Social Cognitive and Affective Neuroscience, 2020, 135–149

doi: 10.1093/scan/nsaa029
Advance Access Publication Date: 12 March 2020
Original article

Neurocomputational mechanisms underlying immoral decisions benefiting self or others

Chen Qu,[1] Yang Hu,[2,3] Zixuan Tang,[2,4] Edmund Derrington,[2,4] and Jean-Claude Dreher[2,4]

[1]School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China, [2]Neuroeconomics Laboratory, Institut des Sciences Cognitives Marc Jeannerod, CNRS, Bron 69675, France, [3]School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China, and [4]Université Claude Bernard Lyon 1, Lyon 69100, France

Correspondence should be addressed to Chen Qu, School of Psychology, South China Normal University, 510631 Guangzhou, China.
Email: fondest@163.com; Yang Hu, School of Psychological and Cognitive Sciences, Peking University, 100871 Beijing, China.
Email: huyang200606@gmail.com; Jean-Claude Dreher, Institut des Sciences Cognitives Marc Jeannerod, Bron 69100, France.
E-mail: dreher@isc.cnrs.fr.
Chen Qu and Yang Hu contributed equally to this study

Abstract

Immoral behavior often consists of weighing transgression of a moral norm against maximizing personal profits. One important question is to understand why immoral behaviors vary based on who receives specific benefits and what are the neurocomputational mechanisms underlying such moral flexibility. Here, we used model-based functional magnetic resonance imaging to investigate how immoral behaviors change when benefiting oneself or someone else. Participants were presented with offers requiring a tradeoff between a moral cost (i.e. profiting a morally bad cause) and a benefit for either oneself or a charity. Participants were more willing to obtain ill-gotten profits for themselves than for a charity, driven by a devaluation of the moral cost when deciding for their own interests. The subjective value of an immoral offer, computed as a linear summation of the weighed monetary gain and moral cost, recruited the ventromedial prefrontal cortex (PFC) regardless of beneficiaries. Moreover, paralleling the behavioral findings, this region enhanced its functional coupling with mentalizing-related regions while deciding whether to gain morally tainted profits for oneself *vs* charity. Finally, individual differences in moral preference differentially modulated choice-specific signals in the dorsolateral PFC according to who benefited from the decisions. These findings provide insights for understanding the neurobiological basis of moral flexibility.

Key words: immoral choice; beneficiary; model-based fMRI; moral flexibility

Introduction

In almost all cultures and societies, human beings tend to transgress established moral values to obtain material advantages in favor of oneself (Bazerman and Gino, 2012; Gächter and Schulz, 2016; Cohn *et al.,* 2019). This immoral behavior often consists of weighing motives to uphold a moral norm (e.g. honesty, fairness) against the maximization of personal profits. However, our moral standards change in different contexts. There are numerous examples that demonstrate the remarkablemalleability of individuals' immorality. It is intriguing that the decision to engage in immoral actions varies depending on whether the action benefits oneself or someone else. For instance, people lie more readily when the lie benefits a charity

Received: 6 November 2019; **Revised:** 21 January 2020; **Accepted:** 28 February 2020

than when it benefits themselves (Lewis *et al.*, 2012). In contrast, the magnitude of dishonesty has been observed to increase over time when it benefits oneself but not when it harms oneself while benefitting others (Garrett *et al.*, 2016). People also tend to judge others' moral transgressions (e.g. unfairness) more harshly than their own (Valdesolo and DeSteno, 2007, 2008). Although recent model-based neuroimaging studies have greatly improved our understanding of the neural substrates of (im)morality *per se* (Hutcherson *et al.*, 2015; Crockett *et al.*, 2017), the neurocomputational mechanisms that guide flexible immoral decision making depending on the beneficiaries of immoral actions remains poorly understood.

Why do people vary their immoral behaviors depending on who receives the benefits, even when, as perpetrators, they receive no punishment for their behavior? According to the self-concept maintenance theory (Mazar *et al.*, 2008), people are often torn between two competing motivations: gaining from immoral actions *vs* maintaining their positive self-concept as a moral individual (Aronson, 1969; Baumeister, 1998; Mazar *et al.*, 2008). Thus, individuals may perform immoral actions to benefit themselves financially at the expense of moral self-concept, or, they may forgo financial benefits to maintain their moral self-concept. In order to resolve this moral dilemma, the theory proposes that people often incorporate a level of immorality into their behavior that can be described as 'just enough.' This strategy allows for a balance between maintaining a relatively intact self-concept (i.e. I am a morally good person) and pursuing personal profit (Mazar *et al.*, 2008; Shalvi *et al.*, 2011). This behavioral theory provides a useful framework for investigating flexible immorality. It explains the reason that people vary their moral standards, i.e. the relative weight of financial gain and moral cost differs depending on who benefits from the immoral actions. However, a computational account of this theory that specifies how people vary the trade-offs between monetary gains and moral costs, according to the characteristics of the beneficiaries, has yet to be defined. Here, we developed and compared computational models of moral decisions incorporating the beneficiary (self/other) of an immoral action, elucidating which variables are computed, how they interact and how they are implemented in the brain during immoral decision-making.

At the brain-system level, a substantial body of literature from social neuroscience and value-based decision-making established a consensus that the ventromedial prefrontal cortex (vmPFC) plays a key role in value computation for different types of goods (Padoa-Schioppa, 2011; Sescousse *et al.*, 2013). Exactly how this region is involved in value computation concerning decisions in social contexts is still debated. While the vmPFC may construct subjective values (SV) during decision-making across domains, irrespective of contexts (Levy and Glimcher, 2012; Bartra *et al.*, 2013; Ruff and Fehr, 2014), other evidence has revealed its unique involvement when people decide for themselves rather than their partners in a delay-discounting task (Nicolle *et al.*, 2012). It is therefore important to directly test whether the vmPFC is engaged in the same way independently of whether it is self or another that benefits from one's immoral action. In addition to the mass-univariate approach to identify common (or different) neural correlates of the value computation in (or between) the two conditions, we also adopted a representational similarity analysis approach (RSA; Kriegeskorte *et al.*, 2008) to assess whether the neural patterns of vmPFC during value computation are similar in the two contexts. The RSA takes advantage of information of multiple voxels to describe the neural pattern similarity between conditions (Kriegeskorte *et al.*, 2008; Kriegeskorte and Kievit, 2013), and this approach has been recently applied to the field of social and decision neuroscience (van Baar *et al.*, 2019).

Here, we developed a novel paradigm in which participants were asked to make a series of decisions involving trade-offs (i.e. offers) between two parties in the MRI scanner. One party had been established to be a morally bad cause in the opinions of the participants, namely an organization severely violating the moral values of caring for the safety and life of others (a gun-holding/hunting advocacy group; see SI: association selection for details). The other party (i.e. the beneficiary) was either the participant (self) or a charity considered to be morally positive. Crucially, by accepting offers in both types of dilemma, either the participant or the charity would be better off; however, it was always accompanied by the moral cost of also profiting the bad cause. When offers were rejected, neither party earned any benefit offered (Figure 1). Notably, to capture how individuals weigh the financial gain and moral cost depending on beneficiaries, we independently varied the monetary payoff for each party (i.e. self/charity *vs* bad cause) in a parametric manner.

This novel experimental design allows us to go beyond traditional behavioral analyses by proposing a series of computational models to elucidate how the human brain computes a decision value integrating moral values and monetary payoffs, and to provide a mechanistic account of flexible immoral choices. We tested and compared a number of such models assuming that immoral choices are made by computing an overall SV as a weighted combination of monetary gains for oneself or the charity and the moral cost of benefiting the morally bad cause. This type of value calculation captures a wide range of behavioral patterns in moral choice (Crockett *et al.*, 2014; Zhu *et al.*, 2014; Volz *et al.*, 2017). By comparing the weights related to different beneficiaries, we were able to characterize the computational processes underlying flexible immoral choices.

Moreover, it is of key importance to establish how the vmPFC, likely to compute SV of an immoral action, interacts with other brain regions during immoral decisions and whether such functional connectivity changes dependent on the beneficiary of the immoral decision. One of the candidate regions is the temporoparietal junction (TPJ), a crucial region proposed to represent other's mental states (Schaafsma *et al.*, 2014; Schurz *et al.*, 2014) and guide other-regarding behaviors (Hampton *et al.*, 2008; Hill *et al.*, 2017). In the moral domain, it has been well established that the TPJ, especially the right part, involves the implementation of the computation concerning how individuals weigh personal gains over other's profits during the altruistic decision-making process (Morishima *et al.*, 2012; Hutcherson *et al.*, 2015; Park *et al.*, 2017). Intriguingly, recent evidence has further clarified a flexible role of the right TPJ in resolving the conflict between personal gains and moral costs depending on the moral contexts (i.e. losing money to benefit a charity or gaining money to benefit a gun rights advocacy group; Obeso *et al.*, 2018). Regarding its link to vmPFC, previous neuroimaging studies have shown increased functional connectivity between the vmPFC and the TPJ in a charity-donation task (Hare *et al.*, 2010) and in a self-other money-split task (Strombach *et al.*, 2015), which consisted of a trade-off between self-profit and benefiting others. We thus investigated whether the functional connectivity between the vmPFC and TPJ is increased when the decider's own interest is not involved at all in an immoral context.

