Moral transgressions corrupt neural representations of value

Molly J. Crockett[1,2], Jenifer Z. Siegel[1], Zeb Kurth-Nelson[3,4], Peter Dayan[5], & Raymond J. Dolan[3,4]

[1]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom
[2]Department of Psychology, Yale University, New Haven, CT 06520, USA
[3]Max Planck–University College London Centre for Computational Psychiatry and Ageing, London WC1B 5EH, United Kingdom
[4]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG
[5]Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom

Correspondence should be addressed to M.J.C. (mj.crockett@yale.edu).

Abstract
Moral systems universally prohibit harming others for personal gain. However, we know little about how such principles guide moral behavior. Using a task that assesses the financial cost participants ascribe to harming others versus themselves, we probed the relationship between moral behavior and neural representations of profit and pain. Most participants displayed moral preferences, placing a higher cost on harming others than themselves. Moral preferences correlated with neural responses to profit, where participants with stronger moral preferences had lower dorsal striatal (DS) responses to profit gained from harming others. Lateral prefrontal cortex (LPFC) encoded profits gained from harming others, but not self, and tracked the blameworthiness of harmful choices. Moral decisions also modulated functional connectivity between LPFC and the profit-sensitive region of DS. The findings suggest moral behavior in our task is linked to a neural devaluation of reward realized by a prefrontal modulation of striatal value representations.

Despite the diversity of human moral values, there is a universal prohibition on harming others for personal gain[1,2]. Humans avoid harming others to a remarkable degree compared with other species[3], and are even willing to incur significant personal costs to alleviate others' suffering[4,5]. Why, and how, people forgo self-interest for the sake of others' welfare remains an enduring puzzle. Recent work has implicated specific brain regions in moral decision making[6–9] and probed how moral behavior relates to social cognitive processes such as empathy and mentalizing[10–14]. However, little is known about the neural computations supporting moral decisions to avoid harming others for personal gain, and whether individual differences in these computations predict variation in actual moral behavior.

We measured moral preferences in a task where participants could trade personal profits against pain experienced by either themselves or an anonymous other person (**Fig. 1a**). Most people required more financial compensation to increase others' pain compared with their own[15,16]. In other words, profiting from another's pain had lower subjective value than profiting from one's own pain. One possible explanation for this moral preference is that another's pain is

more aversive than one's own pain. Alternatively, profits gained from harming another may engender less pleasure than the very same profits gained from harming oneself[17].

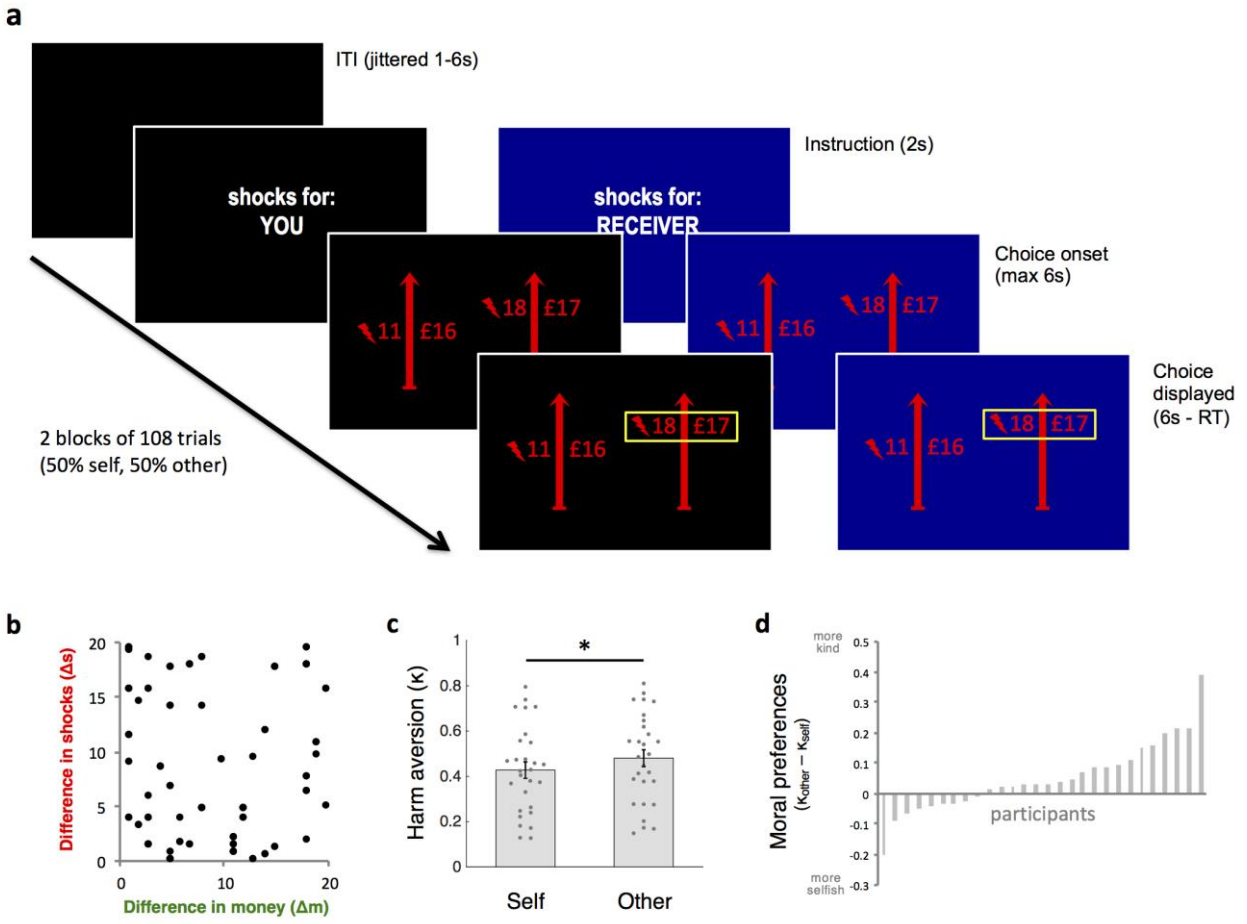
Because the moral behavior we are interested in here reflects a tradeoff between profit and pain, these competing explanations are not easily resolved from behavioral observation alone. However using neuroimaging we can ask whether individual differences in moral preferences are better explained by differential neural representations of pain or profit, in the context of harming others versus oneself. Previous neuroimaging studies of moral decision making have attributed activity in several brain regions to a range of cognitive processes. Activity in insula, anterior cingulate cortex (ACC) and temporoparietal junction (TPJ) is linked to empathy and mentalizing[7,8,10–14]; activity in striatum and ventromedial prefrontal cortex (vmPFC) is linked with value computation[8,12,13]; and lateral prefrontal cortex (LPFC) activity is considered to reflect cognitive control[6–9]. However, decomposing the cognitive mechanisms supporting moral decisions is difficult without resorting to reverse inference. Here we address this challenge by independently manipulating the amounts of profit and pain resulting from participants' decisions (**Fig. 1b**). This in turn allowed us to extract neural representations of profit and pain and ask whether the former was suppressed, or the latter boosted, as a function of the behavioral expression of a moral aversion to harming others for profit. We avoid reverse inference by asking whether, in line with moral behavior, *any* brain region shows a greater response to others' pain compared to one's own pain, or a weaker response to profits gained from harming others relative to profits gained from harming oneself. Our prediction, based on prior literature, was that moral decisions involving a tradeoff between pain and profit would engage regions either implicated in pain processing or value-based decision making respectively, with pain encoded in insula, ACC and TPJ[18,19], and profit encoded in the striatum and vmPFC[20,21].

Higher-order goals represented in regions such as lateral prefrontal cortex (LPFC) are known to modulate value computations in striatum and vmPFC[22]. In particular, LPFC is implicated in orchestrating an influence of moral norms on behavior[23–29]. This region shows increased activation to the extent that people choose to comply with fairness norms[24], reciprocate trust[26] and avoid harming others for personal gain[8]. Disrupting LPFC activity impairs the integration of moral blame assessments into punishment decisions[28]. However, notwithstanding these observations the precise computational role of LPFC in promoting norm compliance remains unanswered. One proposal is that LPFC regulates moral behavior by re-weighting the inputs to policy decisions[30], for example by modulating the subjective value of harmful actions represented in striatum[23,31]. This view is grounded in evidence for major anatomical projections from LPFC to dorsal striatum (DS)[32,33] as well as LPFC modulation of DS value signals during temporal discounting[34]. On the bases of these prior data, we hypothesized that value-sensitive areas of DS would show a reduced response to profits gained from harming others, relative to harming oneself, and that moral decisions would modulate functional connectivity between these same value processing regions and LPFC.

We tested our hypotheses in an fMRI study (N=28) where participants played the role of "decider" who chose whether to profit by inflicting painful electric shocks on either themselves or an anonymous other "receiver" (**Fig. 1a-b**). Crucially, deciders faced identical choice sets when deciding to profit from harm to others vs. self, enabling us to ask how potential moral transgressions modulate neural value computations of profit and pain. To mitigate concerns about reciprocity and reputation, deciders were instructed their choices would be private with respect to the receivers and experimenters, and post-study questionnaires confirmed they believed this. In a

second behavioral study (N=49) involving a similar design we asked participants to provide blame judgments in addition to moral decisions (**Supplementary Fig. 1**). This enabled us to build a model linking blame judgments and moral decisions, which we used to test hypotheses about moral norm representation in LPFC.

