Behavioral/Cognitive

Selective Increase of Intention-Based Economic Decisions by Noninvasive Brain Stimulation to the Dorsolateral Prefrontal Cortex

Tsuyoshi Nihonsugi,[1,2] **Aya Ihara,**[2] **and Masahiko Haruno**[2,3,4,5]

[1]Department of Economics and Information, Gifu Shotoku University, Gifu 500-8288, Japan, [2]Center for Information and Neural Networks, National Institute of Information and Communications Technology, Osaka 565-0871, Japan, [3]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan, [4]Graduate School of Frontier Biosciences, Osaka University, Osaka 565-0871, Japan, and [5]Brain Science Institute, Tamagawa University, Tokyo 194-8610, Japan

The intention behind another's action and the impact of the outcome are major determinants of human economic behavior. It is poorly understood, however, whether the two systems share a core neural computation. Here, we investigated whether the two systems are causally dissociable in the brain by integrating computational modeling, functional magnetic resonance imaging, and transcranial direct current stimulation experiments in a newly developed trust game task. We show not only that right dorsolateral prefrontal cortex (DLPFC) activity is correlated with intention-based economic decisions and that ventral striatum and amygdala activity are correlated with outcome-based decisions, but also that stimulation to the DLPFC selectively enhances intention-based decisions. These findings suggest that the right DLPFC is involved in the implementation of intention-based decisions in the processing of cooperative decisions. This causal dissociation of cortical and subcortical backgrounds may indicate evolutionary and developmental differences in the two decision systems.

Key words: **computational modeling; cooperation; fMRI; social neuroscience; tDCS**

Introduction

How we behave during social interactions depends not only on the behavioral outcome but also on the underlying intention of others (Weber, 1919; Falk et al., 2003; McCabe et al., 2003; Fehr and Schmidt, 2006; Buckholtz and Marois, 2012). In other words, what a person believes are the beliefs of others affects how that person makes choices even with the same contingency. Indeed, ample economic behavioral evidence has shown that two systems, an outcome-based (distributional preferences) and an intention-based (belief-dependent) system, work during social decision making (Fehr and Schmidt, 2006). However, although the dissociation of the outcome-based and intention-based decisions seems to be critical in many real-life situations, such as in determining how to allocate money (Falk et al., 2003; McCabe et al., 2003; Fehr and Schmidt, 2006), deciding how to make judicial decisions (Buckholtz and Marois, 2012), and in deciding whom to vote for, very few studies have attempted to dissociate the neural substrates of the two systems.

Guilt aversion is a representative and quantifiable form of an intention-based economic decision in which an individual dislikes disappointing others relative to what others believe they should receive (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006). More specifically, the belief of the other person (e.g., player A) is formalized as the product of his (A's) belief probability of the partner's (e.g., player B's) cooperation and his (A's) reward obtained from B's cooperation. Guilt associated with disappointing the other is then defined as the difference between the other's belief and the actual outcome (see Materials and Methods for details). By contrast, inequity aversion (Fehr and Schmidt, 1999; Knoch and Fehr, 2007) is the most commonly perceived form of outcome-based decision making, which is defined as the propensity to avoid an imbalance between outcomes for the self and the other.

Previous imaging studies have reported activity in largely overlapping brain structures for both guilt aversion and inequity aversion. Specifically, a functional magnetic resonance imaging (fMRI) study of guilt aversion revealed that the insula, supplementary motor area, dorsolateral prefrontal cortex (DLPFC), and temporal parietal junction are involved (L.J. Chang et al., 2011), while studies of inequity aversion (Sanfey et al., 2003; Hsu

Received Sept. 18, 2014; revised Jan. 1, 2015; accepted Jan. 6, 2015.

Author contributions: T.N. and M.H. designed research; T.N. and A.I. performed research; T.N. and M.H. analyzed data; T.N. and M.H. wrote the paper.

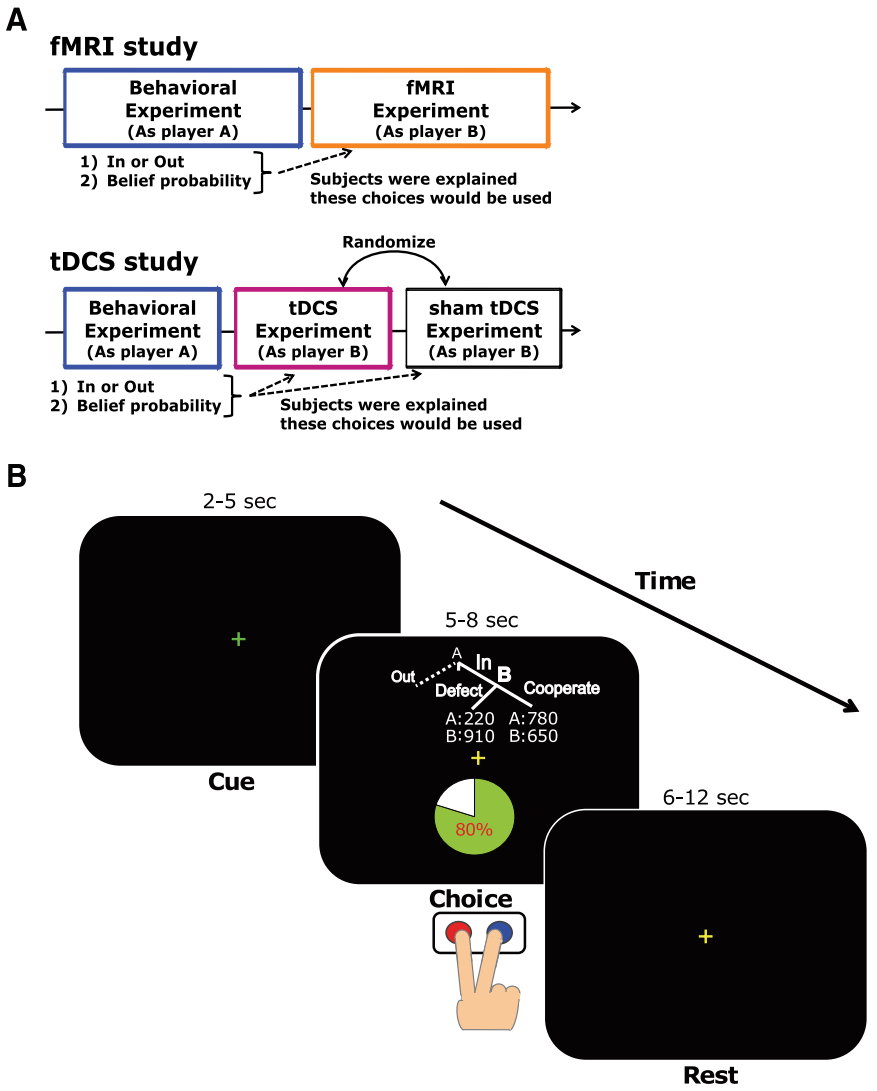
This work was supported by Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency (M.H.), a Kakenhi grant from the Ministry of Education in Japan (#22300139; M.H.), Osaka University COI (M.H.), and a grant-in-aid for Japan Society for the Promotion of Science Fellows (T.N.). We thank Peter Dayan for insightful discussion, Peter Karagiannis for editing the manuscript, and Tomoki Haji for technical assistance with the fMRI.