Additionally, this model-based functional magnetic resonance imaging (fMRI) approach enabled us to understand the links between inter-subject variability regarding flexible immorality and brain activity depending on the beneficiaries

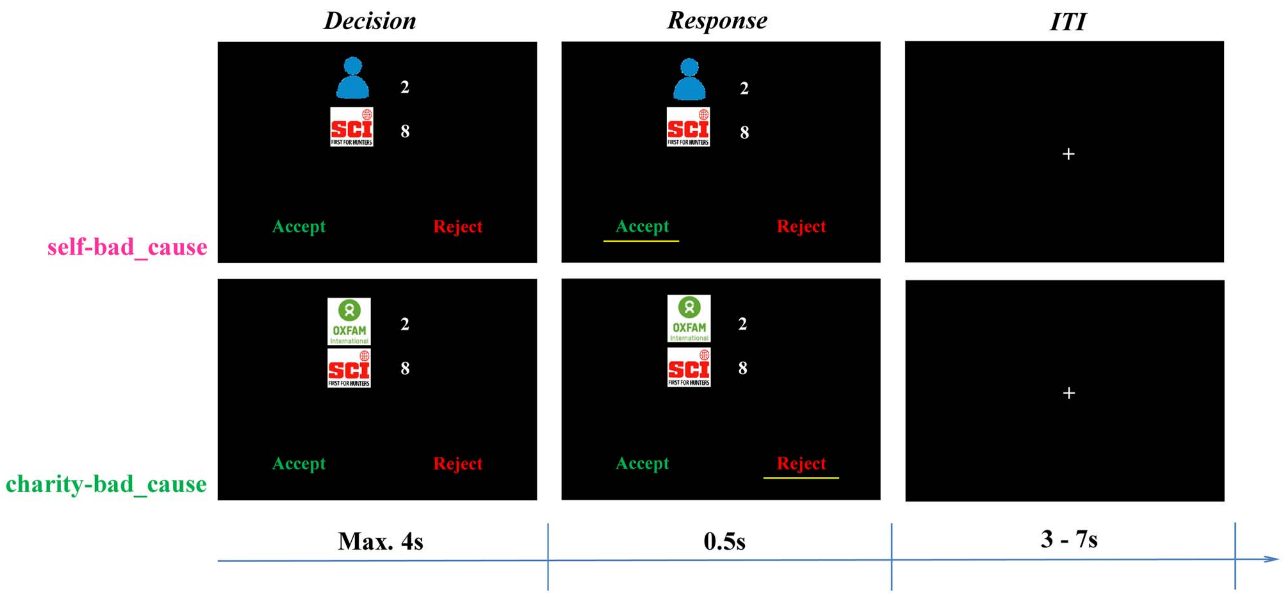

Fig. 1. Experimental design. In each trial, participants were first presented with an immoral offer, in which monetary payoffs for two parties were orthogonally varied. One of the parties was always a morally bad cause (i.e. Safari Club International, SCI), whereas the other party was either the participant himself (i.e. a self-bad cause dilemma) or the preferred charity (i.e. Oxford Committee for Famine Relief, OXFAM; a charity-bad cause dilemma). Participants needed to decide whether to accept or reject the offer within 4 s. If they accepted the offer, both themselves/the charity and the morally bad cause would earn the money as proposed. Otherwise neither party would profit. Each trial was ended with an ITI showing a jittered fixation (3–7 s).

of the morally bad action. More precisely, we aimed to use the model parameters to characterize moral preference across participants (i.e. some subjects are more altruistic and others more selfish), and then to investigate how such inter-individual difference influences the brain activity when accepting that oneself or a charity would benefit (monetarily) from profiting a bad cause (moral cost). It is clear from previous studies that immoral behavior varies from person to person: some subjects never cheat or always cheat but most only cheat sometimes (Fischbacher and Föllmi-Heusi, 2013; Rosenbaum *et al.*, 2014). At the brain level, a causal role of the lateral prefrontal cortex (lPFC) has been reported to be the representation of moral goal pursuit (Carlson and Crockett, 2018). Enhancing activity of the right dorsal lPFC (dlPFC) via anodal transcranial direct current stimulation significantly reduced the probability of cheating, providing a causal role of the lPFC in gating immoral behaviors (Maréchal *et al.*, 2017). Furthermore, neural signals in lPFC often predict inter-individual difference in immoral behaviors such as self-serving lying (Dogan *et al.*, 2016; Yin and Weber, 2018). In a different setting, Crockett *et al.* (2017) revealed an association between lPFC response to immoral earning and inter-individual differences in other-oriented harm aversion. In the light of this literature, we hypothesized that the correlation between inter-individual differences in moral preference and dlPFC activity observed when accepting moral dilemmas would depend upon the beneficiary of the immoral choice.

Methods

Participants

Forty undergraduate or graduate students (25 females; mean age: 20.0 ± 2.0 years, ranging from 18 to 27 years; two left handed) were recruited via online fliers for the fMRI experiment. All participants had normal or corrected-to-normal vision and reported no prior history of psychiatric or neurological disorders. The study took place at the Imaging Center of South China Normal University and was approved by the local ethics committee. All experimental protocols and procedures were conducted in accordance with the IRB guidelines for experimental testing and were in compliance with the latest revision of the Declaration of Helsinki.

Stimuli

Four charities (i.e. Oxford Committee for Famine Relief, Save the Children, First Aid Africa and Oceania) and four non-profit associations advocating gun rights/hunting (i.e. American Rifle Association, The Society for Liberal Weapons Rights, Safari Club International, The European Federation of Associations for Hunting and Conservation of the EU) were selected as the charities and morally bad causes for the current fMRI study, respectively, based on the ratings by an independent group of participants ($N = 30$; see Supplementary data: Association selection for details).

The payoff matrix used in the current fMRI study consisted of 64 different combinations between the monetary gain for participants themselves or their preferred charity (i.e. 1–8 in steps of 1 monetary unit; 1 MU = CNY 9, same below) and the moral cost (i.e. payoffs for the pre-selected morally bad cause: 4–32 in step of 4 MU). The ratio between the monetary gain and the moral cost was deliberately set to 1 : 4 given previous studies in our lab (Obeso *et al.*, 2018) as well as the results of the pilot study (see Supplementary data: Pilot behavioral study for details).

Task

Before scanning, participants were asked to choose one charity and one morally bad association, respectively among the four candidates mentioned above. Similarly, they read the vignette (with logo) first and then indicated the degree of familiarity (0 = 'not at all', 10 = 'very much') as well as liking (−10 = 'not at all' or 'very negative', 0 = 'no preference', 10 = 'very much' or 'very positive') towards each association on a Likert rating scale. To rule out the possibility of equal preference, we also explicitly asked them to indicate the charity (bad cause) he/she likes (dislikes) the most and feels the most (least) willingness to donate to.

An event-related design was adopted in the present fMRI study, including 128 trials in total with half in each of two conditions (see below). Trials were presented pseudo-randomly by using M-sequence to improve the efficiency of estimation of hemodynamic responses (Buračas and Boynton, 2002). On each trial, participants were presented with an immoral offer, benefiting two parties with different amount of monetary payoffs. One party was always the morally bad cause, the other party was either the participants themselves (i.e. a self-bad cause dilemma), or the preferred charity (i.e. a charity-bad cause dilemma). In either dilemma, participants were faced with two options, i.e., 'accept' or 'reject,' with the position counterbalanced across participants but fixed within the participant. If they chose to accept the offer, both themselves/the charity and the morally bad cause would earn the money as proposed. Otherwise neither party would profit. Participants were asked to respond within 4 s by pressing a corresponding button on the button box with the left/right index finger. If an invalid response was made (i.e. no response in 4 s or response less than 200 ms), a warning screen showed up and this trial was repeated at the end of the scanning session. Each trial ended up with an inter-trial interval (ITI) showing a jittered fixation (3–7 s).

Participants were told that their decisions were independent from trial to trial and that once the present task was chosen to be paid (see Procedure for details), one trial in each dilemma would be randomly selected to determine their final payoff and the corresponding donation made to the preferred charity. The final amount donated to the pre-selected morally bad cause was randomly determined between one of the two selected trials mentioned above. In fact, we only paid participants accordingly and no donations were made to these associations. Participants were informed of this at the very end of the experiment.

Procedure

On the day of scanning, participants signed a written informed consent and were explained the procedure, which included the present task and another independent task which will be reported elsewhere. To rule out the possibility of hedging the income risk across two tasks, they were informed that besides the participation fee (i.e. 80 CNY $\approx$12.7 USD), only one task, randomly chosen by the computer at the end of the experiment, would be paid in addition to their basic fee.

For the current task, participants were first provided with the instructions and then they selected their favorite charity as well as the morally bad cause they disliked the most. Before the fMRI task, participants completed a series of comprehension questions to ensure that they fully understood the task and also performed a practice session to get familiar with the paradigm as well as the response button in the scanner. The scanning part included one functional session lasting around 15 min, which was followed by a 6-min structural scan. After that, participants filled out a battery of questionnaires by indicating degrees on a Likert rating scale, including the degree of moral conflict when they made the decisions (0 = 'not at all', 100 = 'very much') and that of moral inappropriateness if they accepted the offer (0 = 'not at all,' 100 = 'very much'), for each dilemma separately. They also filled out several scales of personality traits used for the exploratory analyses. After completing this, participants were debriefed, paid and acknowledged.

Behavioral analyses

All behavioral analyses were conducted using R (http://www.r-project.org/) and relevant packages (R Core Team, 2014). All the reported P values are two-tailed and $P < 0.05$ was considered to be statistically significant. Data visualization was performed via 'ggplot2' package (Wickham, 2016).

Regarding the choice data, we performed a repeated mixed-effect logistic regression on the decision of choosing the 'accept' option by the glmer function in 'lme4' package (Bates *et al.*, 2013), with dilemma (dummy variable; reference level: self-bad cause dilemma; same below) and payoffs for both parties involved in each dilemma (i.e. the monetary gain and the moral cost; mean-centered continuous variable; same below) as the fixed-effect predictors. In addition, we included intercepts varying across participants as the random effects. For the statistical inference on each predictor, we performed the Type II Wald chi-square test on the model fits by using the Anova function in 'car' package (Fox *et al.*, 2016), and reported the odds ratio as relevant effect size.

For decision time (DT), we first did a log-transformation due to its non-normal distribution (Anderson–Darling normality test: $A = 91.90$, $P < 0.001$) and then performed a mixed-effect linear regression on the log-transformed DT by the lmer function in 'lme4' package, with decision (dummy variable; reference level: accept), dilemma, decision × dilemma, as well as payoffs for both beneficiaries as the fixed-effect predictors. Random-effect factors were specified in the same way as above. Similar analyses were also performed on the post-scanning rating except that dilemma was added as the only fixed-effect predictor. We followed the procedure recommended by Luke (2017) to obtain the statistics for each predictor by applying the Satterthwaite approximations on the restricted maximum likelihood model fit via the 'lmerTest' package (Luke, 2017). In addition, we computed the Cohen's d of each predictor via the 'EMAtools' package (Kleiman, 2017), which provided the effect size measure especially for the mixed-effect regressions. For likeness ratings of the selected associations, we compared whether the ratings significantly differed from 0 in each type of selected associations (i.e. charity or morally bad causes), respectively, by the one-sample T-test, and computed the Cohen's d as effect size.

Computational modeling

To examine how participants integrated the payoffs of both parties in two different dilemmas into a SV, we compared the following 10 models with different utility functions.

Model 1 was adapted from a recent study on moral decision making by Crockett *et al.* (2014, 2017). The model described that the SV was calculated by the gains for participants or the charity relative to that for the bad cause, which could be formally represented as follows:

$$\mathrm{SV}\,(G, C) = \alpha G - (1-\alpha)\,Cs$$

where G represents the monetary gain for the participant or the charity, while C represents the moral cost, measured by the monetary gain to the morally bad cause. α is the unknown parameter of social preference that characterizes the relative weight on the payoff for either party involved in the dilemma ($0 < \alpha < 1$).