Figure 1. Moral decision task and behavioral results. (a) In the fMRI study, participants assigned to the role of "decider" (N=28) chose between a *harmful option* containing more money and shocks, and a *helpful option* containing less money and fewer shocks. On half the trials the shocks were for the decider (left) and on the other half the shocks were for the receiver (right). **(b)** Example trial set where each point represents a trial. Across trials we independently manipulated the difference in pain and difference in profit between the two options, which allowed us to separate neural signals related to pain and profit. **(c)** Harm aversion (κ) was greater for others than self ($t_{(27)}$=-2.40, *P*=0.024). **(d)** Distribution of moral preferences ($\kappa_{other} - \kappa_{self}$) among deciders. Error bars depict s.e.m. *P<0.05.



Results

Computational model of moral decisions

In the moral decision task we modeled deciders' choices by adapting a model we previously validated in four behavioral studies[15,16]. The model again explained the current data well, correctly predicting 87% of deciders' choices (95% confidence interval [85%-88%]; mean pseudo-R^2=0.692) and outperformed a range of alternative models (**Supplementary Modeling Note**). The model described the difference in subjective value between the harmful and helpful options as follows:

$$\Delta V = (1-\kappa)\Delta m - \kappa \Delta s$$

$$\kappa = \begin{cases} \kappa_{self} & \text{if self trial} \\ \kappa_{other} & \text{if other trial} \end{cases}$$

where Δ*V* is the difference in subjective value between the harmful and helpful options, and Δ*m* and Δ*s* are the objective differences in money and shocks between the harmful and helpful options, respectively. Δ*V* is based on a weighted average of these two quantities, where the relative weighting is determined by a harm aversion parameter κ. When κ=0, deciders will accept any number of shocks to gain profit. As κ approaches 1, deciders become maximally harm averse and will sacrifice increasing amounts of profit to avoid an additional shock. The setting of κ depends on who is receiving the shocks, where κ_{self} and κ_{other} capture the subjective cost of harm to self and others, respectively. Trial-by-trial subjective value differences were transformed into choice probabilities using a softmax function[35]. Consistent with previous findings[20,21] BOLD responses at choice onset correlated with model estimates of the subjective value of the chosen relative to the unchosen option (relative chosen value; GLM1) in a network including vmPFC (P_{FWE}<0.0001), mid-posterior cingulate (P_{FWE} <0.0001), precuneus (P_{FWE}<0.0001), and bilateral clusters encompassing amygdala, striatum and insula (P_{FWE}<0.0001; all results whole brain familywise error (FWE) corrected at the cluster level after voxel-wise thresholding at *p*<0.001).; **Supplementary Fig. 2a** and **Supplementary Table 1**). Relative chosen value signal in these regions did not significantly differ between the self and other conditions.

Previous studies using this task showed that most participants displayed moral preferences involving a greater harm aversion for others than for self[15,16]. We replicated this effect again in the current study ($\kappa_{other} > \kappa_{self}$, M=0.053, SD=0.116, $t_{(27)}$=-2.40, p=0.024; **Fig. 1c**). This pattern of moral preferences was observed in 68% of participants (**Fig. 1d**). Analysis of raw choice data indicated that participants were more likely to choose the harmful option for themselves than for the receiver (difference score, M=5%, SD=12%; $t_{(27)}$=2.18, p=0.038). Moral preferences (computed as the difference in harm aversion for self and others, i.e., $\kappa_{other} - \kappa_{self}$) resulted in participants paying, on average, an extra 17p per shock to prevent shocks to others relative to themselves.

Responses in vmPFC reflected these moral preferences. We regressed participant-specific subjective values for the harmful option against BOLD responses at decision time, and extracted the parameter estimates from the value-sensitive vmPFC region identified above. In this region of vmPFC, BOLD responses were less correlated with the value of harming others than the value of harming oneself ($t_{(27)}$=2.51, p=0.019; **Supplementary Fig. 2b**). Nevertheless, as seen previously there was wide individual variation in the degree of expression of moral preferences, which we exploited to probe the neural computations that guide moral decisions.

Neural representation of pain is uncorrelated with moral behavior

Our first aim was to test whether moral behavior is explained by a greater neural sensitivity to anticipated pain for others relative to self, or a lesser neural sensitivity to profit gained from harming others relative to self. We tested these hypotheses in a GLM that identified regions responding parametrically at decision onset, irrespective of participants' choices, to the objective amounts of profit and pain (Δm and Δs) that would result from choosing the harmful option, relative to the helpful option, in the self and other conditions (GLM2).

For the pain analysis, we identified voxels where activity varied parametrically with Δ*s*, irrespective of choice. Region-of-interest (ROI) analyses revealed neural responses to Δs_{self} and Δs_{other} in ACC and TPJ, respectively, but these were not correlated with individual differences in behavior (**Supplementary Fig. 3**). For completeness, we regressed individual differences in moral preferences onto the group-level maps of parametric responses to anticipated pain for others relative to self ($\Delta s_{other} > \Delta s_{self}$). In a whole brain analysis, we found no significant clusters in a whole brain analysis exceeding a significance level of p<0.05 FWE corrected or within any *a priori* ROIs (**Supplementary Table 2**). Thus, we did not find evidence supporting a relationship between individual differences in neural responses to anticipated pain and variation in moral behavior. Although this null association could reflect an absence of robust neural responses to anticipated pain, this is unlikely because there was a robust relationship between individual differences in κ_{self} and neural responses to Δs_{self} in the insula (P_{FWE}=0.011, whole brain FWE corrected at the cluster level after voxel-wise thresholding at p<0.001; **Supplementary Table 3)**.

Finally, we investigated a possible relationship between other-related pain signals in TPJ and ACC and choice-related value signals in vmPFC. To test this, we correlated vmPFC responses to the value of harming others (**Supplementary Fig. 2b**) with TPJ and ACC responses to Δs_{other} (**Supplementary Fig. 3b & 3e**).The correlations were not significant (vmPFC-TPJ: robust correlation, r = -0.04, 95% CI=[-0.43 0.41]; vmPFC-ACC: robust correlation, r = -0.02, 95% CI=[-0.41 0.40]).

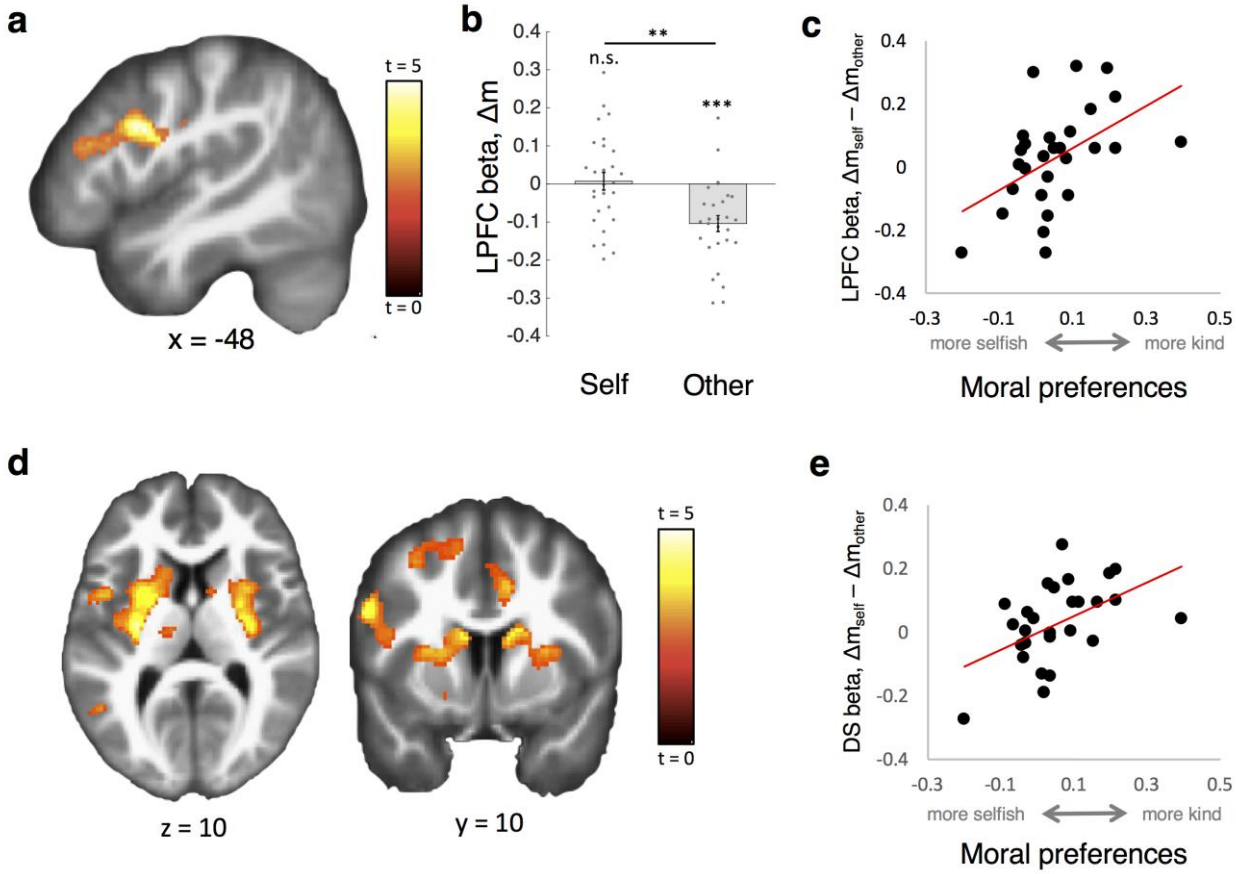
Neural representation of profit predicts moral behavior

