7680-E7689.

885 Gitelman DR, Penny WD, Ashburner J, Friston KJ (2003) Modeling regional and
886 psychophysiologic interactions in fMRI: the importance of hemodynamic
887 deconvolution. Neuroimage 19:200-207.

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

891 Greenberg DM, Warrier V, Allison C, Baron-Cohen S (2018) Testing the empathizing–
892 systemizing theory of sex differences and the extreme male brain theory of autism in
893 half a million people. Proc Natl Acad Sci U S A 115:12152-12157.

894 Grosch K, Rau HA (2017) Gender differences in honesty: The role of social value
895 orientation. J Econ Lit 62:258-267.

896 Güth W, Schmidt C, Sutter M (2007) Bargaining outside the lab–a newspaper experiment
897 of a three-person ultimatum game. Econ J (London) 117:449-469.

898 Hackel LM, Wills JA, Van Bavel JJ (2020) Shifting prosocial intuitions: neurocognitive
899 evidence for a value-based account of group-based cooperation. Soc Cogn Affect
900 Neur 15: 371-381.

Haidt J (2003) The moral emotions. In: Handbook of Affective Sciences (Davidson RJ, Scherer KR, Goldsmith HH, eds.), pp852–870, Oxford: Oxford UP.

Hare TA, Camerer CF, Knoepfle DT, O'Doherty JP, Rangel A. (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J Neurosci 30:583-590.

Hare TA, Camerer, C. F., & Rangel, A. (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324: 646-648.

Haruno M, Frith CD (2010) Activity in the amygdala elicited by unfair divisions predicts social value orientation. Nat Neurosci 13:160-161.

Haruno M, Kimura M, Frith CD (2014) Activity in the nucleus accumbens and amygdala underlies individual differences in prosocial and individualistic economic choices. J Cogn Neurosci 26:1861-1870.

Hoffman M (1982) Development of prosocial motivation: empathy and guilt. In: The Development of Prosocial Behavior (Eisenberg N, ed), pp281–313, New York: Academic Press.

Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MF, Behrens TE (2012) Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci 15:470-476.

Kamas L, Preston A (2015) Can social preferences explain gender differences in economic behavior? J Econ Behav Organ 116:525-539.

Keil B, Blau JN, Biber S, Hoecht P, Tountcheva V, Setsompop K, Triantafyllou C, Wald LL (2013) A 64-channel 3T array coil for accelerated brain MRI. Magn Reson Med 70:248–258.

924 Khalmetski K (2016) Testing guilt aversion with an exogenous shift in beliefs. Game
925 Econ Behav 97: 110-119.

926 Kosmidis I (2019) brglm: Bias Reduction in Binary-Response Generalized Linear
927 Models. https://cran.r-project.org/package=brglm.

928 Linderman SW, Gershman SJ (2017) Using computational theory to constrain statistical
929 models of neural data. Curr Opin Neurobiol 46: 14-24.

930 Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for
931 neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data
932 sets. Neuroimage 19:1233-1239.

933 Murakami Y, Murakami C (1999) The standardization of a Big Five personality
934 inventory for separate generations. Pasonariti Kenkyu 8:32-43.

935 Nettle D (2007) Empathizing and systemizing: What are they, and what do they
936 contribute to our understanding of psychological sex differences? Br J
937 Psychol 98:237-255.

938 Nicolle A, Klein-Flügge MC, Hunt LT, Vlaev I, Dolan RJ, Behrens TE (2012) An agent
939 independent axis for executed and modeled choice in medial prefrontal
940 cortex. Neuron 75:1114-1121.

941 Nihonsugi T, Ihara A, Haruno M (2015) Selective increase of intention-based economic
942 decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. J
943 Neurosci 35:3412-3419.

944 Okada N, Kasai K, Takahashi T, Suzuki M, Hashimoto R, Kawakami N (2014) Brief
945 rating scale of socioeconomic status for biological psychiatry research among

946 Japanese people: A scaling based on an educational history. Japanese Journal of
947 Biological Psychiatry 25:115-117.