Model 3 was based on the study by Park *et al.* (2011), which initially examined the integration of positive and negative values and recently was adapted to a donation task (Lopez-Persem *et al.*, 2017):

$$\mathrm{SV}\,(G, C) = \alpha G + \beta C$$

where α and β are the unknown parameters, which capture the weight of the payoff for either party involved in the dilemma ($-10 < \alpha < 10$, $-10 < \beta < 10$).

Model 5 was based on the Fehr–Schmidt model (Fehr and Schmidt, 1999):

$$\mathrm{SV}\,(G, C) = G - \alpha\ \max\ (C - G, 0) - \beta \max\,(G - C, 0)$$

where α and β measure the degree of aversion to payoff inequality in disadvantageous and advantageous situation, respectively (i.e. how much participants dislike when they themselves/the charity earned less/more than the bad cause; $0 < \alpha < 5$, $0 < \beta < 1$).

In addition, we also included Model 7 assuming that people are aversive to the absolute payoff inequality between two beneficiaries, captured by a parameter θ ($0 < \theta < 5$):

$$\mathrm{SV}\,(G, C) = G - \theta\,|G - C|$$

Models with even index mimic corresponding models with odd index (i.e. Models 2, 4, 6, 8 match with Models 1, 3, 5, 7, respectively) except that those unknown parameters varied dependently on the two dilemmas.

For all models above, when the moral cost (C) and the benefit (G) are different from 0, participants bear these cost/benefit [i.e. SV = SV (G,C)]. However, when G = C = 0, participants do not have to bear the cost nor the benefit, and SV = 0 [i.e. SV = SV (0,0) = 0]. Thus, given the softmax rule, the probability of accepting or rejecting the offer is written as below:

$$p\,(\text{accept}) = \frac{e^{\tau \mathrm{SV}_{\text{accept}}}}{e^{\tau \mathrm{SV}_{\text{accept}}} + e^{\tau \mathrm{SV}_{\text{reject}}}} = \frac{e^{\tau \mathrm{SV}}}{1 + e^{\tau \mathrm{SV}}}$$

$$p\,(\text{reject}) = 1 - p\,(\text{accept}) = \frac{1}{1 + e^{\tau \mathrm{SV}}}$$

where τ refers to the inverse softmax temperature ($0 < \tau < 5$), which denotes the sensitivity of individual's choice to the difference in SV between options of acceptance and rejection.

We used the 'hBayesDM' package (Ahn *et al.,* 2017) to fit all aforementioned candidate models using the hierarchical Bayesian analysis (HBA) approach (Gelman *et al.,* 2014). The 'hBayesDM' package is developed based on the Stan language (Stan Development Team, 2016), which utilizes a Markov Chain Monte Carlo (MCMC) sampling scheme to perform full Bayesian inference and obtain the actual posterior distribution. We adopted HBA rather than maximum likelihood estimation (MLE) because HBA provides more stable and accurate estimates than MLE (Ahn *et al.,* 2011). Following the approach in 'hBayesDM' package, we assumed the individual-level parameters ϕ were drawn from a group-level normal distribution: $\phi \sim \text{Normal}\ (\mu_\phi, \sigma_\phi)$, where μ_ϕ and σ_ϕ refer to the group-level mean and standard deviation, respectively. Weakly informative priors were adopted for both these group-level parameters, i.e. $\mu_\phi \sim \text{Normal}\ (0, 1)$ and $\sigma_\phi \sim \text{half-Cauchy}\ (0, 2)$ (Ahn *et al.,* 2017). In HBA, all group-level parameters and individual-level parameters are simultaneously estimated through the Bayes rule given the behavioral data. We fit each candidate model with four independent MCMC chains using 1000 iterations after 1500 iterations for initial algorithm warmup per chain, resulting in 4000 valid posterior samples. Convergence of the MCMC chains was assessed through the Gelman–Rubin R-hat Statistics (Gelman and Rubin, 1992).

For model comparison, we computed the widely applicable information criterion (WAIC) score per candidate model (Vehtari *et al.,* 2016). WAIC score provides the estimate of out-of-sample predictive accuracy in a fully Bayesian way, which is more reliable compared to the point-estimate information criterion (e.g. AIC). By convention, the lower WAIC score indicates better out-of-sample prediction accuracy of the candidate model. A difference score of 10 on the information criterion scale is considered decisive (Burnham and Anderson, 2004). We selected the model with the lowest WAIC as the winning model for subsequent analysis. In addition, we also implemented a posterior predictive check to further examine the absolute performance of the winning model, i.e. whether the prediction of the winning model could characterize the features of real choices. In specific, we employed each individual's joint posterior MCMC samples (i.e. 4000 times) to generate new choice datasets correspondingly (i.e. 4000 choices per trial per participant), given the actual trial-wise stimuli sequences presented to each participant. Thus, we obtained the model prediction by calculating the average rejection proportion of these new datasets in terms of two dilemmas for each subject, respectively. We tested to what degree the individual model prediction correlated with the actual rejection proportion using Pearson correlation. Based on the winning model and its parameter estimation, we derived the mean of the trial-wise SV for each option and defined the relative SV (rSV) by subtracting the SV of the non-chosen option from that of the chosen option (i.e. rSV = SV_chosen − SV_unchosen). These trial-wise rSVs were used as parametric modulators (PMs) for model-based fMRI analyses (see below for details).

fMRI data acquisition and analyses

The imaging data were acquired on a 3-Tesla Siemens Trio MRI system (Siemens, Erlangen, Germany) with a 32-channel head coil at the Imaging Center of South China Normal University. Functional data were acquired using T2∗-weighted echo-planar imaging (EPI) sequences employing a BOLD contrast (TR = 2000 ms, TE = 30 ms; flip angle = 90°; slice thickness = 3.5 mm, slice gap = 25%, matrix = 64 × 64, FoV = 224 × 224 mm^2) in 32 axial slices. Slices were axially oriented along the AC–PC plane and acquired in an ascending order. A high-resolution structural T_1-weighted image was also collected for every participant using a 3D MRI sequence (TR = 1900 ms, TE = 2.52 ms; flip angle = 9°; slice thickness = 1 mm, matrix = 256 × 256, FoV = 256 × 256 mm^2).

Three participants were excluded from later analyses due to excessive head movements (>3 mm), thus leaving a total of 37 participants whose data were analyzed for the fMRI study (24 females; mean age ± SD = 19.9 ± 2.0 years, ranging from 18 to 27 years; two left handedness). Functional imaging data were analyzed using SPM12 (Wellcome Trust Centre for Neuroimaging, University College London, London, UK). The preprocessing procedure followed the pipeline recommended by SPM12. In particular, functional images (EPI) were first realigned to the first volume to correct motion artifacts, unwarped and corrected for slice timing. Next, the structural T_1 image was segmented into white-matter, gray-matter and cerebrospinal fluid with the skull removed, and co-registered to the mean functional images. Then, all functional images were normalized to the MNI space, resampled with a 2 × 2 × 2 mm^3 resolution, based on parameters generated in the previous step. Last, the normalized functional images were smoothed using an 8-mm isotropic full width half maximum based on Gaussian kernel.

General linear models (GLMs) analyses. For all GLMs below, the canonical hemodynamic response function (HRF) was used and a high-pass temporal filtering was performed with a default cut-off value of 128 s to remove low-frequency drifts.

For each participant, we constructed the following GLMs. GLM1 focused on investigating brain regions encoding the rSV, which integrates the monetary gain and moral cost during decision-making period in each dilemma. Thus, we included the following regressors of interest, namely onsets of the decision period in each dilemma with the duration of actual DT. Each regressor of onset was associated with the rSV based on the winning model as the PMs. For the completeness of analyses, we also established GLM2 to identify regions parametrically encoding monetary gain and moral cost during the decision period in each dilemma. We included the same regressor of onsets as in GLM1, except that each regressor of onsets associated with two PMs, i.e. the monetary gain (self-bad cause dilemma: payoff for the participant; charity-bad cause dilemma: payoff for the charity) and the moral cost (in both dilemmas: payoff for the bad cause). Notably, the default orthogonalization option in SPM12 was switched off to ensure the competition for variance during estimation of two PMs. For these two GLMs, we built up the contrasts of each PM against implicit baseline, and that between two dilemmas for the group-level analyses. GLM3 was established to estimate the choice-specific neural activities in different dilemmas (i.e. the dilemma × decision interactive signals). Four participants were excluded from this analyses due to the missing accept ($N = 2$) or reject ($N = 2$) decisions in the charity-bad cause dilemma. GLM3 was constructed in the same way as GLM2, except that we sorted the onsets of decision period by different decision in each dilemma. We built up the dilemma × decision contrasts (i.e. differential neural activities between accept *vs* reject between two dilemmas) for the group-level analyses.

Regarding the regressors of non-interest, we modeled the onset of button press to rule out the movement effect for all GLMs. Besides, once the participant showed invalid responses, an additional regressor modeling relevant events (i.e. other) was included, which contained decision onsets of invalid trials (i.e. for trials which DTs are less than 200 ms, duration equals the actual DT; for trials of no response, duration equals 4 s) as well as the warning feedback (duration equals 1 s). Furthermore, the six movement parameters were added to all models as covariates to account for artifact of head motion.

RSA. The RSA was carried out in Python 3.6.3 with the NLTools package (v.0.3.6; http://github.com/ljchang/nltools), which aimed to further examine whether the neural patterns in vmPFC during value computation in the self-bad cause dilemma could mimic the one in the charity-bad cause dilemma. For each participant, we established a neural dissimilarity matrix (DM) within the vmPFC (defined based on the conjunction activation in two dilemmas; see Results for details) between value-related contrast maps in the self-bad cause dilemma and in the charity-bad cause dilemma (i.e. the PM contrasts of rSV in GLM 1 in respective dilemmas). The neural DM was calculated by one minus the Pearson correlation between contrast value vectors of vmPFC pattern of two dilemmas. Next, we transformed the individual dissimilarity value back to the correlation coefficient, and then performed the Fisher's z transformation for statistical analyses. Besides one-sample *t*-test, we also did permutation analysis by shuffling individual labels and running the same analysis for 5000 times, and finally calculated the proportion of cases as the significance level of such correlation in which the permuted mean *z*-value exceeded the true mean *z*-value.