To determine whether moral behavior relates to a differential neural sensitivity to profits gained from harming others vs. self, we recapitulated the previous analysis of pain, identifying voxels where activity varied parametrically with Δ*m*, irrespective of choice. We then asked whether these profit-sensitive regions showed differential sensitivity to profit gained from harming others, relative to harming oneself ($\Delta m_{self} > \Delta m_{other}$). This contrast revealed a strong effect in left LPFC (P_{FWE}=0.027, whole brain FWE corrected at the cluster level after voxel-wise thresholding at p<0.001; **Fig. 2a and Supplementary Table 4**). To probe the nature of this effect we extracted mean signal from an independently defined ROI in LPFC separately for the self and other conditions. This revealed an insensitivity to profit gained from harming oneself ($\beta_{\Delta m\_self}$=0.007 $\pm$ 0.02, $t_{(27)}$=0.31, p=0.76; **Fig. 2b**) but a negative parametric response in LPFC to profit gained from harming others ($\beta_{\Delta m\_other}$=-0.10 $\pm$ 0.02; $t_{(27)}$=-4.97, p=0.00003). The LPFC response to ill-gotten gains did not significantly differ on trials where participants chose to harm vs. help others ($t_{(27)}$=0.16, p=0.87). The differential response in LPFC to profits gained from harming others vs. self was more pronounced as a function of the strength of expressed moral preferences, with more moral participants showing a stronger differential response to profit from self- vs. other-harm in LPFC (robust correlation, r=0.53, 95% CI [0.14 0.74]; **Fig. 2c**).

We then asked whether there were additional regions expressing differential activity for profits gained from harming self vs. others that in turn correlated with moral preferences. We regressed individual differences in moral preferences onto the group level maps of the parametric response to profiting from harming others relative to self ($\Delta m_{self} > \Delta m_{other}$). In addition to the previously observed effect in LPFC, we observed a robust effect in bilateral DS extending into insula ($P_{FWE}<0.0001$), superior temporal gyrus ($P_{FWE}=0.007$), posterior cingulate ($P_{FWE}<0.0001$), and posterior medial PFC ($P_{FWE}=0.0003$; all results whole brain FWE corrected at the cluster level after voxel-wise thresholding at $p<0.001$; **Fig. 2d** and **Supplementary Table 5**). This network overlapped substantially with regions where activity correlated with relative chosen value (conjunction analysis, $P_{FWE}<0.0001$, whole brain FWE corrected at the cluster level after voxel-wise thresholding at $p<0.001$; **Supplementary Table 6**). Furthermore, other-related profit signals in DS were significantly correlated with relative chosen value signals in vmPFC (robust correlation, r=0.40, 95% CI [0.02 0.77]).

To further investigate the relationship between moral preferences and DS responses to profits, we examined parametric responses in DS (mean signal extracted from independently defined ROI) to the amount of profit gained from harming self and others separately. Moral participants ($\kappa_{other} > \kappa_{self}$) showed a positive parametric response in DS to the amount of profit gained from harming oneself ($\beta_{\Delta m\_self}=0.05 \pm 0.02$, $t_{(17)}=2.60$, p=0.018), but not to profit gained from harming others ($\beta_{\Delta m\_other}=0.004 \pm 0.03$, $t_{(17)}=0.15$, p=0.88), and reductions in the DS response to profits gained from harming others (relative to self) correlated positively with moral preferences (robust correlation, r=0.49, 95% CI [0.20 0.72]; **Fig. 2e**). This suggests moral behavior might arise via an attenuation of DS responses to profits gained from harming others.

Figure 2. Moral transgressions modulate corticostriatal responses to profit. (a) At choice onset, left LPFC activity negatively correlated with the relative amount of profit gained from harming others, but not self ($\Delta m_{self} > \Delta m_{other}$; $P_{FWE}=0.027$, whole brain FWE corrected at the cluster level after voxel-wise thresholding at $p<0.001$). Image displayed at p<0.005, uncorrected to show extent of activation. **(b)** Mean signal from an independently defined ROI in LPFC, separately extracted for the self and other conditions, was uncorrelated with Δm_{self} ($t_{(27)}=0.31$, p=0.76) but negatively correlated with Δm_{other} ($t_{(27)}=-4.97$, p=0.00003). **(c)** Differential LPFC response to profits gained from harming self vs. others positively correlated with individual differences in moral preferences (robust correlation, r=0.53, 95% CI [0.14 0.74]). **(d)** Image shows a second-level parametric map of moral preferences regressed against the contrast $\Delta m_{self} > \Delta m_{other}$. For participants showing stronger moral preferences, DS was less responsive to profits gained from harming others than profits gained from harming self ($P_{FWE}<0.0001$, whole brain FWE corrected at the cluster level after voxel-wise thresholding at $p<0.001$). Image displayed at p<0.005, uncorrected to show extent of activation. **(e)** Parameter estimates for Δm_{other} and Δm_{self} were extracted from an independently defined ROI in DS. Reduced DS responses to profits gained from harming others (relative to self) were positively correlated with moral preferences (robust correlation, r=0.49, 95% CI [0.20 0.72]). Error bars depict s.e.m. **P<0.01; ***P<0.0001; n.s., nonsignificant.



Computation of moral value in LPFC

During moral decision making LPFC responded more strongly on trials where participants could harm others for a low profit, relative to trials where harming resulted in a high profit. A similar pattern has been reported for blame judgments, where blame is higher for moral transgressions resulting in lower profits[36]. Thus, people may anticipate more blame for harmful decisions yielding lower profits, and this anticipated blame could be encoded by LPFC, in line with previous work showing LPFC responses to moral norm violations[24–30]. Directly testing this hypothesis required us to construct, for each participant, a trial-by-trial trajectory of anticipated blame and regress this against LPFC activity. Although participants in the fMRI study did not provide blame judgments, we hypothesized that within our study population blame could be predicted based on choice features (Δm, Δs) and individual preferences (κ_{self}, κ_{other}). This allowed us to infer blame judgments for the fMRI participants from a model of blame built using data from a separate group of participants who provided blame judgments in addition to moral decisions (**Fig. 3a**).

The model described blame judgments as follows:

$$Blame_t = \beta_0 + \beta_1 \Delta m_t + \beta_2 \Delta s_t + \beta_3 \Delta m_t \kappa_o + \beta_4 \Delta m_t \kappa_s + \beta_5 \Delta s_t \kappa_o + \beta_6 \Delta s_t \kappa_s + \beta_7 \Delta m_t \kappa_o \kappa_s + \beta_8 \Delta s_t \kappa_o \kappa_s$$

where blame on trial t is a linear function of choice features on trial t (Δm_t, Δs_t), individual preferences (κ_s, κ_o), and their interactive combinations (see **Supplementary Modeling Note** for parameter estimates; mean R^2=0.33). Blame was negatively correlated with profit ($\beta_{\Delta m}$=-0.07, 95%