The authors declare no competing financial interests.

Correspondence should be addressed to either of the following: Masahiko Haruno, Center for Information and Neural Networks, NICT, 1-4 Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: mharuno@nict.go.jp; or Tsuyoshi Nihonsugi, Department of Economics and Information, Gifu Shotoku University, 1-38 Nakauzura, Gifu-shi, Gifu 500-8288, Japan. E-mail: t.nihonsugi@gmail.com.

DOI:10.1523/JNEUROSCI.3885-14.2015

 0270-6474/15/53412-08$15.00/0



correlation	Guilt	Inequity
Reward	0.283	0.015
Guilt	-	-0.045

Figure 1. Task design. ***A***, Illustration of a whole experimental paradigm. For the fMRI study, participants (as player A) choose In or Out and reveal their belief probability that player B would choose Cooperation in the behavioral experiment. On different days, participants (as player B) decide to cooperate or defect in the fMRI experiment. For the tDCS study, in the behavioral experiment, participants (as player A) play the same task as the first part of the fMRI study. On different days, the tDCS and sham tDCS experiments are conducted. Participants play the trust game as player B while receiving either anodal or sham tDCS to the right DLPFC. The orders of the anodal and sham tDCS are counterbalanced across participants. ***B***, Trust game task. After the green fixation period (2–5 s; Cue phase), a task condition is presented from 5 to 8 s (Choice phase), and participants are asked to press the Cooperate or Defect button (red and blue). A yellow fixation cross then appears from 6 to 12 s (Rest phase). ***C***, The correlation coefficients between Reward, Guilt, and Inequity. There were no significant correlations ($p > 0.05$).

et al., 2008; Haruno and Frith, 2010; Tricomi et al., 2010; Haruno et al., 2014) showed that the insula, DLPFC, anterior cingulate cortex, ventral striatum, and amygdala are involved. Many brain structures are commonly activated and it is difficult to determine from these studies whether the two systems share a core neural computation. To investigate this question, we introduced a novel task design incorporating a trust game so that the two systems could be separated and quantified using a computational model. Here, we investigated whether the two systems are causally dissociable in the brain by conducting fMRI and transcranial direct current stimulation (tDCS) experiments of the newly developed trust game task.

Materials and Methods

Participants

For the fMRI study, 69 volunteers participated in the behavioral experiments, 49 of whom participated in the fMRI experiments. Data from eight participants were excluded from the analysis of the fMRI experiments because of excessive head movement (i.e., >3 mm) or misunderstanding of the task instruction. Consequently, we analyzed data from 41 participants (mean age, 20.7 years; SD, 1.4 years; 19 male and 22 female). For the tDCS study, the participants were 22 healthy right-handed different volunteers (mean age, 20.5 years; SD, 1.5 years; 13 male and 9 female). Informed consent was obtained from all the participants, and the experimental protocol was approved by the ethics committees of the National Institute of Information Technology and Tamagawa University.

Experimental paradigm and task

We conducted two studies: an fMRI study and a tDCS study (Fig. 1*A*). The fMRI study had two parts: the behavioral experiment and the fMRI experiment. The tDCS study had three parts: the behavioral experiment, the tDCS experiment, and the sham tDCS experiment. We will first explain our task (a trust game) and then move on to the details of the procedures of the fMRI and tDCS studies.

Trust game. Participants performed a trust game adapted from the task originally used by Charness and Dufwenberg (2006). The trust game has two players: player A and player B (Fig. 1*B*). Player A chooses either an Out or In option, and is simultaneously asked to reveal his or her belief about the probability (from 0 to 100%) that Player B would choose Cooperation (probability τ_A). In other words, τ_A is player A's belief that player B will cooperate. If player A chooses Out, player A receives money z_A and player B receives z_B. If player A chooses In, then player B must choose either Cooperate (the term "Roll" was used in the original notation; Charness and Dufwenberg, 2006) or Defect by referencing the allocation of the monetary payoffs and τ_A. 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 trial shown in Figure 1*B*, the belief probability of A is 80%, and if B defects, player A and player B receive ¥220 ($1 is equivalent to ~¥100 yen) and ¥910, respectively. Otherwise, they receive ¥780 and ¥650, respectively, if B cooperates.