Functional connectivity analyses. To address how the functional connectivity between the region encoding value-signals (i.e. vmPFC) and the rest of brain changes between dilemmas, by taking a generalized psycho-physical interaction (gPPI) approach (McLaren *et al.*, 2012). To this end, for each participant, we constructed a PPI–GLM (based on GLM1) using the gPPI toolbox (https://www.nitrc.org/projects/gppi) (i) to extract the de-convolved time series at the group-level peak of the joint activation of vmPFC (i.e. encoding the rSV in both dilemmas; peak MNI: −2/48/−14; see Results for details) within a 6mm radius sphere as the physiological regressor, (ii) to define all regressors (i.e. onsets and PMs) in GLM1 as the psychological regressors and (iii) to multiply the physiological regressor with each psychological regressor as the PPI regressors. These regressors were all convolved with the canonical HRF to model the BOLD signal. In addition, we also incorporated six movement parameters as covariates to account for artifact of head motion. We then built up the individual-level PPI contrasts between two dilemmas and used them for the group-level analyses.

Statistical inference and visualization. Individual-level contrasts mentioned above were fed to the group-level random-effect analyses. One-sample *T*-tests, conjunction analyses (Nichols *et al.*, 2005) and regression analyses were performed to detect the differential neural activities, joint activation and behavioral-brain correlation, respectively. For whole-brain analyses, we adopted $P < 0.05$ at the cluster-level controlling for family wise error (FWE) rate combining with an uncorrected voxel-level threshold of $P < 0.001$ as the analyses of the whole-brain threshold (Eklund *et al.*, 2016). Based on our hypotheses, we also adopted the following regions of interest (ROI) for specific contrasts by performing a small volume correction (SVC), i.e. the vmPFC (2/46/−8) related with value-computation (Bartra *et al.*, 2013), the TPJ (left: −53/−59/20; right: 56/−56/18) related with mentalizing (Schurz *et al.*, 2014) and the dorsolateral PFC (dlPFC: ±46/36/24) related with moral judgment and decision-making (Greene *et al.*, 2001). All these ROIs were defined by 9-mm spheres with corresponding MNI coordinates as the center. Regions were labeled according to the automated anatomical labeling template via the xjView toolbox (http://www.alivelearn.net/xjview8/).

To visualize the effect of PMs on neural activities in relevant ROIs (i.e. vmPFC) over time, we followed the procedure used by (Fleming *et al.*, 2018). In brief, we first extracted the de-noised time courses within the masks mentioned above from 10 s windows time-locked to the onset of decision. Then, we applied a regression with corresponding standardized PMs (i.e. rSV) to the standardized activity of each time point in each dilemma, respectively, resulting in a time course of β weights of PMs. In case of illustrating the effect of dilemma on the modulation of PMs, we ran similar regressions except that we pooled the two conditions together and adopted the dilemma, PMs and their interactions as predictors. We used non-parametric permutation tests (1000 permutations) to assess group-level significance of β weights against 0. Significant effects for individual time points were marked by asterisks if the actual *t*-statistic fell outside the 2.5th or 97.5th percentiles of the null distribution generated by the permutation test.

Results

Behavioral results

Each candidate for charity and morally bad cause was selected by participants at least once (see Supplementary Figure S1a). None of the selected associations was familiar to them, as indicated by the low average scores for familiarity (i.e. less than two on a 0–10 Likert scale; mean ± s.d: charity: 1.68 ± 1.80; morally bad cause: 0.46 ± 1.10). However, participants rated the chosen charities positively [mean ± s.d., (95% confidence interval (CI): 8.57 ± 1.52 (8.06, 9.07); $t(36) = 34.31$, $P < 0.001$, Cohen's $d = 5.64$] and had negative evaluations of the chosen gun/hunting rights advocacy group [i.e. morally bad cause; mean ± s.d. (95% CI): −8.03 ± 2.65 (−8.91, −7.14); $t(36) = -18.42$, $P < 0.0014$, Cohen's $d = 3.03$, see Supplementary Figure S1b).

Although participants stated that it felt less morally inappropriate to accept offers to benefit the charity rather than themselves [48.1 ± 32.5 *vs* 59.1 ± 30.0; b (95% CI) = −11.03 (−21.13, −0.93), SE = 5.09, $t(36) = -2.167$, $P = 0.037$, Cohen's $d = -0.72$] and rated comparable levels of moral conflict during the decision-making period for the self and charity conditions [46.1 ± 32.0 *vs* 48.5 ± 29.6; b (95% CI) = −2.43 (−12.68, 7.81), SE = 5.16, $t(36) = -0.471$, $P = 0.640$, Cohen's $d = -0.16$], their behavior did not tell the same story. Specifically, participants were less likely to accept offers in the charity-bad cause dilemma *vs* self-bad cause dilemma [acceptance rate: 39.7 ± 26.5% *vs* 47.3 ± 35.9%; odds ratio = 0.47, b (95% CI) = −0.75 (−0.93, −0.57), SE = 0.09, $\chi^2(1) = 65.26$, $P < 0.001$]. We also found that higher monetary gain for the participants themselves or the charity [odds ratio = 1.91, b (95% CI) = 0.65 (0.60, 0.70), SE = 0.03, $\chi^2(1) = 640.36$, $P < 0.001$] made participants more likely to accept offers, whereas the moral cost showed the opposite effect [odds ratio = 0.87, b (95% CI) = −0.14 (−0.15, −0.13), SE = 0.01, $\chi^2(1) = 520.91$, $P < 0.001$; see Figure 2].

Concerning the relationship of choice behaviors between the two conditions, we found that participants who accepted more immoral offers in the self-bad cause dilemma also accepted offers more frequently in the charity-bad cause dilemma [r(95% CI) = 0.830 (0.692, 0.910), $t(35) = 8.81$, $P < 0.001$; see Supplementary Figure S2]. Moreover, participants who accepted the previous immoral offer in the self-bad cause dilemma were less likely to reject the current immoral offer in the charity-bad cause dilemma [odds ratio = 0.77, b (95% CI) = −0.26 (−0.45, −0.07), SE = 0.10, $\chi^2(1) = 7.08$, $P = 0.008$], after controlling the effect of the monetary gain [odds ratio = 0.81, b (95% CI) = −0.21 (−0.24, −0.18), SE = 0.02, $\chi^2(1) = 180.69$, $P < 0.001$] and moral cost [odds ratio = 1.05, b (95% CI) = 0.05 (0.04, 0.06), SE = 0.004, $\chi^2(1) = 159.99$, $P < 0.001$] in the current trial.

For DT, we first did a log-transformation due to its non-normal distribution (Anderson–Darling normality test: A = 91.90, $P < 0.001$). Regressions on log-transformed DT revealed a trend-to-significant dilemma × decision interaction [b (95% CI) = −0.04 (−0.07, 0.001), SE = 0.02, $t(4702) = -1.87$, $P = 0.062$, Cohen's $d = -0.05$; see Figure 2]. In addition, higher moral cost accelerated the decision process [b (95% CI) = −0.002 (−0.003, −0.001), SE = 0.0005, $t(4079) = -3.05$, $P = 0.002$, Cohen's $d = -0.09$]. However, participants made decisions more slowly when the earnings for themselves or the charity were large [b (95% CI) = 0.01 (0.008, 0.017), SE = 0.002, $t(4714) = 5.54$, $P < 0.001$, Cohen's $d = 0.16$]. To unpack the marginal significant interaction effect, we ran the same analyses on acceptance and rejection decisions separately. The effect of dilemma on DT in both acceptance choice [self-bad cause dilemma *vs* charity_bad cause dilemma: 1553.6 ± 473.3 *vs* 1558.3 ± 366.5 ms; b (95% CI) = 0.09 (0.06, 0.11), SE = 0.01, $t(2028) = 6.00$, $P < 0.001$, Cohen's $d = 0.27$] and rejection choice [1474.7 ± 304.1 *vs* 1481.1 ± 278.9 ms; b (95% CI) = 0.02 (0.001, 0.05), SE = 0.01, $t(2652) = 2.08$, $P = 0.038$, Cohen's $d = 0.08$] after controlling the effect of payoff [acceptance: monetary gain: b (95% CI) = −0.02 (−0.03, −0.01), SE = 0.003, $t(2027) = -6.23$, $P < 0.001$, Cohen's $d = -0.28$; moral cost: b (95% CI) = 0.005 (0.003, 0.006), SE = 0.0008, $t(2039) = 5.84$, $P < 0.001$, Cohen's $d = 0.26$; rejection: monetary gain: b (95% CI) = 0.037 (0.031, 0.042), SE = 0.003, $t(2659) = 14.06$, $P < 0.001$, Cohen's $d = 0.55$; moral cost: b (95% CI) = −0.005 (−0.006, −0.004), SE = 0.0007, $t(2643) = -7.64$, $P < 0.001$, Cohen's $d = -0.30$].

Computational modeling results

We fitted the computational models noted above to the choice data by adopting the HBA approach (Gelman *et al.*, 2014) via the R package 'hBayesDM' (Ahn *et al.*, 2017). R-hat values of all estimated parameters of all models were close to 1.0 (at most smaller than 1.03 in the current case), indicating adequate convergence of the MCMC chains (Gelman and Rubin, 1992). The hierarchical Bayesian model comparison showed that the Model 4 below was with lowest WAIC scores and outperformed other competitive models (see Supplementary Table S1).

$$SV(G, C) = \alpha G + \beta C$$

$$\alpha = \begin{cases} \alpha_{self\text{-}bad\_cause} \text{ if } self\text{-}bad\_cause \text{ trial} \\ \alpha_{charity\text{-}bad\_cause}, \text{ if } charity\text{-}bad\_cause \text{ trial} \end{cases}$$

$$\beta = \begin{cases} \beta_{self\text{-}bad\_cause}, \text{ if } self\text{-}bad\_cause \text{ trial} \\ \beta_{charity\text{-}bad\_cause}, \text{ if } charity\text{-}bad\_cause \text{ trial} \end{cases}$$