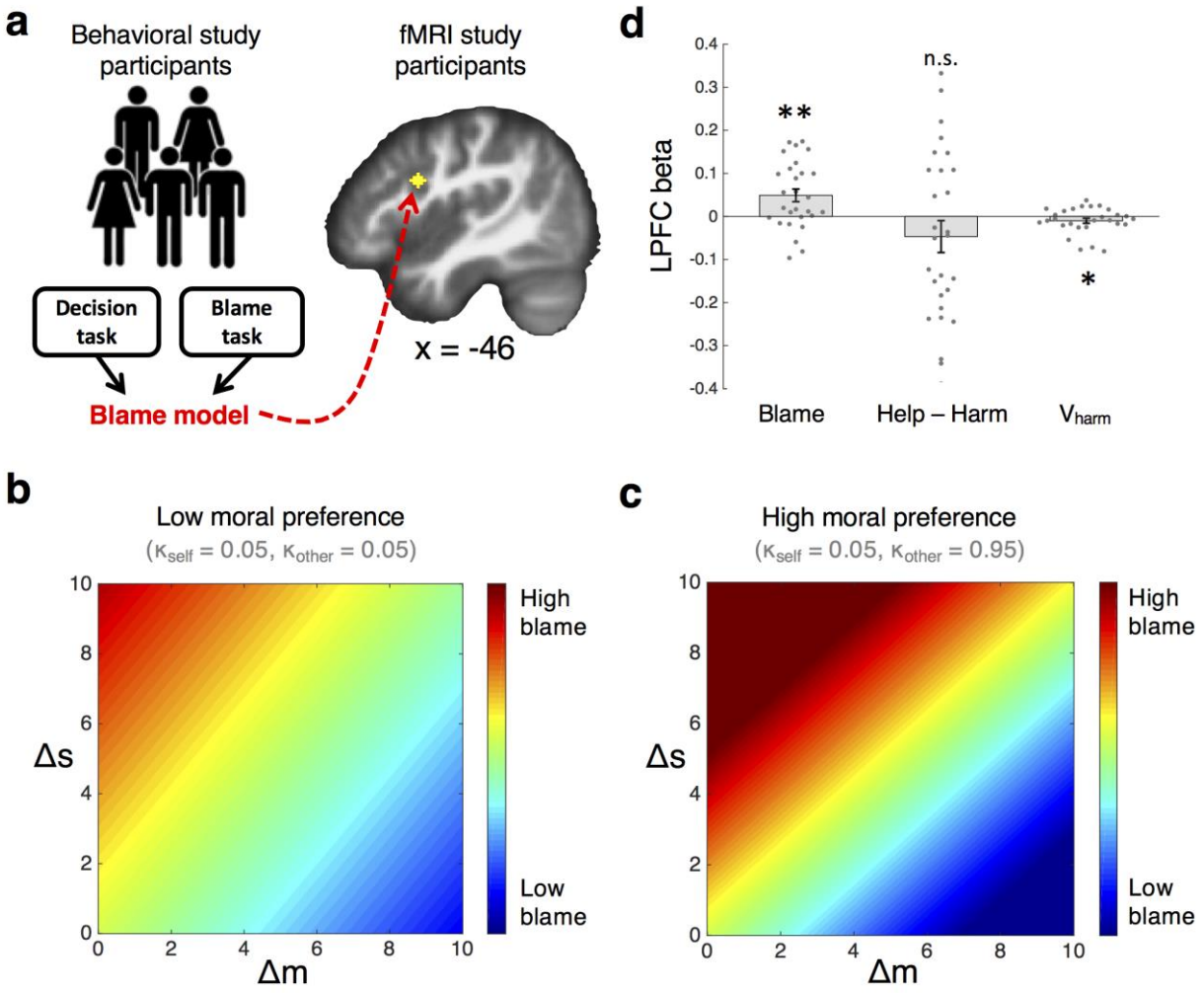
CI=[-0.10 -0.04]) and positively correlated with pain ($\beta_{\Delta s}$=0.05, 95% CI=[0.01 0.09]). Individual preferences modulated the relationship between profit, pain and blame such that participants with stronger moral preferences showed more extreme judgments and a stronger influence of pain (relative to money) on blame (**Fig. 3b-c**).

Next, we performed multiple regression on the BOLD signal extracted from an independently defined ROI in LPFC. Using the blame model described above we constructed a trajectory of anticipated blame for each fMRI participant and then tested whether LPFC signal correlated with anticipated blame estimates in a GLM that included anticipated blame, relative chosen value and total value, which were orthogonalised (GLM3). This analysis showed a positive correlation between LPFC activity and anticipated blame ($t_{(27)}$=2.67, p=0.012). The relationship between LPFC activity and anticipated blame remained significant when controlling for the total amounts of money and shocks on each trial (GLM4; $t_{(27)}$=2.84, p=0.008). If LPFC computes anticipated blame to guide decision making then this value should be choice-independent. Consistent with this, the relationship between LPFC and blame did not differ significantly on trials where participants harmed vs. helped ($t_{(27)}$=-0.86, p=0.40).

An alternative account of LPFC function in prosocial behavior is that this region serves a role akin to a "brake system" for inhibiting self-interested behavior, i.e., influencing policy selection once values are computed[31,37]. To rule out this account, we conducted several analyses, focusing on trials in the other condition. First, we examined participants' response times. If choosing the helpful option involves an inhibition of self-interest, then helpful choices should be slower than harmful choices (RT-GLM1, **Supplementary Table 7**). In fact, RTs were faster for helpful relative to harmful choices ($t_{(27)}$= -3.76, p=0.0008). RT data actually supported a blame computation account. If moral choices involve integrating moral value into overall subjective value, then people with stronger moral preferences should respond slower in the other condition (where moral values must be computed) relative to the self condition (which requires no such computation). We tested this by comparing RTs for other vs. self in a second GLM (RT-GLM2, **Supplementary Table 7**) and found slowing in the other relative to the self condition was indeed positively correlated with moral preferences (robust correlation, r=0.52, 95% CI [0.11 0.80]).

Next, we tested whether LPFC activity differed for harm vs. help trials. If LPFC is involved in inhibiting self-interest, then LPFC activity should be higher on "successful inhibition" trials, i.e., trials where participants chose the helpful option. Because participants were more likely to choose the helpful option on trials where harming would result in more blame ($t_{(27)}$=54.29, p=4 x 10^{-29}), we controlled for blame as well as relative chosen value and total value (GLM5). We found no difference in LPFC activity between help trials and harm trials ($t_{(27)}$=-0.96, p=0.35), while the effect of blame on LPFC activity remained significant ($t_{(27)}$= 2.89, p=0.007; **Fig. 3d**). Finally, we tested whether LPFC responses on help trials depend on the value of the harmful option (GLM6). If LPFC is required for inhibiting self-interest, it should be more active on trials where helping was more difficult (i.e., when the value of the harmful option was high). In fact, LPFC was not more active on help trials when the value of the harmful option was high (help*V_{harm} interaction, $t_{(27)}$= -1.67, p =0.11). Rather, LPFC was *less* active on trials where the value of the harmful option was high, regardless of whether participants harmed or helped (V_{harm} main effect, $t_{(27)}$= -2.21, p =0.04; **Fig. 3d**). Thus, our findings are not consistent with the notion that moral behavior involves an inhibition of self-interest implemented by LPFC or indeed anywhere else in the brain.

Figure 3. Blame computation in LPFC. (a) Behavioral study participants (N=49) completed a moral decision task and a moral blame task. These data were used to construct a model of blame, which we used to construct a unique blame trajectory for each participant in the fMRI study (N=28). We then regressed these blame estimates against activity in an independently defined ROI in LPFC. **(b-c)** Model estimates of blame as a function of profit (Δ*m*) and pain (Δ*s*) for a participant with low **(b)** and high **(c)** moral preferences. Across all participants, blame was highest for choices inflicting high pain for low profit. **(d)** LPFC signal was positively correlated with blame (GLM5; $t_{(27)}$= 2.89, p=0.007); did not significantly differ for helpful vs. harmful choices (GLM5; $t_{(27)}$=-0.96, p=0.35); and was negatively correlated with the subjective value of the harmful option V_{harm} (GLM6; $t_{(27)}$= -2.21, p =0.04). Error bars depict s.e.m. *P<0.05; **P<0.01; n.s., nonsignificant.



Moral decisions modulate corticostriatal connectivity

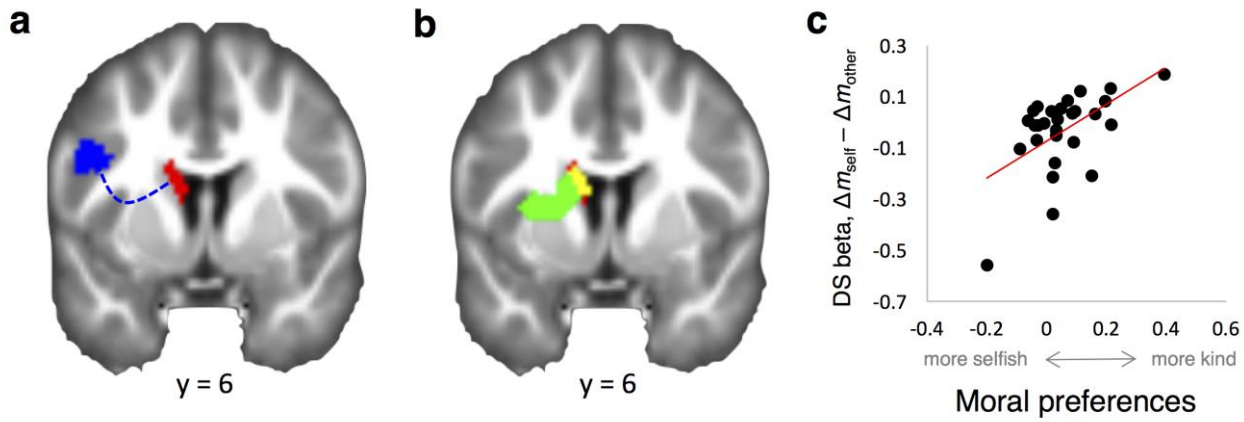
Moral preferences were associated with reduced striatal responses to profit from harming others, relative to self, while LPFC responses correlated with model estimates of blame. This suggests LPFC may modulate DS responses to profit in line with anticipated blame. This would predict that LPFC should show differential functional connectivity with DS during decisions to help others, compared with decisions to harm others and decisions to help oneself. Consequently, we implemented

psychophysiological interaction (PPI) analyses with LPFC as a seed region. The PPI models included regressors for the main effect of LPFC activity, the main effect of decision type, and their interaction. We examined two decision types in two separate PPI models: (i) helpful choices in the other condition relative to harmful choices in the other condition (help-other > harm-other), and (ii) helpful choices in the other condition relative to helpful choices in the self condition (help-other > help-self). This analysis revealed differential functional connectivity between LPFC and DS during help-other choices, relative to harm-other choices and help-self choices (conjunction of contrasts (i) & (ii): P_{FWE}=0.025, whole brain FWE corrected at the cluster level after voxel-wise thresholding at p<0.001; **Fig. 4a, Supplementary Fig. 4a** and **Supplementary Table 8**). This region overlapped with the cluster in DS identified above as showing differential responses to profits from harming self vs. others as a function of moral preferences (**Fig. 4b**). The results are consistent with a model whereby LPFC modulates DS responses to profit in line with moral considerations.