There are several features among the payoffs: (1) $y_A < z_A < x_A$: in this condition it is necessary to signal player A's trust (cooperative) message to player B when player A chooses In; and (2) $z_B < x_B < y_B$: this is necessary to make player B feel guilt upon disappointing the other player relative to the other's belief of what he or she will receive. The actual

assignment of x, y, and z are determined so that they do not reveal correlations among Reward, Guilt, and Inequity, which were calculated based on the values of x, y, and z (Fig. 1*C*). The ranges of x, y, and z were as follows: x_A, 410–1760; x_B, 180–1020; y_A, 220–720; y_B, 510–1270; z_A, 300–800; z_B, 100–800. The standard economic solution to this game is based on backward induction. If player B is rational and selfish, player B chooses Defect. Player A, anticipating this action, will choose Out and produce the allocation of the money payoffs z_A and z_B.

Procedure of the fMRI study. In the first (behavioral) experiment, >20 participants (i.e., 26, 22, and 21) were invited into a room and received written instructions of the rules and procedure of the trust game. Every participant played the trust game as player A and experienced one trial. Participants were instructed that player B was another participant who would take part in the second experiment. However, player A was not informed of player B's identity. Participants chose Out or In, and were simultaneously asked to reveal their belief about the probability (from 0 to 100%) that player B would choose Cooperation. More specifically, participants were asked to answer the following question: "With how much probability do you believe that player B will choose Cooperation in the second experiment? Please report your belief probability from 0 to 100% by 10% steps." Participants were informed that these choices would be used when player B made his or her choice in the second experiment.

The second (fMRI) experiment was conducted on average 8 d (range, 2–10 d) after the first experiment. Each participant was provided with written instructions that explained the rules and procedure of the trust game. Every participant experienced 45 trials. All participants played the game as player B. In each trial, participants (player B) were presented a choice of monetary allocation between A and B with A's belief probability of how likely player B would cooperate (as participants were instructed to assume that player A chose In in this experiment, the Out option was illustrated as a dashed line in Fig. 1*B*). Participants were informed that the other participant (player A) was different in each trial and that the pairings were anonymous (player A; the order of the exposure was randomized across participants). This experimental setting was realistic, because participants themselves took part in the first experiment and made decisions as player A. We did not provide any feedback to participants. After the instructions, participants were briefed about the rules of the game from the experimenter and were then tested to confirm that they understood the rules. The participants were then invited into the scanning room individually and practiced the game using the response buttons on the scanner.

Functional images were acquired as participants played the game. Timelines for a trial are shown in Figure 1*B*. Each trial began with a 2–5 s preparation interval at which time a green fixation was presented for the first 1 s and a yellow fixation (Cue phase) for the remainder. Then, participants were presented with the trust game, which included the allocation of monetary payoffs for each choice and player A's belief, and were required to select Cooperate or Defect by pressing the appropriate button at 5–8 s (Choice phase). In each trial, participants made their choice on the assumption that player A chose In. Then, a fixation cross was presented for a variable time period from 8 to 13 s (Rest phase). The total time of the fMRI session was 870 s.

After scanning, all participants returned the questionnaire and received a cash payment. Payment to the participants was proportional to the number of payoffs earned during the experiment.

Procedure of the tDCS study. The tDCS study was divided into three parts done on different days. In the first part, >10 participants (i.e., 13 and 12) were invited into a room and received instructions on the rules of the game. This task and procedure were the same as the first part of the fMRI study.

The second part was conducted 1 or 2 d after the first part. Participants played the trust game as player B while receiving either anodal or sham tDCS to the right DLPFC. The task was the same as the second experiment of the fMRI study, but only the intertrial intervals were shortened (the total time was 552 s). Then, participants took part in another tDCS experiment 1 or 2 d later (third part). Thus, our tDCS study used a within-participants design. The order of the anodal and sham tDCS were counterbalanced across participants. After completing the third part, participants received a cash payment depending on the money earned during the games.

Model and analysis

Guilt aversion and inequity aversion. Guilt aversion (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; L.J. Chang et al., 2011) assumes that an individual dislikes disappointing another's belief. This model includes social pressure on player B if the profile (In, Defect) is played. Player B is assumed to believe that if player A chooses In, player A believes that he or she will get a return $\tau_A \cdot x_A$ because the setting of player A's payoff is $y_A < z_A < x_A$. The difference, $\tau_A \cdot x_A - y_A$, which is nonnegative in our settings, can measure how much player B believes that he or she disappoints player A relative to player A's belief if player B were to choose Defect. In other words, the difference $\tau_A \cdot x_A - y_A$ is the size of guilt that player B experiences.

However, in addition to this Guilt model, the expected payoff model $[\tau_A \cdot x_A + (1 - \tau_A)y_A - y_A]$ is also a plausible candidate to explain the behavioral data. Therefore, we compared the Bayesian information criterion (BIC) of the Guilt model and the expected payoff model and found that the former had a slightly smaller BIC (1827.442 vs 1837.379). Based on this, we judged the Guilt model to be a plausible model for the current task.

Let us assume that γ_B is the parameter that measures player B's sensitivity to guilt. If $y_B - \gamma_B \cdot (\tau_A \cdot x_A - y_A) < x_B$, a guilt-averse player will choose Cooperate. In the example trial in Figure 1*B*, if $910 - \gamma_B \cdot (0.8 \cdot 780 - 220) < 650$, player B will choose Cooperate.

By contrast, inequity aversion assumes a social preference for equitable payoffs (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Haruno and Frith, 2010; Tricomi et al., 2010; Haruno et al., 2014). An individual is inequity averse if, in addition to his monetary self-interest, his utility decreases when the allocation of monetary payoffs become different. If an inequity-averse player suffers from inequity, he chooses an option that results in a smaller difference between his and the other's monetary payoffs.