This model was adapted from the study by Park *et al.* (2011), which initially examined the integration of positive and negative values and recently was adapted to a donation task (Lopez-Persem *et al.*, 2017). G and C represent the monetary gain for the participant or the charity and the morally bad cause, respectively. α and β are unknown parameters which capture the weight of the monetary gain and moral cost involved in the dilemma, respectively ($-10 < \alpha < 10$, $-10 < \beta < 10$). On top of it, this model distinguished weights on payoffs of both parties in terms of dilemma. The posterior predictive check further showed that the prediction of the winning model highly correlated the actual choice behavior [self-bad cause dilemma: r(95% CI) = 0.998 (0.996, 0.999), $t(35) = 94.06$, $P < 0.001$; charity-bad cause dilemma: r(95% CI) = 0.996 (0.993, 0.998), $t(35) = 0.996$, $P < 0.001$; see Figure 3a]. Taking a closer look at these individual-level posterior mean of key parameters estimated from the winning model, we found that participants weighted positively the monetary gains either for themselves [$\alpha_{\text{self-bad\_cause}}$: mean ± s.d. (95% CI): 5.41 ± 2.79 (4.48, 6.34); $t(36) = 11.80$, $P < 0.001$, Cohen's $d = 1.94$] or the charity [$\alpha_{\text{charity-bad\_cause}}$: mean ± s.d. (95% CI): 5.22 ± 0.63 (5.01, 5.43); $t(36) = 50.41$, $P < 0.001$, Cohen's $d = 8.29$], whereas they weighted negatively the moral cost in both dilemmas [$\beta_{\text{self-bad\_cause}}$: mean ± s.d. (95% CI): −1.79 ± 1.81 (−2.39, −1.19); $t(36) = -6.03$, $P < 0.001$, Cohen's $d = 0.99$; $\beta_{\text{charity-bad\_cause}}$: mean ± s.d. (95% CI): −2.21 ± 1.47 (−2.70, −1.72); $t(36) = -9.12$, $P < 0.001$, Cohen's $d = 1.50$]. Paired-wise t-test further showed that participants weighted the moral cost more negatively in the charity-bad cause dilemma *vs* self-bad cause (95% CI of mean difference: −0.84, −0.01; $t(36) = -2.07$, $P = 0.046$, Cohen's $d = 0.34$), whereas their weights on the monetary gains were comparable for themselves and the charity (95% CI of mean

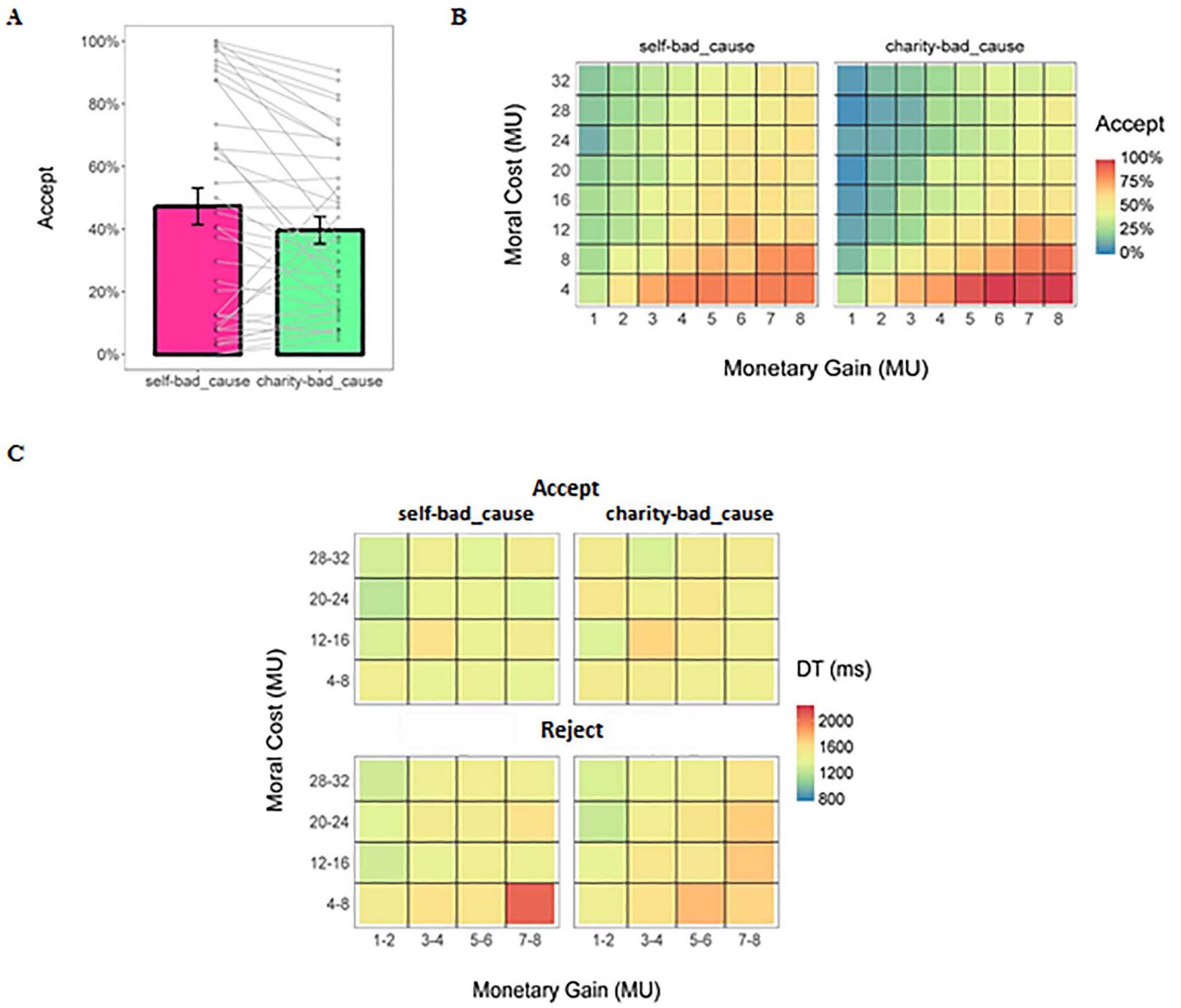

Fig. 2. Behavioral results of the fMRI study. (a) Mean acceptance rate in each dilemma. Each dot refers to the acceptance rate of a single participant. Each line links the acceptance rate of the same participant in two dilemmas. (b) Heat map of the mean acceptance rate (%) as a function of the monetary gain and the moral cost in each dilemma. (c) Heat map of the mean DT as a function of the monetary gain and the moral cost in each dilemma. Data were collapsed into 4-by-4 matrices only for a better visualization. Error bars represent the SEM.

difference: −1.05, 0.68; $t(36) = 0.44$, $P = 0.662$, Cohen's $d = 0.07$; see Figure 3b; also see Supplementary Figure S3 for posterior distribution of individual-level parameters).

To further characterize the inter-individual variance in differential modulatory effect of dilemma on immoral decisions, we computed an index of moral preference in the following way:

$$\text{moralpref} = \left(\alpha_{\text{charity-bad cause}} + \beta_{\text{charity-bad cause}}\right) - \left(\alpha_{\text{self-bad cause}} + \beta_{\text{self-bad cause}}\right)$$

the higher the index is, the stronger the preference of participants to weight monetary gain for the charity higher than for themselves when controlling the weights of the moral cost in the two dilemmas, respectively. Notably, we standardized the original payoffs and re-fitted the winning model (i.e. m4) to the dataset. This made the parameter estimates capturing the weights of both the monetary gains and the moral costs comparable on the same scale. As a *post-hoc* check, we also observed a negative correlation between the moral preference and the total score of Machiavelli scale (Mach-IV; Pearson correlation: r (95% CI) = −0.32 (−0.59, −0.001), $t(35) = -2.03$, $P = 0.049$; see Supplementary Figure S4), where a higher score indicates someone who agrees with the views of achieving one's own purposes or interests by manipulating others even via immoral ways (Christie and Geis, 1970). This finding justified the external validity of the moral preference measured by the present task.

Neuroimaging results

vmPFC encodes rSV during immoral decision-making in both dilemmas (GLM1). The conjunction analyses of both parametric contrasts (against implicit baseline) showed that the activity in vmPFC was positively modulated by rSV generated based on the winning model [peak MNI coordinates: −2/48/−14, $t(72) = 3.08$, p(SVC–FWE) = 0.043; see Figure 4; also see Supplementary

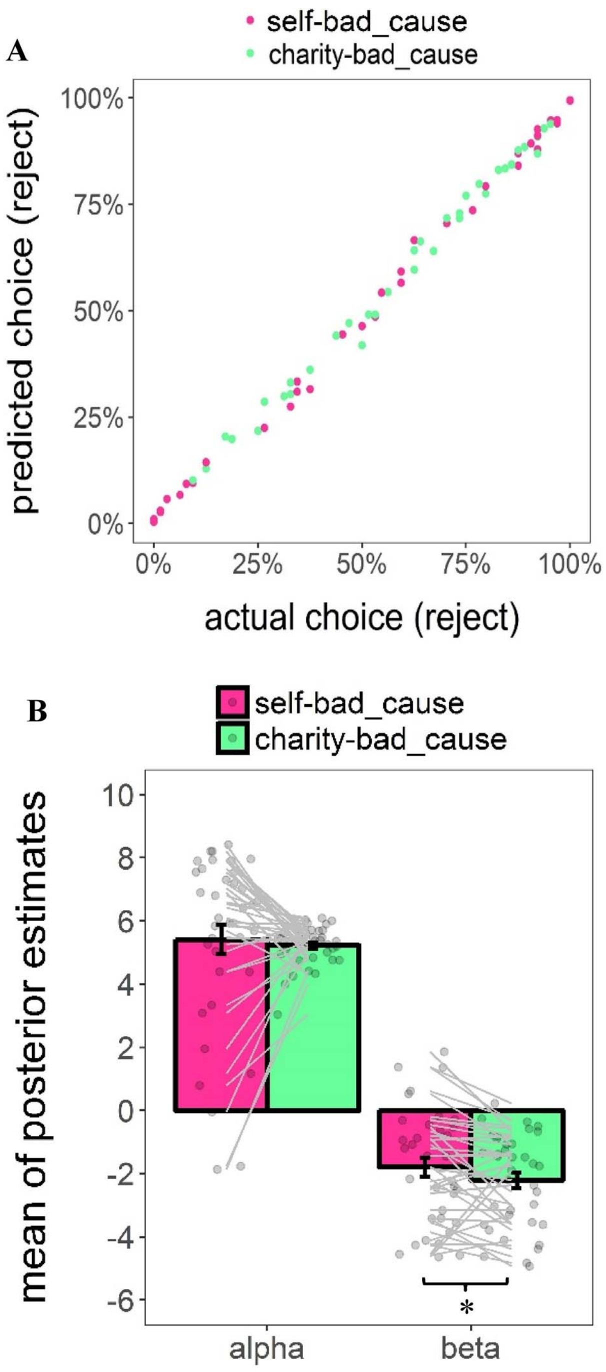

Fig. 3. Results of computational modeling. (a) Scatter plot of the relationship between the actual reject rate (%) and the mean predicted reject rate (%) based on the posterior distribution of the parameters estimated based on the winning model in each dilemma. Each dot represents the data of a single participant. (b) Group-level mean of posterior estimates of individual-level key parameters (i.e. α and β) based on the winning model. Each dot represents the data of a single participant. Error bars represent the SEM; significance: $*P < 0.05$.

Table S2 for details of other activated regions]. This suggests that immoral decisions rely on the same brain circuitry regardless of the beneficiary of the bad action. This was in line with a direct comparison between brain regions modulated by the rSV in the two dilemmas, which did not reveal any significant difference.