Participants with stronger moral preferences showed a greater reduction in the DS response to profits gained from harming others relative to self. If translating moral norms into moral behavior involves changes in functional connectivity between LPFC and DS, then we should also see a relationship between moral preferences and reduced responses to profiting from others' pain in the precise area of DS that was functionally connected with LPFC. To test this hypothesis we extracted for each participant the contrast estimate ($\Delta m_{self} > \Delta m_{other}$) from the DS cluster identified in the PPI analysis, and regressed these estimates against participants' moral preferences. This analysis revealed a positive correlation between moral preferences and reductions in the DS response to profiting from harming others relative to self (robust correlation, r=0.36, 95% CI [0.002 0.68]; **Fig. 4c**). That is, moral decisions modulated functional connectivity between LPFC and DS, and the extent to which the DS showed a reduced response to profiting from others' pain relative to one's own pain predicted moral preferences.

We next tested whether corticostriatal connectivity was associated with choice-related striatal value signals. As an index of value sensitivity in DS, we extracted for each participant the mean signal from the voxels in DS that were sensitive to relative chosen value ($t_{(27)}$ = 3.28, p = 0.003). As an index of corticostriatal connectivity during helpful moral choices, we extracted from this same region of DS the signal from the PPI contrast LPFC * [Help other]. Choice-related value signals in DS were negatively correlated with corticostriatal connectivity (robust correlation, r =-0.51, 95% CI [-0.73 -0.14]; **Supplementary Fig. 4b**), suggesting a negative connectivity between LPFC and DS during moral decisions as a function of value sensitivity in DS.

Figure 4. Corticostriatal connectivity during the exercise of moral choices. (a) Moral decisions modulated functional connectivity between seed region in LPFC (blue) and DS (red). Red cluster in DS (P_{FWE}=0.025, whole brain FWE corrected at the cluster level after voxel-wise thresholding at p<0.001) depicts conjunction of PPI contrasts (LPFC seed * help other > harm other) and (LPFC seed * help other > help self). Image displayed at p<0.005, uncorrected to show extent of activation. **(b)** The area of DS showing differential functional connectivity with LPFC during moral decisions (red) overlapped considerably with the area of DS showing reduced responses to profits gained from harming others vs. self (green; overlap shown in yellow). Image displayed at p<0.005, uncorrected to show extent of activation. **(c)** Within the area of DS showing differential functional connectivity with LPFC during moral decisions (red cluster), reduced responses to profits gained from harming others (relative to self) predicted moral preferences (robust correlation, r=0.36, 95% CI [0.002 0.68]).



Discussion

We offer an account of how a moral prohibition against harming others is translated into moral behavior. Replicating previous findings[15,16], we show most people prefer to harm themselves over others for profit. This moral preference was associated with diminished neural responses in value-sensitive regions to profit accrued from harming others. This observation suggests a neural explanation for why people are reluctant to seek profits from immoral actions[15–17], and disapprove of individuals and organizations who accept money from morally tainted sources[36,38], in revealing that moral transgressions corrupt neural representations of value.

Our findings implicate LPFC in computing the moral value of actions so as to guide moral decision making. LPFC negatively encoded the magnitude of profits gained from harming others but not self, and the strength of this encoding predicted individual differences in moral behavior. LPFC was most active on trials where inflicting pain yielded minimal profit, the very same trials considered most blameworthy by a second group of participants who provided blame judgments of decisions to harm others for profit. A model of blame built from these judgments predicted LPFC activity in fMRI participants, consistent with a role for LPFC in representing moral values.

Previous accounts of LPFC in prosocial behavior have distinguished between inhibitory "braking" functions and executive "overriding" functions, generally attributed to more ventral and dorsal aspects of LPFC, respectively[39]. The region we observed here, situated in the inferior frontal gyrus, did not show activity consistent with inhibitory control. It was not more active during helpful decisions than harmful decisions, nor when helpful choices were more difficult. Instead, its activity pattern suggested an encoding of moral goals, which may be used to modulate action values represented in DS[23,31]. This mechanism has parallels with models of self-control in non-social decision making, whereby long-term goals represented in LPFC modulate neural representations of value[22,34].

We tested this hypothesis with functional connectivity analyses and found reduced connectivity between LPFC and DS during the exercise of helpful choices, relative to harmful choices and non-social choices. The DS, in turn, showed reduced responses to profit from harming others to the extent that people behaved morally. The connectivity findings suggest two possible mechanisms that are not mutually exclusive. First, they could reflect a negative corticostriatal connectivity during helpful decisions, which would suggest LPFC directly down-regulates value representations in DS at the time of choice[34]. Alternatively, they could reflect a greater positive

connectivity during harmful decisions relative to helpful decisions, which could result from LPFC updating action value representations in DS following perceived moral transgressions. This latter account follows recent work suggesting people are uncertain about their preferences and use social information to "learn" what to want[40]. Our data are agnostic on this point as our study was not designed to arbitrate between these mechanisms, but future studies could usefully investigate possible links between moral decision making and moral learning, including the question of whether harm aversion declines over time[41]. We acknowledge that our conclusions are limited by the correlational nature of neuroimaging findings, and suggest future studies employ brain stimulation or lesion-deficit analyses to deduce the causal role of LPFC in blame computation beyond domain-general attentional or control functions.

We previously reported that acutely enhancing dopamine levels with the dopamine precursor levodopa disrupted moral preferences[16]. The current findings hint at a possible mechanism for this effect. Moral preferences were associated with reduced DS responses to profit gained from harming others relative to self, and this region showed differential functional connectivity with the LPFC during moral decisions. Levodopa may disrupt this corticostriatal circuit by amplifying phasic dopamine signals in the striatum that guide action selection[42,43], biasing choice toward immediate rewards (i.e., profits) and away from higher-order values (i.e., norm compliance). Such a mechanism would be consistent with reports that antisocial and aggressive behaviors are associated with heightened striatal dopamine[44,45].

Neural representations of others' pain during moral decision making did not correlate with moral preferences in our study. Although previous work has argued such representations play a prominent role in mediating moral behavior[10,14], these conclusions focused on neural responses to observing others in pain, rather than neural computations during moral decision making. As neural representations of others' pain have been linked to empathy[18,19], our study informs current debates on the extent to which empathy guides moral behavior[46] and highlight the importance of norms in restraining self-interest. Our findings suggest that neural responses elicited by the potential suffering of anonymous strangers may be dissociated from the moral choices people make, in line with evidence that empathy and a motivation to help others are psychologically distinct[46] and the latter (but not the former) predicts LPFC responses during costly altruism[8].

Finally, our results shed light on the question of whether moral decisions engage a specialized set of neural computations, or simply rely on the same circuitry that is involved in generic value-based decision making. Previous studies have shown that moral judgments and decisions engage similar neural circuitry as observed for simple value-based decisions[47], but the encoding of subjective value for moral and non-moral decisions had never been directly compared in the same study. Here we found that the overall subjective value of moral choices was ultimately reflected in the same regions that encoded the subjective value of non-social choices. Moral preferences were reflected in a reduced vmPFC response to the value of harming others, which could reflect either a directly reduced utility for ill-gotten gains, or a partially corrupted mapping of utility onto vmPFC signal for ill-gotten gains. However, we also observed profit-related computations in the LPFC during moral decision making that were not observed during non-social choices of a similar nature. Responses to ill-gotten gains in LPFC correlated with moral preferences, and this region expressed altered functional connectivity with value-encoding regions during moral choices, relative to non-moral choices. Thus, the construction of moral values seems to incorporate additional computations that may represent anticipated or internalized moral

judgments of others. In this way, our conscience, the "great judge and arbiter of our conduct"[2], may influence the values that guide the choices we make.

Acknowledgments

We thank E. Boorman, A. de Berker, L. Hunt, M. Klein-Flugge, C. Mathys, R. Rutledge, B. Seymour, P. Smittenaar, G. Story, I. Vlaev, and J. Winston for helpful feedback. M.J.C. was supported by a Sir Henry Wellcome Postdoctoral Fellowship (092217/Z/10/Z) and a Wellcome Trust Institutional Strategic Support Fund grant. J.Z.S. was supported by a Wellcome Trust Society and Ethics studentship (104980/Z/14/Z). Z.K.-N. was supported by a Joint Initiative on Computational Psychiatry and Ageing Research between the Max Planck Society and University College London. P.D. is funded by the Gatsby Charitable Foundation. RJD holds a Wellcome Trust Senior Investigator Award (098362/Z/12/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Trust Centre for Neuroimaging, where scanning was carried out, is supported by core funding from the Wellcome Trust (091593/Z/10/Z).

Author contributions

M.J.C. conceived the study. M.J.C., J.Z.S., Z.K.N., P.D. and R.D. designed the study. M.J.C. and J.Z.S. collected behavioral and fMRI data. M.J.C., J.Z.S., Z.K.N. and P.D. analyzed the data. M.J.C. wrote the manuscript with edits from J.Z.S., Z.K.N., P.D. and R.D.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Gert, B. *Common Morality: Deciding What to Do*. (Oxford University Press, 2004).