948 Park SA, Sestito M, Boorman ED, Dreher J (2019) Neural computations underlying
949 strategic social decision-making in groups. Nat Commun 10: 5287.

950 Pornpattananangkul N, Zhen S, Yu R (2018) Common and distinct neural correlates of
951 self-serving and prosocial dishonesty. Hum Brain Mapp 39: 3086-3103.

952 Ressler KJ, Mayberg HS (2007) Targeting abnormal neural circuits in mood and anxiety
953 disorders: from the laboratory to the clinic. Nat Neurosci 10:1116-1124.

954 Smith A, Bernheim BD, Camerer CF, Rangel A (2014) Neural Activity Reveals
955 Preferences without Choices. Am Econ J-Microecon 6: 1-36.

956 Solnick SJ (2001) Gender differences in the ultimatum game. Econ Inq 39:189-200.

957 Soutschek A, Burke CJ, Beharelle AR, Schreiber R, Weber SC, Karipidis II, ten Velden J,
958 Weber B, Haker H, Kalenscher T, Tobler PN (2017) The dopaminergic reward
959 system underpins gender differences in social preferences. Nat Hum
960 Behav 1:819-827.

961 Steinbeis N, Haushofer J, Fehr E, Singer T (2016) Development of behavioral control and
962 associated vmPFC–DLPFC connectivity explains children's increased resistance to
963 temptation in intertemporal choice. Cereb Cortex 26: 32-42.

964 Sutterer MJ, Koscik TR, Tranel D (2015) Sex-related functional asymmetry of the
965 ventromedial prefrontal cortex in regard to decision-making under risk and
966 ambiguity. Neuropsychologia 75:265-273.

967 Suzuki, S, Adachi, R, Dunne, S, Bossaerts, P, O'Doherty, JP (2015) Neural mechanisms
968 underlying human consensus decision-making. Neuron, 86: 591-602.

969 Tanaka T, Yamamoto T, Haruno M (2017) Brain response patterns to economic inequity
970 predict present and future depression indices. Nat Hum Behav 1:748-756.

971 Tranel D, Damasio H, Denburg NL, Bechara A (2005) Does gender play a role in
972 functional asymmetry of ventromedial prefrontal cortex? Brain 128:2872-2881.

973 Tricomi E, Rangel A, Camerer, CF, O'Doherty JP (2010) Neural evidence for
974 inequality-averse social preferences. Nature 463:1089-1091.

975 van Baar JM, Chang LJ, Sanfey AG (2019) The computational and neural substrates of
976 moral strategies in social decision-making. Nat Commun 10:1483.

977 van Overwalle, F. Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp.*
978 30, 829–858 (2009).

979 Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN (2008)
980 Prefrontal-subcortical pathways mediating successful emotion
981 regulation. Neuron 59:1037-1050.

982 Wakabayashi A, Kawashima H (2015) Is empathizing in the E–S theory similar to
983 agreeableness? The relationship between the EQ and SQ and major personality
984 domains. Pers Individ Dif 76:88-93.

985 Yarkoni T (2009) Big correlations in little studies: Inflated fMRI correlations reflect low
986 statistical power—Commentary on Vul et al. (2009). Perspect Psychol Sci 4:
987 294-298.

988 **Figure legends**

989 **Figure. 1.** Task design. ***A***, Design of the trust game. First, player A chooses *In* or *Out*,
990 which reveals a belief probability of the likeliness that player B will choose *Cooperate*. If
991 player A chooses *Out* (i.e., does not trust player B), player A and B receive z_A and z_B,