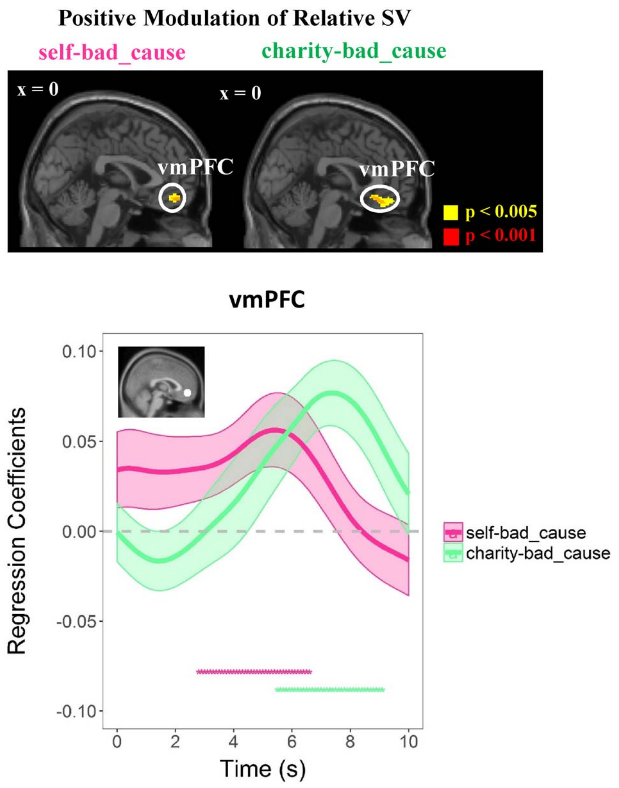

Fig. 4. vmPFC encodes rSV during decision-making period regardless of beneficiaries. Top panel: positive modulation of relative SV on the vmPFC in both dilemmas (GLM2). Bottom panel: regression analysis of the effects of relative SV on an independent ROI (i.e. Bartra *et al.*, 2013: center MNI coordinates: x/y/z = −2/46/−8; a sphere with a radius of 9 mm); activity time courses in each dilemma; the significant effect was indicated by the magenta or green dots. Regression coefficients are represented as group-level mean ± SEM (shaded areas). Dots below the time course indicate significant excursions of t statistics assessed using two-tailed permutation tests. Display threshold: $P < 0.001$ and $P < 0.005$ uncorrected at the voxel-level.

Furthermore, we performed an additional RSA to directly test whether the neural patterns of vmPFC during value computation identified in the self-bad cause dilemma is similar to the that in the charity-bad cause dilemma. Here, we defined the ROI in vmPFC by constructing a 6-mm sphere with the peak coordinate of the conjunction analysis as the center. Consistent with the conjunction analysis, the neural patterns of vmPFC in two conditions were significantly correlated [i.e. for distribution of the neural pattern similarity across participants, see Supplementary Figure S5; r (mean ± s.d.): = 0.151 ± 0.417, Fisher's z (mean ± s.d.) = 0.223 ± 0.570, $t(36) = 2.38$, $P = 0.023$, $p(\text{permutation}) = 0.012$].

Functional connectivity varies across different dilemmas in vmPFC-related network (GLM-PPI). By using the conjunction findings of vmPFC as the seed region (see Supplementary data: fMRI study methods for details), we observed that the functional connectivity between vmPFC and clusters including the dorsomedial PFC (dmPFC) extending to the supplementary motor area (whole-brain level corrected), and the left TPJ [peak MNI coordinates: −56/−62/22, $t(36) = 3.53$, $p(\text{SVC–FWE}) = 0.026$] was significantly higher in the self-bad cause dilemma when making immoral decisions [*vs* charity-bad cause; see Figure 5; also see Supplementary Table S3 for details of PPI results in each dilemma separately].

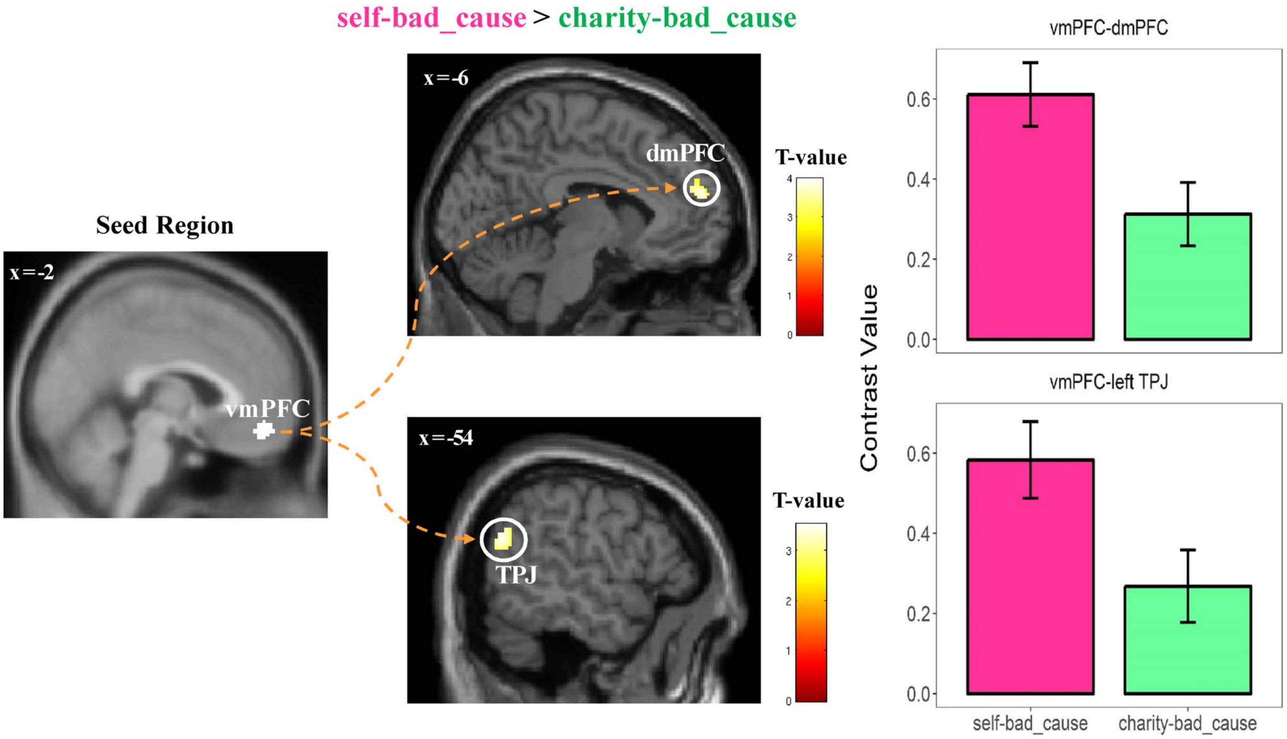

Fig. 5. Reduced connectivity between vmPFC and mentalizing network (i.e. dmPFC; left TPJ) during decision-making period in the charity-bad_cause dilemma (*vs* self-bad_cause). To visualize these connectivity results, we extracted the contrast value of the PPI regressor in both regions (i.e. masks are defined by the intersection between the activated cluster and the independent ROI; dmPFC: center MNI: x/y/z = −1/56/24; left TPJ: center MNI: x/y/z = −53/−59/20; spheres with a radius of 9 mm) in each dilemma. No further statistics were performed to avoid of double dipping. Display threshold: $P < 0.001$ uncorrected at the voxel-level with $k = 100$.

Neural correlates of single attributes (i.e. monetary gains and moral cost; GLM2). For the completeness of the analyses, we also examined the neural correlates of single attributes in each dilemma separately. In the self-bad cause dilemma, we found a positive modulation of personal monetary gains on the neural signals in bilateral ventral striatum and medial prefrontal areas from mid-cingulate cortex to the ventral part of the anterior cingulate cortex (ACC), whereas a negative modulation of moral costs was observed in activity of bilateral inferior parietal lobules (IPL) and left orbitofrontal cortex. In the charity-bad cause dilemma, we observed a positive modulation of money with increasing donations to the charity on the activities in the dorsal part of the ACC, the SMA and the left IPL, whereas a negative modulation of moral costs was observed in the dmPFC and the right IPL (see Supplementary Figure S6 and Supplementary Table S4 for details of all activated regions). No difference on the parametric effect of moral cost was observed between the two dilemmas. The conjunction analysis on the positive modulation of the monetary gains in two dilemmas and that on the negative modulation of the moral costs did not reveal any significant brain region.

Individual differences in moral preference modulate on context-dependent choice-specific decision-relevant neural activation (GLM3). The right dlPFC [i.e. peak MNI coordinates: 40/36/18, $t(31) = 4.52$, $p(\text{SVC–FWE}) = 0.003$] was the only brain region showing a positive correlation between moral preferences and the dilemma × decision interaction (i.e. contrast comparing accept *vs* reject in the interaction between self-bad cause *vs* charity-bad cause dilemma). To better understand this correlation, we extracted contrast values in right dlPFC (defined by an independent mask; see Methods for details) and ran correlation analyses between the right dlPFC signals during acceptance (*vs* reject) choice and moral preference in each dilemma separately, using the lmrob function in the 'rubustbase' package in R (Rousseeuw *et al.*, 2015), which rules out the effect of outlier data points. As a result, the moral preference positively correlated with the right dlPFC activity during acceptance (*vs* reject) in the self-bad cause dilemma [robust correlation: r (95% CI) = 0.31 (0.14, 0.49), $t(31) = 3.64$, $P < 0.001$], whereas a trend-to-significant negative correlation was observed in the charity-bad cause dilemma [r (95% CI) = −0.29 (−0.61, 0.02), $t(31) = -1.93$, $P = 0.063$; see Figure 6].