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

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

where α_B is a constant that measures player B's sensitivity to inequity. A narrowly self-interested agent is given by 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 (Luce, 1959; Hensher et al., 2005). The probability that player B chooses Cooperate can be expressed as $P_{B,\,Cooperate} = 1/1 + e^{u_B(X) - u_B(Y)}$. Based on this logistic model, we used logistic regression as follows: $Utility_t = \beta_0 + \beta_1 Reward_t + \beta_2 Guilt_t + \beta_3 Inequity_t$, where $Reward_t$ is the size of the reward and is calculated as $x_B - y_B$, at time t, $Guilt_t$ is the size of guilt and is calculated as $-[0 - (\tau_A \cdot x_A - y_A)]$, and $Inequity_t$ is the size of inequity and is calculated as $-(x_A - x_B - y_A - y_B)$. For convenience, β_1, β_2, and β_3 will be denoted as $\beta(Reward)$, $\beta(Guilt)$, and $\beta(Inequity)$, respectively hereafter.

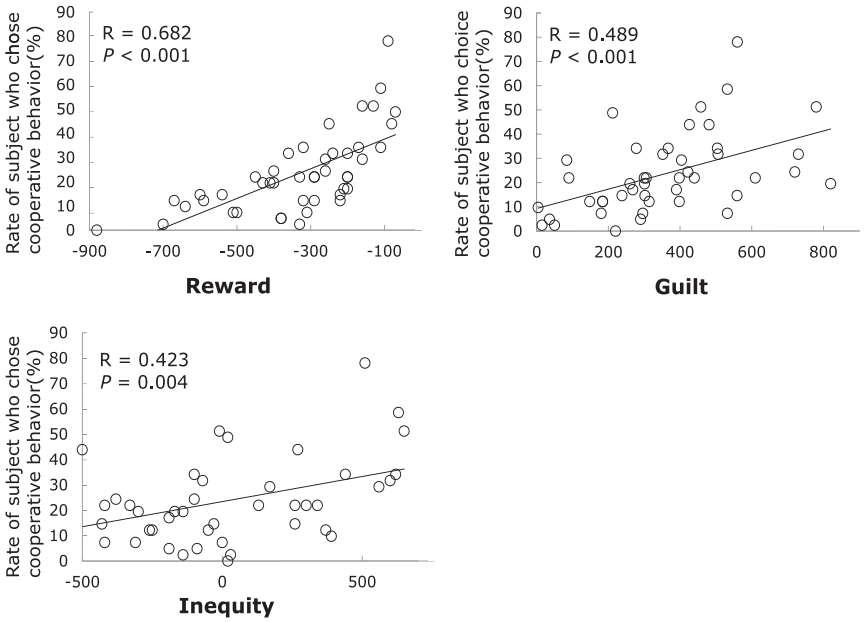
As explained previously, to dissociate the computational processes for guilt aversion and inequity aversion, values of guilt and inequity were designed to be orthogonal (the correlation coefficient of two variables was −0.045 and not significant ($p = 0.772$), and the correlation coefficients among the three explanatory variables were <0.3 with no significant differences found among them (Fig. 1*C*).

We adopted the absolute difference for Inequity in contrast to Fehr and Schmidt's (1999) models that split the inequity into positive and

A

	fMRI study	tDCS study
First half	24.39%	32.56%
Second half	22.51%	29.60%

B



C

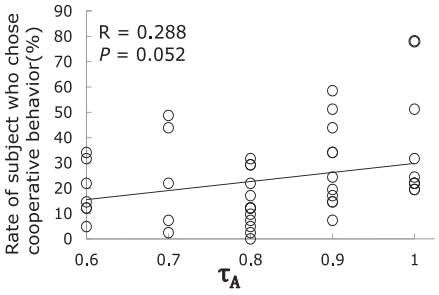


Figure 2. Behavioral results. ***A***, The cooperation ratio in the first and second half in each study. The cooperation ratio was not different between the first and second half (fMRI study, $p = 0.3505$; tDCS study, $p = 0.1747$). ***B***, Behavior. Plots show the rate that participants chose cooperation against reward, guilt, or inequity. ***C***, The relationship between the belief probability, τ_A, and the rate that participants chose cooperation showed no significant correlation ($p = 0.052$).

negative terms. We compared the BIC values for the "Inequity (absolute difference)" and Fehr and Schmidt's (1999) models using behavioral data, finding that total BIC was slightly smaller with the Inequity model (1827.442 vs 1831.376) and that 31 of the 41 participants had a smaller BIC in the Inequity model. Thus, we adopted the absolute difference for Inequity in our analysis. However, all subsequent results were the same even if the inequity was split into positive and negative terms.

All behavioral statistics were computed using the R statistical package (R Development Core Team, 2008). We used the brglm package to conduct a maximum likelihood with bias-reduction method (Kosmidis, 2013).

fMRI image acquisition. Scanning was performed on a Siemens 3T Trio scanner at the Brain Science Institute, Tamagawa University, 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; 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. We used SPM8 (http://www.fil.ion.ucl.ac.uk/spm) for MRI data preprocessing and analysis. Preprocessing included motion correction, coregistering 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. Coregistered EPIs were normalized using an anatomical normalization parameter. Spatial smoothing used an 8 mm Gaussian kernel.

General analysis methods. We focused on the neural basis of Guilt and Inequity and performed general linear model (GLM) analyses on the functional data. To model the BOLD signal driven by Guilt and Inequity, these two variables were convolved with a hemodynamic response function (spm_hrf function with TR equal to 3.0 s).