2. Smith, A. *The Theory of Moral Sentiments*. (1759).

3. Boehm, C. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. (Basic Books, 2012).

4. Rand, D. G. & Epstein, Z. G. Risking Your Life without a Second Thought: Intuitive Decision-Making and Extreme Altruism. *PLoS ONE* **9,** e109687 (2014).

5. Marsh, A. A. *et al.* Neural and cognitive characteristics of extraordinary altruists. *Proc. Natl. Acad. Sci.* **111,** 15036–15041 (2014).

6. Greene, J. D. in *The Cognitive Neurosciences V* 1013–1023 (MIT Press, 2014).

7. FeldmanHall, O. *et al.* Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Soc. Cogn. Affect. Neurosci.* **7,** 743–751 (2012).

8. FeldmanHall, O., Dalgleish, T., Evans, D. & Mobbs, D. Empathic concern drives costly altruism. *Neuroimage* **105,** 347–356 (2015).

9. Greene, J. D. & Paxton, J. M. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci.* **106,** 12506–12511 (2009).

10. Yu, H., Hu, J., Hu, L. & Zhou, X. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc. Cogn. Affect. Neurosci.* **9,** 1150–1158 (2014).

11. Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C. & Fehr, E. Linking Brain Structure and Activation in Temporoparietal Junction to Explain the Neurobiology of Human Altruism. *Neuron* **75,** 73–79 (2012).

12. Hutcherson, C. A., Bushong, B. & Rangel, A. A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron* **87,** 451–462 (2015).

13. Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P. & Rangel, A. Value Computations in Ventral Medial Prefrontal Cortex during Charitable Decision Making Incorporate Input from Regions Involved in Social Cognition. *J. Neurosci.* **30,** 583–590 (2010).

14. Hein, G., Silani, G., Preuschoff, K., Batson, C. D. & Singer, T. Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron* **68,** 149–160 (2010).

15. Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P. & Dolan, R. J. Harm to others outweighs harm to self in moral decision making. *Proc. Natl. Acad. Sci.* **111,** 17320–17325 (2014).

16. Crockett, M. J. *et al.* Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision-making. *Curr. Biol.* (2015).

17. Stellar, J. E. & Willer, R. The Corruption of Value Negative Moral Associations Diminish the Value of Money. *Soc. Psychol. Personal. Sci.* **5,** 60–66 (2014).

18. Zaki, J. & Ochsner, K. N. The neuroscience of empathy: progress, pitfalls and promise. *Nat. Neurosci.* **15,** 675–680 (2012).

19. Lamm, C., Decety, J. & Singer, T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* **54,** 2492–2502 (2011).

20. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76,** 412–427 (2013).

21. Clithero, J. A. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9,** 1289–1302 (2014).

22. Rangel, A. & Hare, T. Neural computations associated with goal-directed choice. *Curr. Opin. Neurobiol.* **20,** 262–270 (2010).

23. Buckholtz, J. W. & Marois, R. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* **15,** 655–661 (2012).

24. Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. & Fehr, E. The Neural Signature of Social Norm Compliance. *Neuron* **56,** 185–196 (2007).

25. Ruff, C. C., Ugazio, G. & Fehr, E. Changing Social Norm Compliance with Noninvasive Brain Stimulation. *Science* **342,** 482–484 (2013).

26. Chang, L. J., Smith, A., Dufwenberg, M. & Sanfey, A. G. Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron* **70,** 560–572 (2011).

27. Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C. & Fehr, E. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nat. Neurosci.* **14,** 1468–1474 (2011).

28. Buckholtz, J. W. *et al.* From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. *Neuron* **87,** 1369–1380 (2015).

29. Treadway, M. T. *et al.* Corticolimbic gating of emotion-driven punishment. *Nat. Neurosci.* **17,** 1270–1275 (2014).

30. Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex. *Science* **314,** 829–832 (2006).

31. Buckholtz, J. W. Social norms, self-control, and the value of antisocial behavior. *Curr. Opin. Behav. Sci.* **3,** 122–129 (2015).

32. Haber, S. N., Kim, K.-S., Mailly, P. & Calzavara, R. Reward-Related Cortical Inputs Define a Large Striatal Region in Primates That Interface with Associative Cortical Connections, Providing a Substrate for Incentive-Based Learning. *J. Neurosci.* **26,** 8368–8376 (2006).

33. Choi, E. Y., Yeo, B. T. T. & Buckner, R. L. The organization of the human striatum estimated by intrinsic functional connectivity. *J. Neurophysiol.* **108,** 2242–2263 (2012).

34. Bos, W. van den, Rodriguez, C. A., Schweitzer, J. B. & McClure, S. M. Connectivity Strength of Dissociable Striatal Tracts Predict Individual Differences in Temporal Discounting. *J. Neurosci.* **34,** 10298–10310 (2014).

35. Daw, N. D. in *Decision Making, Affect, and Learning: Attention and Performance XXIII* (eds. Delgado, M. R., Phelps, E. A. & Robbins, T. W.) (Oxford University Press, 2011).

36. Xie, W., Yu, B., Zhou, X., Sedikides, C. & Vohs, K. D. Money, moral transgressions, and blame. *J. Consum. Psychol.* **24,** 299–306 (2014).

37. Dolan, M. The neuropsychology of prefrontal function in antisocial personality disordered offenders with varying degrees of psychopathy. *Psychol. Med.* **42,** 1715–1725 (2012).

38. Inbar, Y., Pizarro, D. A. & Cushman, F. Benefiting From Misfortune When Harmless Actions Are Judged to Be Morally Blameworthy. *Pers. Soc. Psychol. Bull.* **38,** 52–62 (2012).

39. Feng, C., Luo, Y.-J. & Krueger, F. Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Hum. Brain Mapp.* **36,** 591–602 (2015).

40. Moutoussis, M., Dolan, R. J. & Dayan, P. How People Use Social Information to Find out What to Want in the Paradigmatic Case of Inter-temporal Preferences. *PLOS Comput. Biol.* **12,** e1004965 (2016).

41. Garrett, N., Lazzaro, S. C., Ariely, D. & Sharot, T. The brain adapts to dishonesty. *Nat. Neurosci.* **19,** 1727–1732 (2016).

42. Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.* **15,** 1281–1289 (2012).

43. Macpherson, T., Morita, M. & Hikida, T. Striatal direct and indirect pathways control decision-making behavior. *Front. Psychol.* **5,** (2014).

44. Buckholtz, J. W. *et al.* Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. *Nat. Neurosci.* **13,** 419–421 (2010).

45. Couppis, M. H., Kennedy, C. H. & Stanwood, G. D. Differences in aggressive behavior and in the mesocorticolimbic DA system between A/J and BALB/cJ mice. *Synapse* **62,** 715–724 (2008).

46. Jordan, M. R., Amir, D. & Bloom, P. Are empathy and concern psychologically distinct? *Emotion* **16,** 1107–1116 (2016).

47. Ruff, C. C. & Fehr, E. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15,** 549–562 (2014).

Online Methods

Participants

Healthy volunteers were recruited from the University College London (UCL) Psychology department and the Institute of Cognitive Neuroscience subject pools. Participants with a history of systemic or neurological disorders, psychiatric disorders, medication/drug use, pregnant women, previous participation in studies involving social interactions and/or electric shocks, or more than two years' study of psychology were excluded from participation. For the fMRI study we recruited right-handed participants only.

For the fMRI study, we recruited thirty-seven pairs of participants, with one participant in each pair completing a moral decision task in the fMRI scanner. No statistical methods were used to pre-determine sample sizes but our sample size was based on estimated effect size for moral preferences observed in two previous behavioral studies using the same task[15]. Two participants indicated they did not find the shocks aversive, three participants fell asleep in the scanner, one participant failed to follow task instructions, one participant expressed doubts as to whether the receiver would receive the shocks, and one participant requested to exit the scanner during the first run. A power cut resulted in data loss for another participant. These participants were excluded from further analysis, leaving a total of 28 participants in the role of decider whose data were analyzed for the fMRI study (16 males, mean age 21.9y).

For the behavioral study, fifty-four participants participated in the role of decider. These participants completed the moral blame task after completing a moral decision task similar to the fMRI task; moral decision data from these participants has been published previously[15]. Five

participants did not provide sufficient variation in their blame judgments to allow for model fitting (more than 75% identical judgments or a standard deviation in judgments < 0.03); these participants were excluded from further analysis, leaving a total of 49 participants whose data were analyzed for the behavioral study (18 males, mean age 23y).

Procedure

Both studies took place at the Wellcome Trust Centre for Neuroimaging in London and was approved by the UCL Research Ethics Committee (4418/001). Participants completed a battery of online trait questionnaires approximately 1 week before attending a single testing session. Two individuals participated in each session. They arrived at staggered times and were led to separate testing rooms without seeing one another to ensure complete anonymity.

In both studies, after providing informed consent, participants completed a pain thresholding procedure that has been described in detail elsewhere[15]. This procedure allowed us to (i) control for heterogeneity of skin resistance between participants, thus enabling us to deliver shocks of matched subjective intensity to different participants; (ii) administer a range of potentially painful stimuli in an ethical manner during the task itself; and (iii) provide participants with experience of the shocks before the decision task. Participants were then randomly assigned to roles of either decider or receiver using a role assignment procedure that has been described in detail elsewhere[15].