Discussion

It is well established that people modify their immoral behaviors depending on exactly who will benefit (Lewis *et al.*, 2012; Garrett *et al.*, 2016), but the mechanisms allowing this plasticity in immorality have yet to be described. To address this issue, we designed a novel task in which participants in the scanner were asked to make a series of decisions benefiting either themselves or a charity while simultaneously yielding an immoral consequence in both conditions. Even though participants reported that they felt it was more morally inappropriate to accept an immoral offer to benefit themselves rather than a charity, they nevertheless accepted self-serving immoral offers more frequently and more quickly than those benefitted a preferred charity. These findings are in accordance with previous studies revealing that people exhibit self-serving bias in moral judgment (Bocian and Wojciszke, 2014) and decision-making

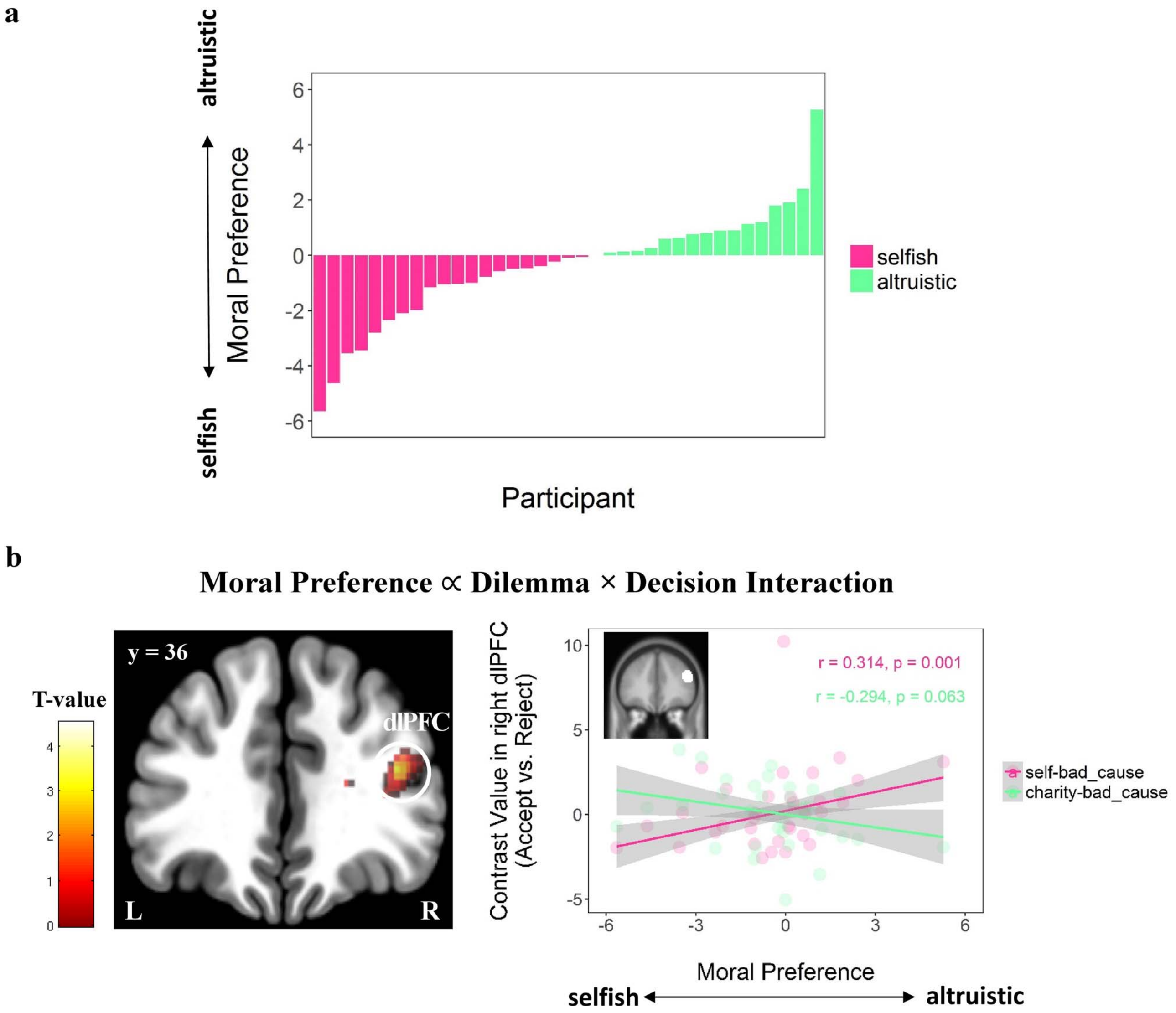

Fig. 6. Individual differences in moral preference modulate on context-dependent choice-specific decision-relevant neural activation. (a) Histogram of the distribution of moral preference across participants. The higher (i.e. more positive) the moral preference is, the more the participant weight on the relative gain for the charity compared with themselves (i.e. more altruistic). (b) Individual difference in moral preference modulates the decision × dilemma interactive effect on the right dlPFC (GLM3; $N = 33$). To unpack the modulation effect, we extracted the contrast value averaged within the right dlPFC (masked by an independent ROI; center MNI: $x/y/z = 46/36/24$; spheres with a radius of 9 mm) from accept *vs* reject contrasts in each dilemma, showing that more altruistic participants tend to elicit higher activity in the right dlPFC when they accept (*vs* reject) the offer in the self-bad_cause dilemma but lower activity when making the same choice in the charity-bad_cause dilemma. Robust correlation was used to rule out the effect of outlier. Lines represent the robust linear fits. Shaded areas represent the 95% CI. Display threshold: $P < 0.001$ uncorrected at the voxel-level with $k = 50$.

(Engel, 2011). Moreover, these results agree with the proposal of moral license, such that displaying altruism via another outlet (i.e. accepting the immoral offers to benefit a charity) would license people to behave more selfishly by accepting more self-serving immoral offers (Merritt *et al.*, 2010). Corroborating this interpretation, we observed a strong inter-individual positive correlation between the acceptance rates of the two types of moral dilemma (involving self or charity as beneficiaries). This compensatory mechanism was also reflected in the fact that participants accepting a self-serving immoral offer in a previous trial were more likely to accept subsequent other-serving immoral offers. These results suggest that individuals may justify their immoral behaviors (i.e. earning morally tainted profits for themselves) by performing alternative altruistic acts (i.e. accepting an immoral offer to benefit a charity), which presumably contributes to a positive self-concept (i.e. the way people view and perceive themselves; Aronson, 1969; Baumeister, 1998).

By adopting a computational approach, we further offer a mechanistic explanation for the behavioral differences underlying these immoral decisions. The winning computational model among those tested assumed the SV of an immoral offer as a linear summation of the weighed monetary gain and moral cost. Such weights varied in different dilemmas: although participants weighed the value of monetary gains for themselves or the charity along similar lines, they differed with respect to how they treated moral cost. That is, participants weighed the moral cost more negatively in dilemmas benefiting a charity (*vs* self), thus explaining why participants accepted offers less frequently in this condition. Notably, our winning computational model also extends a previously reported model accounting for the prosocial choices engaging a tradeoff between personal gains and donations to a charity (Lopez-Persem *et al.*, 2017) to the immoral domain. The present findings also provide a computational account and direct support for the theory of self-concept maintenance (Aronson, 1969; Mazar *et al.*, 2008), which, until this study, has only been tested without computational modeling. This theory proposes that individuals generally value morality (e.g. honesty), and therefore want to maintain this aspect of their self-concept (Greenwald, 1980; Sanitioso *et al.*, 1990). Here, if a person fails to comply with his internal standards for morality by profiting himself with ill-gotten gains, he will need to negatively update his self-concept by decreasing the weight on the moral cost (Mazar *et al.*, 2008).

At the brain-system level, we showed that valuation of immoral offers was associated with rSV computation in the vmPFC, integrating monetary benefits and moral cost regardless of the beneficiaries (self or charity). This finding is consistent with a unified neural representation of value in which the vmPFC is regarded as the key hub (Bartra *et al.*, 2013; Clithero and Rangel, 2014). Furthermore, our RSA analysis directly demonstrated that the neural patterns of vmPFC during value computation identified in the self-bad cause dilemma is similar to the one in the charity-bad cause dilemma. The vmPFC is recruited in the integration of positive (e.g. monetary reward) and negative stimuli (e.g. electrical shock) during value-based decision making processes in non-social contexts (Basten *et al.*, 2010; Park *et al.*, 2011). Previous studies have indicated a similar role for the vmPFC in moral valuation and decision-making (Moll *et al.*, 2005; Greene, 2014). However, these studies did not investigate whether the same valuation regions were engaged for immoral decisions regardless of the beneficiaries. We observed that the functional coupling between the vmPFC and mentalizing regions (i.e. dmPFC together with the TPJ; Schaafsma *et al.*, 2015), was enhanced when making immoral decisions benefiting oneself (as compared to the charity). This differential functional coupling according to the beneficiary of the decision parallels our behavioral findings of a higher acceptance rate for immoral decisions benefiting oneself as compared to a charity. The brain valuation system is known to work together with the mentalizing network during social decision-making (e.g. strategic interactive decisions; Hampton *et al.*, 2008; Hill *et al.*, 2017). In particular, the TPJ is engaged when weighing self-interest with other regarding preferences (Strombach *et al.*, 2015) and it has been shown to be causally necessary to signal the moral conflict between personal profits and moral costs (Obeso *et al.*, 2018). In the present study, the enhanced functional coupling between the vmPFC and mentalizing nodes (i.e. dmPFC and TPJ) in the self-bad cause *vs* charity-bad cause dilemma, reflects that the cross talk between the valuation and the mentalizing network is more sensitive to the signal involving a conflict between personal interests (as compared to the welfare of others) and moral cost. This finding is also interesting in light of recent reports that the mentalizing network supports a key role for social interactions involving oneself as compared to situations in which the participant is an observer rather than an interactive participant (Redcay and Schilbach, 2019).

Furthermore, we investigated the link between choice-specific brain activity and individual differences in moral preference (i.e. an index measuring the extent of behavioral changes in different dilemmas) and how this relationship is influenced by the beneficiary. We found large inter-individual differences in moral preferences at the behavioral level, in line with previous studies (Crockett *et al.*, 2014, 2017; Yin *et al.*, 2017). Investigation of the relationship between this behavioral effect and brain activation revealed, as predicted, that the dlPFC activity was associated with inter-individual differences in moral preference and that this relationship also depended upon the beneficiary of the immoral choice. *Post-hoc* analyses showed that this interaction was a consequence of moral preference increasing with enhanced dlPFC activity in the self-bad cause dilemma when participants accepted (*vs* rejected) the offer, whereas the opposite relationship occurred in the charity-bad cause dilemma. The dlPFC has been shown to impact on a variety of social and moral decisions, such as fairness (Knoch *et al.*, 2006; Spitzer *et al.*, 2007; Ruff *et al.*, 2013), generosity (Hutcherson *et al.*, 2015) and dishonesty (Greene and Paxton, 2009; Maréchal *et al.*, 2017), and has also been linked to individual differences in moral behavior (Crockett *et al.*, 2017). In a recent theoretical framework, the lPFC, including but not limited to the dorsal part, was proposed to act as a flexible guide in the pursuit of moral goals, which depend on the interaction between context and a specific individual (Carlson and Crockett, 2018). In the present study, participants with increasing positive scores in moral preference can be considered as more altruistic while those with decreasing negative scores can be considered to be more selfish. Critically, a negative correlation between this task-based score and the Machiavellianism trait further confirmed the underlying rationale and external validity of our measure of moral preference. Thus, as reflected by stronger dlPFC signals associated with acceptance of self-bad cause dilemmas (*vs* charity-bad cause), altruistic people (i.e. those with higher positive moral preference scores) have to overcome a stronger subjective moral cost when accepting offers that profit themselves. Therefore, our findings provide empirical evidence in support of the hypothesis that dlPFC engagement is key to explaining inter-subject variability in solving conflicts between self-interest and moral cost. In line with the present finding, a recent fMRI investigation found that participants are willing to trade their moral values (i.e. similar to the moral cost in the present study), in exchange for personal profit, and this effect is accompanied by decision value computation engaging the lateral PFC (Qu *et al.*, 2019).