992 respectively. If player A chooses *In* (i.e., trusts player B), then with the knowledge of
993 player A's belief probability, player B decides whether to *Cooperate* or *Defect*. If player
994 B chooses *Defect*, players A and B receive y_A and y_B, respectively; if *Cooperate*, players
995 A and B receive x_A and x_B, respectively. The actual assignment of x, y, z and
996 τ_A for the 45 trials is shown in Extended Data Figure 1-1. ***B***, An outline and example of
997 experimental trials. After the green fixation period (2–5 s; cue phase), a task condition is
998 presented for 5 s (choice phase), and participants are asked to press the Cooperate or
999 Defect button (blue and red, respectively). Then, a yellow fixation cross is shown for 6-15
1000 s (rest phase). ***C***, An illustration of the complete experimental paradigm. For both the
1001 fMRI and online studies, in the first experiment, participants (as player A) chose *In* or *Out*
1002 and reveal their belief probability that player B would choose Cooperate. In the second
1003 experiment, participants (as player B) chose to Cooperate or Defect. Participants make
1004 their decisions while being scanned in the fMRI experiment. Instructions for the first and
1005 second experiments are shown in Extended Data Figure 1-2.

1006

1007 **Figure 2.** Behavioral results. ***A***, In the fMRI study (n = 26 men, 26 women), the beta
1008 value for guilt was higher in men than in woman (P = 0.046, t-test), whereas the beta
1009 value for inequity was higher in women than in men (P = 0.039, t-test). ***B***, We validated
1010 and compared the performance of 10 models using the repeated three-fold
1011 cross-validations and found that the model containing three predictors (Reward, Guilt
1012 and Inequity) was best for both fMRI and online studies. Rw: Reward; Gu: Guilt; Iq:
1013 Inequity; Ip: Inequity-positive; In: Inequity-negative. ***C***, BIC also selected the same
1014 model (i.e., RwGuIq in ***B***), with the second best being the Fehr and Schmidt type model
1015 (i.e., RwGuIpIn in ***B***). For the selected model, a majority of participants exhibited the
1016 smallest BIC value for both the fMRI and online experiments. ***D***, $\beta(Guilt)$ had a
1017 significantly or marginally positive correlation with questions *a*, *b*, and *c*.

Figure 3. Activities correlated with Guilt, Inequity and Utility in both genders. ***A***, Activities in the right and left DLPFC and DMPFC were correlated with guilt (right DLPFC, P < 0.001; left DLPFC, P < 0.001; DMPFC, P < 0.001). Activities related with Guilt in both genders are listed in Extended Data Figure 3-1. ***B***, The bilateral ventral striatum activity was correlated with inequity (right ventral striatum, P = 0.035; left ventral striatum, P = 0.042). Activities related with Inequity in both genders are listed in Extended Data Figure 3-2. **C**, (left) Activity in the VMPFC was positively correlated with the value difference (larger utility-smaller utility) (P = 0.040, see also Extended Data Figure 3-3). (Top right) Overlay of the VMPFC × Guilt cluster (green) and the Guilt-correlated region shown in Fig. 3**A** (red). These two areas overlap in the DMPFC (brown). For display purposes, we used a threshold of P < 0.001 uncorrected for the Guilt contrast, and a threshold of P < 0.005 uncorrected for VMPFC × Guilt. Results of the PPI analysis for VMPFC × Guilt in both genders are summarized in Extended Data Figure 3-4. (Bottom right) Overlay of the VMPFC× Inequity cluster (green) and the Inequity-correlated region shown in Fig. 3**B** (red). These two areas overlap in the striatum (brown) at the relaxed threshold. For display purposes, the threshold of the VMPFC × Inequity contrast is uncorrected P < 0.05. Results of the PPI analysis for VMPFC × Inequity in both genders are summarized in Extended Data Figure 3-5.