For first-level GLM analysis, the event was modeled with the duration of the response time in the Choice phase. We introduced two regressors in addition to a response-period constant regressor: (1) an hrf function for Guilt and (2) an hrf function for Inequity. In addition, regressors modeling the head motion as derived from the realignment procedure were included in the model. Serial autocorrelation was modeled as a first-order autoregressive model, and the data were high-pass filtered at a cutoff of 128 s.

These individual contrast images were then processed in a second-level random-effects analysis. To correct for multiple comparisons, we used for each contrast the familywise error (FWE) correction across the whole brain at $p < 0.05$ based on Gaussian random field theory ($k > 10$ voxels). In addition, as we had a priori hypothesis from previous studies that the amygdala and striatum are involved in inequity (Fliessbach et al., 2007; Haruno and Frith, 2010; Tricomi et al., 2010; Gospic et al., 2011; Crockett et al., 2013; Haruno et al., 2014), when analyzing Inequity, we performed region-of-interest analyses for these two subcortical regions at a threshold $p < 0.001$, uncorrected. For display purposes, we present inequity-related statistical maps at a threshold of both $p < 0.001$ and $p < 0.005$.

Methods of tDCS. Direct current was induced using two saline-soaked surface sponge electrodes of 35 cm^2, and delivered by a battery-driven, constant-current stimulator. For stimulation over the right DLPFC, the anode electrode was placed over at MNI coordinate (x, y, z) = (44, 34, 22), which corresponded to the peak voxel in the right DLPFC identified in our fMRI study. The position of the electrode was identified using the T2T-Converter (http://wwwneuro03.uni-muenster.de/ger/t2tconv/). The reference electrode (cathodal) was placed over Oz (electroencephalography 10/20 system). Participants received a constant current of 2 mA intensity. tDCS started 5 min before the task began and was delivered during the whole course of the trust game. For sham stimulation, the electrodes were placed at the same positions as active stimulation, but the stimulator was only turned

on for the initial 30 s. Thus, participants felt the initial itching sensation associated with tDCS, but received no active current for the rest of the stimulation period. This method of sham stimulation has been shown to be reliable (Gandiga et al., 2006).

Results

Behavioral results

We separated the behavioral data into the first and second half and examined whether participants had changed their behavioral selection as the game progressed (Fig. 2*A*). We confirmed that the cooperation ratio was not different between the first and second half (Fisher exact test: fMRI study, $p = 0.3505$; tDCS study, $p = 0.1747$). Then, we conducted a linear regression analysis to see whether each of the reward, guilt (A's belief about self-outcome minus A's actual outcome), and inequity (absolute difference in actual outcomes between A and B) had an effect on the participant's behavior in fMRI study. We found not only that the amount of the reward, guilt, and inequity had a significant positive correlation with the cooperation ratio (Fig. 2*B*), but that guilt had much more significant effect ($r = 0.489$, $p < 0.001$) than the belief probability displayed on the screen ($r = 0.288$, $p = 0.052$; Fig. 2*C*). This result suggests that guilt and inequity play critical roles in the current task and we subsequently analyzed behavioral and fMRI data using the utility described in the Materials and Methods section. The correlation coefficient between $\beta(Reward)$ and $\beta(Guilt)$ and between $\beta(Reward)$ and $\beta(Inequity)$ was estimated to be 0.273 ($p = 0.084$) and 0.285 ($p = 0.071$), respectively.

We also examined whether the behaviors of participants as player A and player B were correlated. There was a significant positive correlation between the belief probability revealed as player A and the frequency of cooperation when acting as player B ($r = 0.321$, $p = 0.041$). These data indicate that participants who expected more cooperation tended to be more cooperative.