In the fMRI study, following role assignment the decider participant completed the moral decision task in the fMRI scanner. In the behavioral study, following role assignment the decider participant completed the moral decision task, followed by the moral blame task. Deciders were instructed their choices and identity would be kept confidential to minimize the extent to which their choices would be based on concerns about reputation or reciprocity.

After completing the decision task, decider participants completed self-report measures concerning their experiences during the experiment, including a measure of how morally conflicted they felt about their decisions (rated on a 7-point Likert scale, 1="not at all", 7="very much"). At the end of the session, one trial was randomly selected and actually implemented. Before departing the laboratory all participants completed debriefing questionnaires that assessed their beliefs about the experimental setup, including a measure of confidence that their choices and identity would remain confidential (rated from 1="fully" to 5="not at all"). Participants reported maximal confidence (decisions confidential: M=1.04, SE=0.04; identity confidential: M=1.04, SE=0.04). Finally, we asked participants to explain, in their own words, how they made their decisions during the experiment (**Supplementary Table 9**). No participant mentioned concerns about their reputation or reciprocity, while 86% of participants used language indicative of value computation (e.g., "worth", "value", "calculate"). Only 7% of participants mentioned concerns about the pain tolerance of the receiver.

Moral decision task

On each trial, deciders had to choose between two options involving pairs of numbers of shocks and amounts of money: a *harmful option* containing more shocks and money, and a *helpful option* containing fewer shocks and less money. The decider always received the money, but the shocks were allocated to the decider in half of the trials (self condition) and to the receiver in the

other half (other condition). Deciders had a maximum of six seconds to select either the left or right side option by pressing a button box with their left or right index finger. Button presses resulted in the selected option being highlighted for the remainder of the six-second decision period. If a response was not made within six seconds, the missed trial was repeated at the end of the session. Transitions between conditions were cued with an instruction screen lasting two seconds. Each trial culminated in an inter-trial interval jittered between one and six seconds. Participants completed a total of 216 trials, delivered across two scanning runs lasting approximately twenty minutes each. To avoid habituation and preserve choice independence no money or shocks were delivered during the task. Instead, one trial was randomly selected and implemented at the end of the experiment. All procedures were fully transparent to participants, and no deception was used in the paradigm.

Our trial set was optimized to jointly satisfy two constraints. First, we aimed to optimize the trials to give the most efficient estimates of potential participants' harm aversion parameters κ_{self} and κ_{other}. Second, we aimed to de-correlate, across trials, the relative amounts of profit and pain that would result from participants' choices. We satisfied the first constraint using a procedure described in detail elsewhere[15] to create a set of 54 trials that efficiently estimated participants' harm aversion parameters. We then repeated this procedure 10,000 times. For each iteration we simulated choices on the trial set across a range of values of κ_{self} and κ_{other}, computed the correlations between the amounts of profit and pain resulting from simulated choices, and selected the trial set that resulted in the lowest correlation between parameters. After creating this optimized trial set, we duplicated it and reversed the left and right options, producing a full set of 108 trials. Participants completed each of these 108 trials in both the self condition and the other condition, for a total of 216 trials. Thus, each money/shock pair appeared four times: twice in each condition (self/other) and twice on each side (left/right). We created four different trial sequences that each contained an equal number of self- and other-trials in the first and second blocks of 108 trials, and participants were randomly assigned to receive one of the four trial sequences.

The trial optimization procedure successfully satisfied our constraints: across trials, the amounts of profit and pain that would result from choosing the more harmful option were uncorrelated (r=0.009, p=0.926; **Fig. 1b**). In addition, there were no significant correlations between the relative and total number of shocks across the two options (r=-0.02, p=0.84), nor between the relative and total amounts of money across the two options (r=-0.13, p=0.18). Across participants, relative and total subjective values were also not significantly correlated (r=-0.14, p=0.15). This suggests our findings related to relative values are unlikely to be explained by overall value.

Moral blame task

Participants evaluated sequences of 30 moral decisions made by two fictional agents, presented in random order, for a total of 60 trials. Trials were self-paced. On each trial, agents faced a choice similar to that faced by deciders in the fMRI study, i.e., they had to choose between delivering more painful electric shocks to another person for a larger profit, and delivering fewer shocks but for a smaller profit. We used our model of moral decision making[15] to simulate the choices of a bad agent with κ_{other}=0.3 who mostly chose the harmful option, and a good agent with κ_{other}=0.7 who mostly chose the helpful option. After observing each choice, participants provided

a moral judgment of the choice on a continuous visual analogue scale ranging from 0 (*blameworthy*) to 1 (*praiseworthy*) (**Supplementary Fig. 1**). Across trials, we independently manipulated the amounts of profit and pain resulting from the agent's choices, which enabled us to examine how profit and pain resulting from harmful choices load onto blame judgments.

fMRI acquisition and preprocessing

fMRI scanning was performed on a 3-Tesla Siemens Allegra scanner with a Siemens head coil at the Wellcome Trust Centre for Neuroimaging at University College London. Functional images were taken with a gradient echo T2*-weighted echo-planar sequence (repetition time=2.40 s, echo time=30 ms, flip angle=90o, 64x64 matrix, field of view=192 mm, slice thickness=2 mm with 1 mm gap). A total of 40 axial slices were acquired in ascending order (in-plane resolution 3x3 mm). 428 volumes were acquired in each of two sessions and the initial five volumes of each session were discarded to allow for steady-state magnetization. Slices were tilted at an orientation of -30 degrees to minimize signal dropout in ventral frontal cortex. Anatomical images were T1-weighted (MDEFT, 1x1x1 mm resolution). We also acquired a field map (double-echo FLASH, short TE=10 ms, long TE=12.46 ms, 3x3x3 mm resolution with 1 mm gap) for distortion correction of functional images. We used a breathing belt and pulse oximeter for collecting physiological data during the imaging sessions.

All image preprocessing and analysis was carried out in SPM8 (Wellcome Department of Imaging Neuroscience). Images were realigned to the first scan of the first session and unwarped using field maps, spatially normalized via segmentation of the T1 structural image into gray matter, white matter, and CSF using ICBM tissue probability maps, and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum).

GLM1: model of relative chosen value

We constructed a GLM to identify regions responding parametrically at decision onset to the subjective value of the chosen and unchosen options, as determined by our computational model of choice. We regressed fMRI time series onto a GLM containing four main event regressors describing the onsets of (i) self trials where the left option was selected; (ii) self trials where the right option was selected; (iii) other trials where the left option was selected; and (iv) other trials where the right option was selected. All four events were modeled with a duration corresponding to the participant's RT on that trial and were each associated with two parametric modulators: the subjective value of the chosen and unchosen options, derived from each participant's choice model. Critically, custom scripts ensured that these four parametric modulators competed for variance during the estimation, rather than being serially orthogonalized as is standard in SPM. The GLM contained four additional event regressors of no interest, describing the onsets of: (i) left button presses; (ii) right button presses; (iii) screen signaling transition to self condition; and (iv) screen signaling transition to other condition. These events were modeled as stick functions with duration zero. Finally, a total of 23 nuisance regressors were included to control for motion and physiological effects of no interest. These included the six motion regressors obtained during realignment, as well as 17 physiological regressors derived from a physiological noise model, constructed using an in-house Matlab toolbox[48]: ten for cardiac phase, six for respiratory phase, and one for respiratory volume.

GLM2: model of decision parameters

We built a general linear model (GLM) to identify regions responding parametrically at decision onset, irrespective of participants' choices, to the objective amounts of profit and pain that would result from choosing the harmful option, relative to the helpful option, in the self and other conditions. We regressed fMRI time series onto a GLM containing four main event regressors describing the onsets of (i) self trials where the left option was selected; (ii) self trials where the right option was selected; (iii) other trials where the left option was selected; and (iv) other trials where the right option was selected. All four events were modeled with a duration corresponding to the participant's RT on that trial and were each associated with four parametric modulators: the amount of profit and pain for the harmful and the helpful option, irrespective of what the participant chose. Again we ensured that these two parametric modulators competed for variance during the estimation. There were four additional event regressors of no interest, indicating the onsets of button presses and transitions between task conditions, and 23 nuisance regressors controlling for motion and physiological effects of no interest.

GLM3-6: ROI analysis in LPFC

To test different accounts of the role of LPFC in moral decision making, we extracted time course data from a 4mm sphere surrounding independently defined ROI coordinates in LPFC. Because we were interested in moral decisions we restricted these analyses to trials in the other condition. Custom MATLAB scripts were used to orthogonalize each participant's time series with respect to motion & physiological regressors and apply a high-pass filter. The time series were then normalized, up-sampled at 100 ms, and time-locked to decision onsets, creating a data matrix with dimensions nTrials x nTimepoints. Next we fit a GLM across trials separately for each participant, resulting in parameter estimates at each time point for each GLM regressor. To test the significance of each regressor in LPFC, we convolved the regressor time series with a canonical hemodynamic response function aligned to decision onset and calculated resulting *t* statistics and *p* values. The latency for the canonical hemodynamic response function was estimated using the CANlab Core Tools package[49].