Figure 4. Results of gender differences for guilt in neural activity. ***A***, Men showed greater VMPFC activity than women (P = 0.029). As displayed in the box plot, the extracted contrast estimates in the VMPFC demonstrate that men showed increased VMPFC activity in response to guilt (P < 0.001, t-test). Importantly, the VMPFC seed exhibited positive correlation with activity in the right DLPFC as guilt increases for men but not for women (P < 0.001, uncorrected). Differences of activities related to guilt between men and women are listed in Extended Data Figure 4-1. ***B***, Overlay of the

VMPFC, which is related to gender difference in Guilt (blue), and the Guilt-correlated region (red). For display purposes, the threshold for the Guilt areas is $P < 0.001$ uncorrected and the VMPFC threshold is $P < 0.005$ uncorrected. The activation of the VMPFC involved in gender difference in Guilt largely overlaps with the clusters of activation correlated with guilt (overlap area; brown). ***C***, Overlay of the VMPFC cluster shown in Fig. 3***C***, which was positively correlated with the value difference (green), and the VMPFC cluster shown in Fig. 4***A***, which showed differential activation in the guilt contrast (men > women; blue). These two areas are close, but do not overlap. ***D***, Using a PPI analysis, a comparison of men and women showed enhanced functional connectivity of the VMPFC with the right DLPFC during the processing of guilt only in men (orange areas). This activation area (DLPFC) largely overlaps with the clusters of activation correlated with guilt shown in Fig. 3***A*** (shown in this figure as red areas). Results of the PPI analysis for guilt when testing for gender differences are shown in Extended Data Figure 4-2. ***E***, Mediation analysis of the relationship of gender, DLPFC-VMPFC connectivity and $\beta(Guilt)$ shows that DLPFC-VMPFC connectivity is a complete mediator of the interaction between gender and guilt-aversion behavior. Path coefficients are shown next to arrows with standard errors in parentheses. *$P <$ 0.05, ***$P < 0.001$. ***F***, Diagram summarizing the results of our analyses. Activities in the DLPFC and DMPFC were correlated with guilt in both genders. The blue line represents a stronger connectivity between the VMPFC and right DLPFC in men than in women depending on VMPFC × guilt, and the green line represents stronger positive coupling between the VMPFC and DMPFC depending on VMPFC × value difference.

Figure 5. Results of gender differences in neural activity for inequity. ***A***, Women showed greater ventral striatum activity than men ($P = 0.008$). The box plot illustrates the contrast estimates in the right ventral striatum and shows that only women showed increased activity in response to inequity ($P < 0.001$, *t*-test). Differences of activities related to inequity between men and women are summarized in Extended Data Figure

5-1. ***B***, A mediation analysis shows that the mediation effect of the striatum is significant (a*b, $P < 0.001$). Path coefficients are shown next to the arrows with standard errors in parentheses. *$P < 0.05$, ***$P < 0.001$.

Table 1. Descriptive statistics for online sample (between genders).

	Men (n = 1986)	Women (n = 2737)
Variable	Mean (SD)	Mean (SD)
Age	39.585 (15.318)	36.751 (15.273)
Neuroticism	47.433 (9.7900)	46.731 (9.9606)
Extraversion	45.426 (9.1431)	46.244 (9.2942)
Openness	50.699 (9.4570)	47.640 (9.3566)
Agreeableness	42.780 (10.636)	44.333 (10.448)
Conscientiousness	49.068 (9.3706)	48.076 (9.3299)
SelfEduHistory	5.3197 (1.1422)	5.0431 (1.0410)
ParentsEduHistory	4.6511 (1.4190)	4.7947 (1.3323)
Income	2.7296 (1.4772)	1.6153 (0.9299)
Occupation	2.2477 (1.6859)	3.4439 (1.8253)
Subjective SES	5.2513 (2.0408)	5.2700 (1.7901)

Notes: SD: Standard deviation. All scores were raw values

Table 2. Logistic regression models predicting decision to Cooperate or Defect.