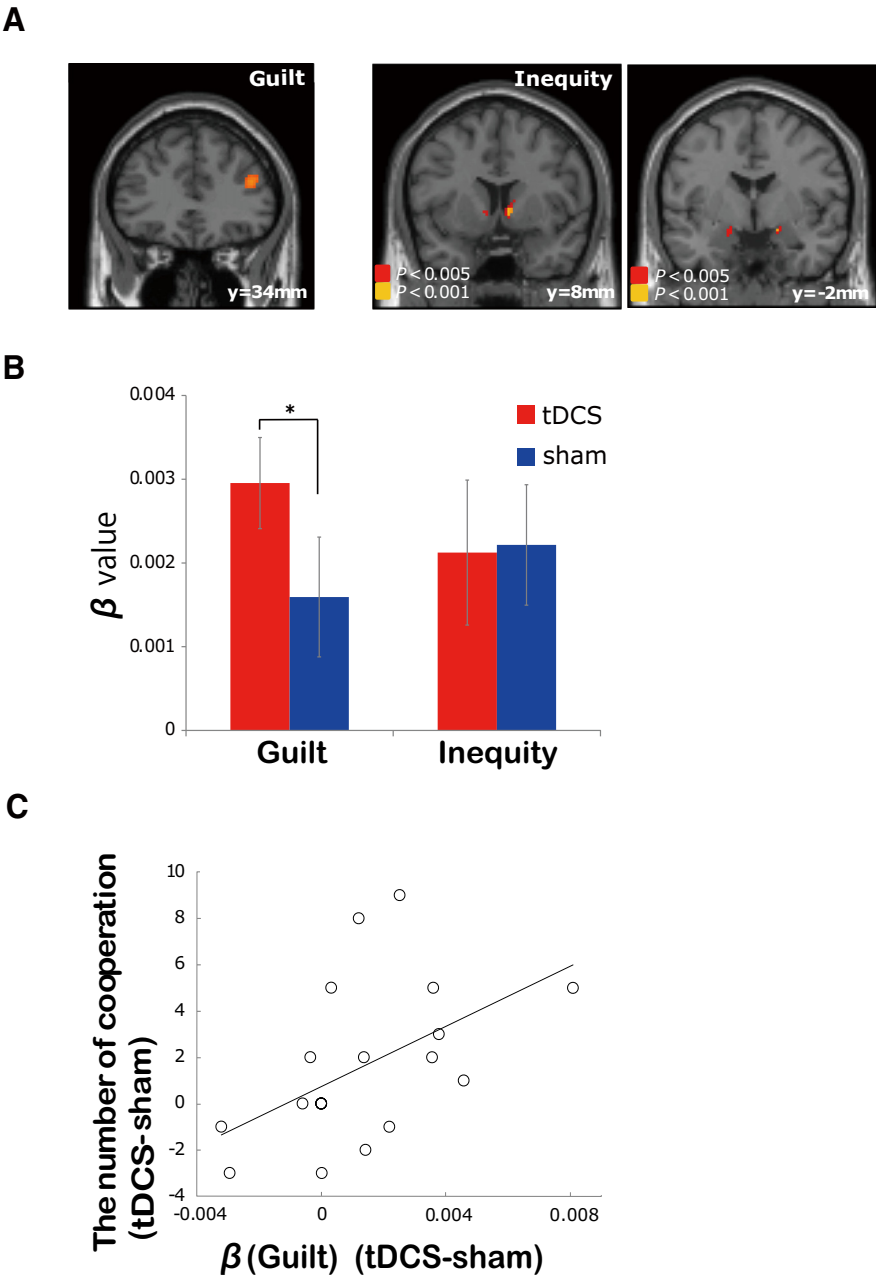


Figure 3. fMRI and tDCS results. ***A***, fMRI. Activity in the right DLPFC was correlated with guilt ($p = 0.015$, FWE), and the ventral striatum and amygdala with inequity ($p < 0.001$). ***B***, tDCS. $\beta(Guilt)$ and $\beta(Inequity)$ are displayed for the tDCS and sham conditions separately. $\beta(Guilt)$ in the tDCS condition was significantly higher than in the sham condition (*$p < 0.05$), but $\beta(Inequity)$ and $\beta(Reward)$ were not. Error bars represent SEs. ***C***, Effect of tDCS on the frequency of cooperation. A significant positive relationship ($r = 0.51$, $p = 0.019$) between the difference (tDCS-sham) in $\beta(Guilt)$ and difference (tDCS-sham) in the frequency of cooperation was seen.

fMRI results

To elucidate the neural substrates for guilt aversion and inequity aversion, a model-based GLM analysis of the fMRI data was done using SPM8 (Friston et al., 1995). More specifically, based on the behavioral analysis, we included guilt and inequity as parametric modulators to the event of game presentation whose onset and duration were the onset timing of game presentation and response time, respectively. We found significant correlation of activity with the amount of guilt in the right DLPFC ($p = 0.015$, FWE corrected; Fig. 3*A*; Table 1), and the amplitude of this DLPFC activity for each participant showed a significant correla-

Table 1. Activities correlated with the parametric modulator for size of Guilt during decision making[a]

Region	MNI coordinates (*x*, *y*, *z*)	Voxel size (voxels)	*t* value
Right DLPFC	44, 34, 22	87	7.15
Right DLPFC	44, 12, 32	78	6.06
Left primary motor cortex	−42, 4, 30	45	6.02
Right hippocampus	24, −26, 0	34	6.79
Right occipital lobe	22, −94, 6	4033	12.2
Left occipital lobe	−12, −96, −4	3933	15.0

[a]MNI coordinates (*x*, *y*, *z*) indicate the location of the peak correlation. The threshold is set at an FWE $p < 0.05$. The voxel size shows the number of suprathreshold voxels, and the *t* values are shown for the peak activation voxel.

Table 2. Activities correlated with the parametric modulator for size of Inequity during decision making[a]

Region	MNI coordinates (*x*, *y*, *z*)	Voxel size (voxels)	*t* value
Right striatum	10, 8, 0	22	3.87
Right amygdala	16, −2, −12	12	3.57

[a]MNI coordinates (*x*, *y*, *z*) indicate the location of the peak correlation. The thresholds are set at FWE $p < 0.05$ for the whole-brain analysis and at uncorrected $p < 0.001$ for region-of-interest analysis. The voxel size shows the number of suprathreshold voxels, and the *t* values are shown for the peak activation voxel. No suprathreshold cluster for the whole-brain analysis.

tion with $\beta(Guilt)$ ($r = 0.351$, $p < 0.01$). By contrast, the activity in the right DLPFC was not correlated with Reward and Inequity (even at uncorrected $p < 0.001$). These data indicate a critical role of the right DLPFC in the implementation of guilt aversion. By contrast, correlation with the amount of inequity was found in the right amygdala and ventral striatum ($p < 0.001$, uncorrected; Fig. 3*A*; Table 2). Similar to the DLPFC in the case of guilt aversion, activity in the right ventral striatum and amygdala showed significant correlation with $\beta(Inequity)$ (striatum, $r = 0.31$; amygdala, $r = 0.36$; $p < 0.01$). The activity in the striatum was not correlated with Reward ($p < 0.001$, uncorrected). This dissociation of the brain areas made us further hypothesize that if the right DLPFC is specifically related to guilt aversion, tDCS to the DLPFC may selectively increase guilt aversion-based cooperation.

tDCS results

Twenty-two different participants participated in the tDCS experiment using the same task as the fMRI experiment. In these later experiments, an anodal electrode was put on the right DLPFC of each participant [cathodal electrode on the occipital lobe (Oz); see Materials and Methods]. To identify the process that tDCS changed, we contrasted β values between the tDCS and sham conditions (the order of exposure was counterbalanced across participants), and found that $\beta(Guilt)$ in the tDCS condition was significantly higher ($p = 0.011$, paired *t* test; Fig. 3*B*) than in the sham condition, while $\beta(Reward)$ and $\beta(Inequity)$ were comparable in the two conditions ($p = 0.25$ and 0.40, respectively). Furthermore, we also found that participants who showed larger increases in $\beta(Guilt)$ in the tDCS condition tended to exhibit a higher cooperation ratio as well ($r = 0.51$, $p = 0.019$; Fig. 3*C*; linear regression with Grubbs' test). These results demonstrate that guilt aversion and inequity aversion have dissociable neural substrates, and that computational model-based tDCS can selectively increase guilt-aversion-based cooperation.