Despite that obtaining personal gains *via* an immoral approach is universally forbidden by moral rules, people still break such moral rules by trading-off self-interest and moral values to maintain a positive self-image. However, the process of moral decision-making can be highly flexible across different contexts, meaning that the balance point between different cost-benefit tradeoffs is not fixed, but context-dependent. Our findings identify the neurocomputational mechanisms and the brain circuitry underlying this flexible immoral behavior to provide a mechanistic understanding of the neurobiological architecture encompassing value computation and integration of contextual information. In particular, an overall value signal is constructed from independent attributes and this integration between single attributes is computed in the vmPFC (computing the decision value) regardless of the beneficiary. In turn, stronger functional connectivity between the vmPFC and components of the mentalizing network (i.e. dmPFC and TPJ) during the dilemma weigh up concerns for the bad cause and oneself as compared to the bad cause and the charity. This represents the neural signature of flexible moral behavior depending upon the beneficiary of the immoral action. Our results also shed light on the question of whether moral decisions rely on the same valuation circuitry engaged during value-based decision making or requires a specialized set of brain regions. From a broader perspective, these findings may have societal implications because the way we value immorality fundamentally affects our behaviors and further shapes the functioning of our societies.

Supplementary data

Supplementary data are available at *SCAN* online.

Acknowledgements

We would like to thank Si-ying Li, Yan-er Su and Fang-jie Xu for their assistance during data collection, Dr Chen Hu, Dr Lei Zhang, Dr Xiao-xue Gao as well as Dr Martin Hebart for their help on data analyses, Dr Chun-liang Feng for comments on an earlier draft.

Funding

This research has benefited from the financial support of IDEX-LYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. J.-C.D. was funded by the grant 'ANR-NSF CRCNS "SOCIAL_POMDP" no. 16-NEUC'. Q.C. was funded by the Program for National Science Foundation of China (31470995).

Conflict of interest: The authors declare that no competing interests exist.

References

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J.R., Brown, J.W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, **4**, 95.

Ahn, W.-Y., Haines, N., Zhang, L. (2017). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, **1**, 24–57.

Aronson, E. (1969). The theory of cognitive dissonance: a current perspective. In: Berkowitz, L., editor. *Advances in Experimental Social Psychology*, New York: Academic Press.

van Baar, J.M., Chang, L.J., Sanfey, A.G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, **10**, 1483.

Bartra, O., McGuire, J.T., Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, **76**, 412–27.

Basten, U., Biele, G., Heekeren, H.R., Fiebach, C.J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, **107**, 21767–72.

Bates, D., Maechler, M. Bolker, B. (2013) *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. 2012*. Available: http://CRAN.R-project.org/package=lme4.

Baumeister, R.F. (1998). The self. In: Gilbert, D.T., Fiske, S.T., Lindzey, G., editors. *Handbook of Social Psychology*, New York: McGraw-Hill.

Bazerman, M.H., Gino, F. (2012). Behavioral ethics: toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science*, **8**, 85–104.

Bocian, K., Wojciszke, B. (2014). Self-interest bias in moral judgments of others' actions. *Personality and Social Psychology Bulletin*, **40**, 898–909.

Buračas, G.T., Boynton, G.M. (2002). Efficient design of event-related fMRI experiments using M-sequences. *NeuroImage*, **16**, 801–13.

Burnham, K.P., Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**, 261–304.

Carlson, R.W., Crockett, M.J. (2018). The lateral prefrontal cortex and moral goal pursuit. *Current Opinion in Psychology*, **24**, 77–82.

Christie, R., Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Clithero, J.A., Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, **9**, 1289–302.

Cohn, A., Maréchal, M.A., Tannenbaum, D., Zünd, C.L. (2019). Civic honesty around the globe. *Science*, **365**, 70–3.

Crockett, M.J., Kurth-Nelson, Z., Siegel, J.Z., Dayan, P., Dolan, R.J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, **111**, 17320–5.

Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Dayan, P., Dolan, R.J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, **20**, 879–85.

Dogan, A., Morishima, Y., Heise, F., *et al.* (2016). Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. *Scientific Reports*, **6**, 33263.

Eklund, A., Nichols, T.E., Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, **113**, 7900–5.

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, **14**, 583–610.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, **114**, 817–68.

Fischbacher, U., Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, **11**, 525–47.

Fleming, S.M., Putten, E.J., Daw, N.D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, **21**, 617–24.

Fox, J., Weisberg, S., Adler, D., *et al.* (2016). *Package 'car'*.

Gächter, S., Schulz, J.F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, **531**, 496.

Garrett, N., Lazzaro, S.C., Ariely, D., Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, **19**, 1727.

Gelman, A., Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–72.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2014). *Bayesian Data Analysis*, FL, USA: Chapman & Hall/CRC Boca Raton.

Greene, J.D. (2014). The cognitive neuroscience of moral judgment and decision-making. In: Gazzaniga, M.S., editor. *The Cognitive Neurosciences V*, Cambridge, MA: MIT Press.

Greene, J.D., Paxton, J.M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, **106**, 12506–11.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, **293**, 2105–8.

Greenwald, A.G. (1980). The totalitarian ego: fabrication and revision of personal history. *American Psychologist*, **35**, 603.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, **105**, 6741–6.

Hare, T.A., Camerer, C.F., Knoepfle, D.T., O'Doherty, J.P., Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience*, **30**, 583–90.

Hill, C.A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J.P., Ruff, C.C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, **20**, 1142.

Hutcherson, C., Bushong, B., Rangel, A. (2015). A Neurocomputational model of altruistic choice and its implications. *Neuron*, **87**, 451–62.

Kleiman, E. (2017). *EMAtools: data management tools for real-time monitoring/ecological momentary assessment data*.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, **314**, 829–32.

Kriegeskorte, N., Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, **17**, 401–12.

Kriegeskorte, N., Mur, M., Bandettini, P.A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, **2**, 4.

Levy, D.J., Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, **22**, 1027–38.

Lewis, A., Bardis, A., Flint, C., *et al.* (2012). Drawing the line somewhere: an experimental study of moral compromise. *Journal of Economic Psychology*, **33**, 718–25.

Lopez-Persem, A., Rigoux, L., Bourgeois-Gironde, S., Daunizeau, J., Pessiglione, M. (2017). Choose, rate or squeeze: comparison of economic value functions elicited by different behavioral tasks. *PLoS Computational Biology*, **13**, e1005848.

Luke, S.G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, **49**, 1494–502.

Maréchal, M.A., Cohn, A., Ugazio, G., Ruff, C.C. (2017). Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, **114**, 4360–4.

Mazar, N., Amir, O., Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *Journal of Marketing Research*, **45**, 633–44.

McLaren, D.G., Ries, M.L., Xu, G., Johnson, S.C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *NeuroImage*, **61**, 1277–86.

Merritt, A.C., Effron, D.A., Monin, B. (2010). Moral self-licensing: when being good frees us to be bad. *Social and Personality Psychology Compass*, **4**, 344–57.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, **6**, 799.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C., Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, **75**, 73–9.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, **25**, 653–60.

Nicolle, A., Klein-Flügge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., Behrens, T.E. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, **75**, 1114–21.

Obeso, I., Moisa, M., Ruff, C.C., Dreher, J.-C. (2018). A causal role for right temporo-parietal junction in signaling moral conflict. *eLife*, **7**, e40671.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual Review of Neuroscience*, **34**, 333–59.

Park, S.Q., Kahnt, T., Rieskamp, J., Heekeren, H.R. (2011). Neurobiology of value integration: when value impacts valuation. *Journal of Neuroscience*, **31**, 9307–14.

Park, S.Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., Tobler, P.N. (2017). A neural link between generosity and happiness. *Nature Communications*, **8**, 15964.

Qu, C., Météreau, E., Butera, L., Villeval, M.C., Dreher, J.-C. (2019). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLoS Biology*, **17**, e3000283.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*.

Redcay, E., Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, **20**, 495–505.

Rosenbaum, S.M., Billinger, S., Stieglitz, N. (2014). Let's be honest: a review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, **45**, 181–96.

Rousseeuw, P., Croux, C., Todorov, V., *et al.* (2015). *Robustbase: Basic Robust Statistics. R Package Version 0.92–3'*.

Ruff, C.C., Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, **15**, 549–62.

Ruff, C.C., Ugazio, G., Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, **342**, 482–4.

Sanitioso, R., Kunda, Z., Fong, G.T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology*, **59**, 229.

Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R. (2014). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, **19**, 65–72.

Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, **19**, 65–72.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, **42**, 9–34.

Sescousse, G., Caldú, X., Segura, B., Dreher, J.-C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, **37**, 681–96.

Shalvi, S., Dana, J., Handgraaf, M.J., De Dreu, C.K. (2011). Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, **115**, 181–90.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, **56**, 185–96.

Stan Development Team (2016). *Stan: A C++ Library for Probability and Sampling*.

Strombach, T., Weber, B., Hangebrauk, Z., *et al.* (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, **112**, 1619–24.

Valdesolo, P., DeSteno, D. (2007). Moral hypocrisy. *Psychological Science*, **18**, 689–90.

Valdesolo, P., DeSteno, D. (2008). The duality of virtue: deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, **44**, 1334–8.

Vehtari, A., Gelman, A., Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–32.

Volz, L.J., Welborn, B.L., Gobel, M.S., Gazzaniga, M.S., Grafton, S.T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, **114**, 7963–8.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Yin, L., Weber, B. (2018). I lie, why don't you: neural mechanisms of individual differences in self-serving lying. *Human Brain Mapping*, **40**, 1–13.

Yin, L., Hu, Y., Dynowski, D., Li, J., Weber, B. (2017). The good lies: altruistic goals modulate processing of deception in the anterior insula. *Human Brain Mapping*, **38**, 3675–90.

Zhu, L., Jenkins, A.C., Set, E., *et al.* (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, **17**, 1319–21.