	Dependent variable: $logit(P_{B,Cooperate})$	
Explanatory variable	fMRI	Online
Reward	0.0033317***	0.0098942***
Guilt	0.0014490***	0.0029833***
Inequity	0.0011182***	0.0059259***
Constant	−0.47831**	0.0973352***
McFadden's R^2	0.09147	0.01479

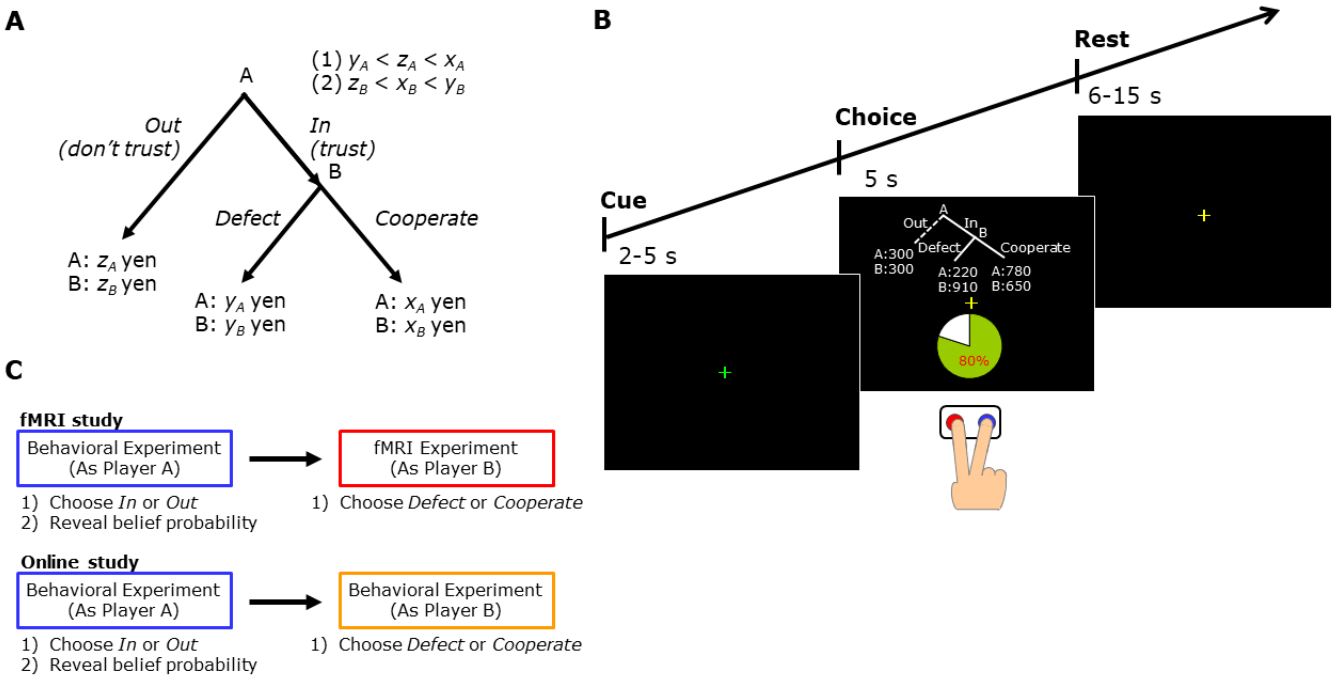
Observations	2340	212535

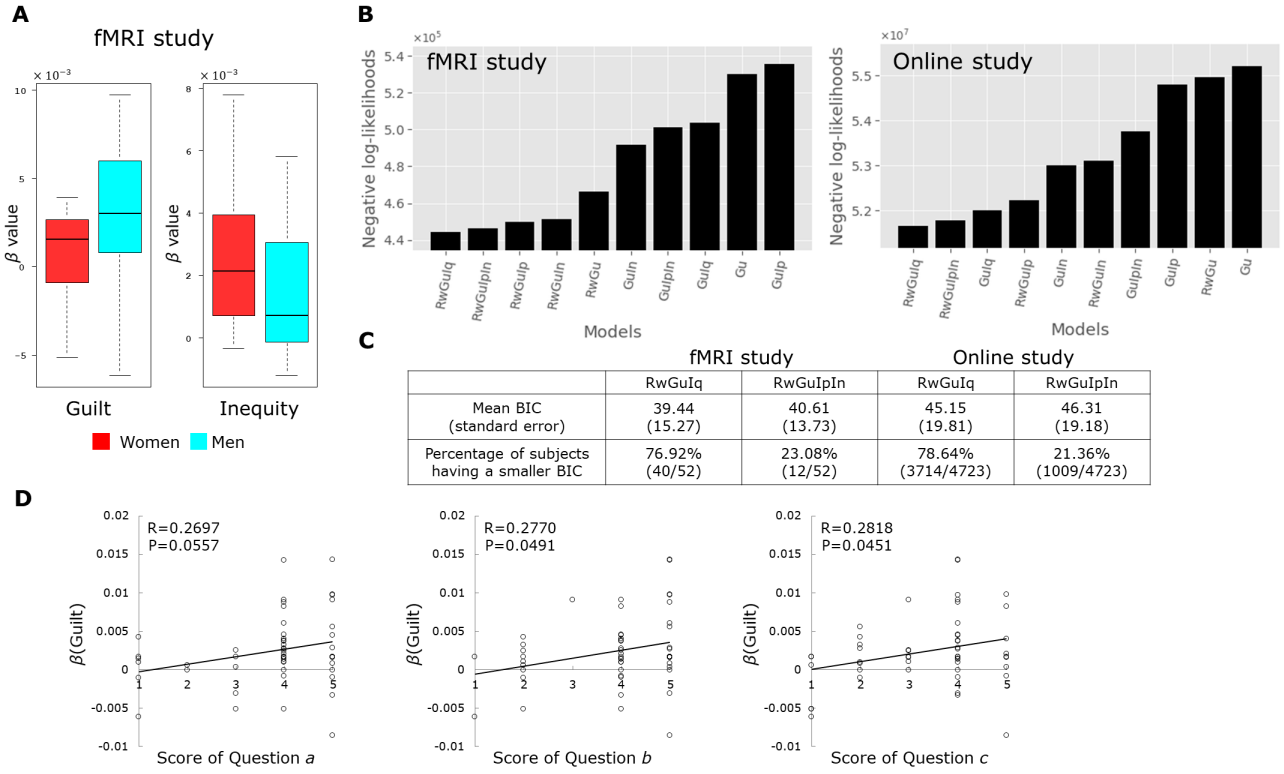
Note: Significance: $^{***}P < 0.001$. $^{**}P < 0.01$.

Table 3. GLM analyses of Guilt.

Explanatory variable	Dependent variable: beta value for Guilt	
	Coefficient	Standard errors
Neuroticism	0.0002667	0.0006187
Extraversion	-0.0008909	0.0006630
Openness	0.0004214	0.0007000
Agreeableness	0.0016424**	0.0006193
Conscientiousness	0.0009482	0.0006558
Age	0.0006831	0.0006325
SelfEduHistory	-0.0007013*	0.0005958
ParentsEduHistory	0.0006947	0.0006141
Income	-0.0001091	0.0008564
Occupation	0.0008336	0.0007700
SubjectiveSES	-0.0014377*	0.0006257
Sex × Neuroticism	0.0019212	0.0029112
Sex × Extraversion	-0.0020649	0.0032372
Sex × Openness	0.0004377	0.0036139
Sex × Agreeableness	-0.0035545	0.0025872
Sex × Conscientiousness	0.0089586**	0.0033272
Sex × Age	-0.0028621	0.0016136
Sex × SelfEduHistory	-0.0005493	0.0026958
Sex × ParentsEduHistory	0.0008173	0.0020632
Sex × Income	-0.0013636	0.0015921
Sex × Occupation	0.0002769	0.0013028
Sex × SubjectiveSES	-0.0005761	0.0018053
Adjusted R^2	-0.07467266	
Observations	4723	

Notes: Significance level: * 0.05, ** 0.01.





C

	fMRI study		Online study	
	RwGuIq	RwGuIpIn	RwGuIq	RwGuIpIn
Mean BIC (standard error)	39.44 (15.27)	40.61 (13.73)	45.15 (19.81)	46.31 (19.18)
Percentage of subjects having a smaller BIC	76.92% (40/52)	23.08% (12/52)	78.64% (3714/4723)	21.36% (1009/4723)