Discussion

The present study demonstrated dissociable neural substrates for outcome-based (distributional preferences) and intention-based (belief-dependent) economic decisions by integrating computational modeling, fMRI, and tDCS in a newly developed trust game. We found that the right DLPFC is involved in the implementation of intention-based economic cooperation, and the amygdala and ventral striatum are related to the implementation of outcome-based cooperation. Furthermore, tDCS stimulation to the right DLPFC selectively enhanced intention-based economic decisions but did not affect the outcome-based decisions. These findings show that the two systems have a causally dissociable neural basis, consolidating the hypothesis that dual (distributional preferences and belief-dependent) neural processes are involved in economic decision making. Although guilt aversion is dependent on the other's belief, we did not observe brain activation involved with the theory of mind in the current results. This may have happened because we explicitly provided participants with player A's belief probability and participants did not have to estimate it.

For outcome-based economic decision, previous imaging studies have reported a key role of the amygdala and ventral striatum in the context of resource allocation and inequity aversion. The ventral striatum is not only positively correlated with the ratio of the two-player (self and other's) payoff (Fliessbach et al., 2007), but is also activated when inequity between the self and others was reduced (Tricomi et al., 2010). Furthermore, by integrating pharmacological and functional neuroimaging, it was found that participants under serotonin depletion rejected a significantly higher proportion of unfair offers in the ultimatum game and such depletion reduced the response to fairness in the ventral striatum and increased the response during rejection in the dorsal striatum (Crockett et al., 2013). The amygdala response to inequity was reported to be correlated with how strongly each participant dislikes the inequity (Haruno and Frith, 2010; Haruno et al., 2014). Furthermore, the administration of benzodiazepines, enhancers of the GABA A receptor, reduced the rejection rate of unfair offers in the ultimatum game (Gospic et al., 2011) and the activity in the amygdala was simultaneously attenuated. These studies are consistent with the current study and indicate the essential contribution of these brain structures to inequity-based fairness consideration, although the previous studies did not take intention-based economic decisions into account in their task design.

Previous imaging studies have shown that the DLPFC is an important component of working memory (Miller and Cohen, 2001). In our task, the calculation of guilt requires working memory as participants have to maintain and compare several alternatives. However, it is also important to note that the activity in the right DLPFC (Fig. 3*A*) was parametrically correlated with $\beta(Guilt)$. This indicates that the function of the DLPFC is not limited to working memory alone but rather is involved in the implementation of guilt aversion, where working memory is a key building block.

The importance of the DLPFC in economic games has also been implicated in executive function (top-down regulation), including self-control, rule-guided response selection, and fairness preference (Buckholtz and Marois, 2012). For instance, the DLPFC is reported to be involved in behavior constrained by fairness norm (Sanfey et al., 2003; Spitzer et al., 2007; Ruff et al., 2013) and self-controlled behavior (Hare et al., 2009). In parallel with these, stimulation studies that disrupt the right DLPFC function with repetitive transcranial magnetic stimulation (rTMS) and cathodal tDCS in the ultimatum game have shown to decrease the rejection rates of unfair offers (Knoch et al., 2006; Knoch and Fehr, 2007; Knoch et al., 2008). These results could be interpreted as evidence of the DLPFC directly suppressing neural activity that represents a self-interested impulse. However, it is also reasonable to assume that the DLPFC is a part of a network that modulates the relative impact of prosocial motives and self-interest goals. In the current tDCS experiment, $\beta(Reward)$, which represents self-interest, was not significantly different between the tDCS and sham conditions ($p = 0.25$). This result favors the latter view, and we would propose that the right DLPFC is involved not only in the computation of the guilt, but also in the modulation of the relative weights for guilt (note that β value of the right DLPFC in correlation with Guilt was correlated with $\beta(Guilt)$, as discussed in the next paragraph) rather than the top-down regulation of self-interest.

Several studies of value computation during decision making have suggested a single value system (i.e., a common currency

system) that integrates information about all stimulus attributes and then determines choice (Kable and Glimcher, 2007; Hare et al., 2009). Information from both cortical and subcortical structures is assumed to be combined into a single common value representation in ventromedial prefrontal cortex (VMPFC; Levy and Glimcher, 2012). In this view, behavioral decisions depend on how VMPFC value computation weights different stimulus attributes. $\beta(Guilt)$ and $\beta(Inequity)$ in our tDCS experiment showed significant positive correlation with the β values of the right DLPFC in correlation with guilt, and striatum and amygdala in correlation with inequity, respectively. It is therefore possible that the DLPFC, modulates the relative weight of the intention-based economic decision (guilt-aversion), while the striatum and amygdala modulates the relative weight of the outcome-based decision (inequity aversion). This view seems to be consistent with the observation that rTMS to the right DLPFC did not decrease rejection rates when the offers were made by a computer instead of a human (Knoch et al., 2006), since a computer does not have intention. Thus, our data suggest that the right DLPFC besides being involved in cooperative decision making, as demonstrated in previous studies (Spitzer et al., 2007; Ruff et al., 2013), is involved in the implementation of intention-based economic decision and in modulating the relative weight of intention-based economic behavior.

It is intriguing to put the neural mechanism for intention-based economic decisions in broader social contexts. The DLPFC has often been implicated in social-norm adhesion (Spitzer et al., 2007; Buckholtz and Marois, 2012; Ruff et al., 2013), particularly when sanctions are imposed. The similar involvement of the DLPFC in intention-based economic decisions and norm adhesion may arise as the amount of sanction is regarded as the difference between the belief of the society's belief and the actual behavioral outcome. Previous imaging (Greene et al., 2004; Moll and Schulkin, 2009) and TMS (Jeurissen et al., 2014) studies have shown that the DLPFC plays an important role in moral judgment. Moral judgment also concerns how the society perceives the difference between its expectation and the actual outcome. Thus, the integration of computational-model-based analysis of intention-based economic decision making and stimulation techniques may provide a powerful quantitative tool for looking closer into a broader range of social behavior by manipulating a specific computational process in the brain.

Finally, as guilt aversion is related to one's beliefs about the beliefs of others, it may have coevolved with the development of the human symbolic system or language (Charness and Dufwenberg, 2006) and is potentially linked with a broad range of human social cognitive functions. Similarly, the DLPFC may also underlie reputation management (Tennie et al., 2010). On the basis of these facts, we assume that intention-based cooperation requires higher brain function, such as the DLPFC. However, as mentioned above, it is widely recognized that people exhibit strong inequity aversion. This function develops early, between the ages of 3 and 7 years old (Fehr et al., 2008), and the relative importance of intentions versus outcome increases with age (Sutter, 2007). Moreover, inequality aversion (in particular, protest against disadvantage inequity) is widespread in species that cooperate outside kinship and mating bonds, including the capuchin monkey and chimpanzee (S.W. Chang et al., 2011; Brosnan and Waal, 2014). From these observations, we assume that subcortical structures are mainly responsible for modulating the outcome-based cooperation. Such evolutionary and developmental differences may be the basis of the causally dissociable neural substrates we describe here.

References

Bolton GE, Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. Am Econ Rev 90:166–193. CrossRef

Brosnan SF, de Waal FB (2014) Evolution of responses to (un)fairness. Science 346:314. CrossRef Medline

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

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

Chang SW, Winecoff AA, Platt ML (2011) Vicarious reinforcement in rhesus macaques (Macaca mulatta). Font Neurosci 5:27. CrossRef Medline

Charness G, Dufwenberg M (2006) Promises and partnership. Econometrica 74:1579–1601. CrossRef

Charness G, Rabin M (2002) Understanding social preferences with simple tests. Q J Econ 117:817–869. CrossRef

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

Dufwenberg M, Gneezy U (2000) Measuring beliefs in an experimental lost wallet game. Games Econ Behav 30:163–182. CrossRef

Falk A, Fehr E, Fischbacher U (2003) On the nature of fair behavior. Econ Inq 41:20–26. CrossRef

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

Fehr E, Schmidt KM (2006) The economics of fairness, reciprocity and altruism? Experimental evidence and new theories. In: Handbook of the economics of giving, altruism and reciprocity (Kolm S-C, Ythier JM, ed), pp 615–691. Amsterdam: Elsevier.

Fehr E, Bernhard H, Rockenbach B (2008) Egalitarianism in young children. Nature 454:1079–1083. CrossRef Medline

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

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS (1995) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Gandiga PC, Hummel FC, Cohen LG (2006) Transcranial DC stimulation (tDCS): a tool for double-blind sham-controlled clinical studies in brain stimulation. Clin Neurophysiol 117:845–850. 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

Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron 44:389–400. CrossRef Medline

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

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

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. CrossRef Medline

Hensher DA, Rose JM, Greene WH (2005) Applied choice analysis: a primer. Cambridge, United Kingdom: Cambridge UP.

Hsu M, Anen C, Quartz SR (2008) The right and the good: distributive justice and neural encoding of equity and efficiency. Science 320:1092–1095. CrossRef Medline

Jeurissen D, Sack AT, Roebroeck A, Russ BE, Pascual-Leone A (2014) TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. Front Neurosci 8:18. CrossRef Medline

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625–1633. CrossRef Medline

Knoch D, Fehr E (2007) Resisting the power of temptations. Ann NY Acad Sci 1104:123–134. 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

Knoch D, Nitsche MA, Fischbacher U, Eisenegger C, Pascual-Leone A, Fehr E (2008) Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. Cereb Cortex 18:1987–1990. CrossRef Medline

Kosmidis I (2013) brglm: bias reduction in binomial-response generalized linear models. http://www.ucl.ac.uk/~ucakiko/software.html.

Levy DJ, Glimcher PW (2012) The root of all value: a neural common currency for choice. Curr Opin Neurobiol 22:1027–1038. CrossRef Medline

Luce RD (1959) Individual choice behavior: a theoretical analysis. New York: Wiley.

McCabe KA, Rigdon ML, Smith VL (2003) Positive reciprocity and intentions in trust games. J Econ Behav Organ 52:267–275. CrossRef

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202. CrossRef Medline

Moll J, Schulkin J (2009) Social attachment and aversion in human moral cognition. Neurosci Biobehav Rev 33:456–465. CrossRef Medline

R Development Core Team (2008) R: a language and environment for statistical computing (R Foundation for Statistical Computing, Vienna).

Ruff CC, Ugazio G, Fehr E (2013) Changing social norm compliance with noninvasive brain stimulation. Science 342:482–484. 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

Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185–196. CrossRef Medline

Sutter M (2007) Outcomes versus intentions: on the nature of fair behavior and its development with age. J Econ Psychol 28:69–78. CrossRef

Tennie C, Frith U, Frith CD (2010) Reputation management in the age of the world-wide web. Trends Cogn Sci 14:482–488. CrossRef Medline

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

Weber M (1919) Politics as a vocation. In: Max Weber: essays in sociology (Gerth HH, Wright MC, eds). New York: Oxford UP.