We tested 4 GLMs in LPFC using the above procedure. **GLM3**: $y=B_1*vdiff + B_2*vtot + B_3*blame + e$; **GLM4**: $y=B_1*vdiff + B_2*m_{tot} + B_3*s_{tot} + B_4*blame + e$; **GLM5:** $y=B_1*help + B_2*vdiff + B_3*vtot + B_4*blame + e$; and **GLM6**: $y=B_1*v_{help} + B_2*v_{harm} + B_3*v_{help}*help + B_4*v_{harm}*help + B_5*vdiff + e$; *vdiff* refers to relative chosen value(i.e., the value of the chosen option relative to the unchosen option), *vtot* refers to sum of the values of the chosen and unchosen options, m_{tot} refers to the total amount of money available on a trial, s_{tot} refers to the total number of shocks available on a trial, v_{harm} refers to the value of the harmful option, v_{help} refers to the value of the helpful option, and *e* denotes an error term. Value regressors and the blame regressor were computed individually for each participant based on their individual preference parameters estimated from the moral decision model.

PPI model: functional connectivity with LPFC

We created LPFC seed regressors by computing individual average time series within 4mm spheres surrounding individual subject peaks within the functional masks of left LPFC as shown in Fig. 2A. The locations of the peak voxels were based on the GLM2 contrast showing parametric effects of profit resulting from choosing the more harmful option in the other condition, relative to the self condition. Variance associated with the six motion regressors was removed from the extracted time series. To construct a time series of neural activity in the left LPFC, the seed time courses were de-convolved with the canonical hemodynamic response function. We then estimated the first PPI model (PPI1) with the following regressors: (1) an interaction between the neural activity in LPFC and a vector coding for the main effect of decision type (1 for help other, -1 for harm other); (2) the main effect of decision type; and (3) the original BOLD eigenvariate (i.e., the average time series from the LPFC seed), as well as the six motion parameters as regressors of no interest. We also estimated a second, complementary PPI model (PPI2) that was identical to the first model, except the first regressor contrasted decisions to help other with decisions to help self (1 for help other, -1 for help self).

Statistical analyses

We used a within-subjects design, so experimental group randomization and blinding were not applicable. Data analysis was not performed blind to the conditions of the experiments. We analyzed behavioral data using *t* tests and multiple linear regression. We analyzed fMRI data using mass univariate methods implemented in SPM8. At the first level we implemented linear regression at each voxel, using generalized least squares with a global approximate AR(1) autocorrelation model, drift fit with Discrete Cosine Transform basis (128s cutoff). At the second level we implemented linear regression at each voxel, using ordinary least squares. All Student's *t* tests were two-tailed. For correlations between brain responses and moral preferences we report the percentage bend correlation, which is robust to outliers, using the Matlab robust correlation toolbox[50]. The toolbox uses the 95% bootstrap confidence interval rather than p-values to make statistical inferences, because the 95% confidence interval is less affected by heteroscedasticity than the traditional t-test.

Moral decision task: choices. We analyzed the moral decision data with a model based on previous studies using the moral decision task[15,16] that explained choices in terms of the value difference (ΔV) between the harmful and helpful options. As the fMRI task required participants to select between two alternatives on each trial, rather than switch from a default to an alternative option as in previous studies, we omitted the loss aversion parameter from the model here. Trial-by-trial value differences were transformed into choice probabilities using a softmax function[35]:

$$P(choose\ alternative) = \left(\frac{1}{1 + e^{-\gamma \Delta V}}\right)$$

where γ is a subject-specific inverse temperature parameter that characterizes the sensitivity of choices to ΔV. We optimized participant-specific parameters across trials using nonlinear optimization implemented in MATLAB (MathWorks) for maximum likelihood estimation. Parameters were estimated individually for each participant, and summary statistics were calculated from these parameter estimates at the group level, treating each parameter estimate as

a random effect[51]. Parametric statistics were used to compare harm aversion for self and others as these parameters were normally distributed (Kolmogorov-Smirnov test statistic for κ_{other}=0.097, p=0.2; for κ_{self}=0.107, p=0.2). See **Supplementary Modeling Note** and **Supplementary Software** for details.

Moral decision task: RTs. We analyzed RT data using a GLM (RT-GLM1) that regressed RTs during the other condition against the following regressors: (1) dummy indicating helpful vs. harmful choices; (2) unsigned value difference; (3) total value; and (4) maximum number of shocks. In a second GLM (RT-GLM2) we regressed RTs during all conditions against the following regressors: (1) dummy indicating self vs. other condition; (2) unsigned value difference; (3) total value; and (4) maximum number of shocks. We optimized participant-specific parameters across trials using the glmfit procedure in Matlab. Parameters were estimated individually for each participant, and summary statistics were calculated from these parameter estimates at the group level, treating each parameter estimate as a random effect[51]. See **Supplementary Table 7** and **Supplementary Software** for details.

Moral blame task. We analyzed the moral blame data using a GLM that regressed participants' z-scored blame judgments onto the amounts of profit and pain resulting from harmful decisions (relative to helpful decisions) as well as individual preference parameters (κ_{self}, κ_{other}) estimated from the moral decision model. We also estimated a second, reduced model that included terms for profit and pain only. Because we were primarily interested in blame judgments for harmful choices, we restricted this analysis to trials on which the bad agent chose the harmful option. Group level parameters were estimated using the regress function in MATLAB. We report F statistics and p values from the full models as well as parameter estimates and 95% confidence intervals for each model in the **Supplementary Modeling Note**. We used the blame model, estimated on data from the behavioral study, to create a unique blame regressor for each participant in the fMRI study (**Fig. 3a**). These were computed by applying the parameters from the blame model to the amounts of profit and pain in the fMRI trials, and the fMRI participants' preferences parameters κ_{self} & κ_{other}. The blame regressors were used in GLM3-5. See **Supplementary Modeling Note** and **Supplementary Software** for details.

fMRI: Correction for multiple comparisons. For whole brain analyses, we tested for statistical significance using whole brain correction (p<0.05, FWE corrected at the cluster level after voxel-wise thresholding at *p*<0.001). This threshold provides an acceptable FWE control[52]. Post hoc analyses of regions identified in the whole brain analyses were carried out using one-sample t-tests on mean signal extracted from 4mm spheres surrounding independently defined ROIs (**Supplementary Table 2**). For ROI analyses, mean signal was extracted from 4mm spheres surrounding coordinates defined from previous studies. For analyses in TPJ, ACC and insula, we took coordinates from previous meta-analyses of empathy for pain and moral judgment[19,53]. For analyses in LPFC we took the mean of peak coordinates from previous studies investigating LPFC modulation of subjective value signals[54–58]. For analyses in DS we took coordinates from anatomical study of corticostriatal connectivity[33]. Images are displayed at a threshold of p<0.005, k>10 to show the extent of activation in significant clusters. Results are reported using the MNI coordinate system.

fMRI: PPI analysis. For the LPFC connectivity analysis we tested for statistical significance by conducting a conjunction analysis of the PPI contrasts from PPI1 & PPI2. To compute an appropriate threshold for the two-way conjunction of contrasts, we employed Fisher's method[59], a procedure that combines probabilities of multiple hypothesis tests using the following formula:

$$\chi^2 = -2\sum_{i=1}^{k} \log_e(p_i)$$

where p_i corresponds to the p-value for the i^{th} test being combined, k corresponds to the number of tests to be combined, and the resulting statistic has a χ^2 distribution with $2k$ degrees of freedom. According to this method, thresholding each contrast at p=0.01 resulted in a combined threshold of p<0.001, uncorrected. We report as significant results surviving whole brain correction for multiple comparisons (cluster-level corrected after voxel-wise thresholding at p<0.001).

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

Analysis code for fitting models to moral decision choice data, moral decision RT data and blame judgment data are provided as **Supplementary Software**. All other analysis code is available from the corresponding author upon reasonable request.

References

48. Hutton, C. *et al.* The impact of physiological noise correction on fMRI at 7 T. *NeuroImage* **57,** 101–112 (2011).

49. Lindquist, M.A., *et al.* Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage* **45**, S187-S198 (2009).

50. Pernet, C. R., Wilcox, R. R. & Rousselet, G. A. Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox. *Front. Psychol.* **3,** (2013).

51. Holmes, A. & Friston, K. Generalisability, random effects & population inference. *NeuroImage* **7,** (1998).

52. Flandin, G. & Friston, K. J. Analysis of family-wise error rates in statistical parametric mapping using random field theory. *ArXiv160608199 Stat* (2016).

53. Bzdok, D. *et al.* Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct. Funct.* **217,** 783–796 (2012).

54. Hare, T. A., Camerer, C. F. & Rangel, A. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science* **324,** 646–648 (2009).

55. Rudorf, S. & Hare, T. A. Interactions between Dorsolateral and Ventromedial Prefrontal Cortex Underlie Context-Dependent Stimulus Valuation in Goal-Directed Choice. *J. Neurosci.* **34,** 15988–15996 (2014).

56. Hare, T. A., Malmaud, J. & Rangel, A. Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice. *J. Neurosci.* **31,** 11077–11087 (2011).

57. Hare, T. A., Hakimi, S. & Rangel, A. Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. *Front. Neurosci.* **8,** (2014).

58. Maier, S. U., Makwana, A. B. & Hare, T. A. Acute Stress Impairs Self-Control in Goal-Directed Choice by Altering Multiple Functional Connections within the Brain's Decision Circuits. *Neuron* **87,** 621–631 (2015).

59. Fisher, R.A. Statistical methods for research workers. *Genesis Publishing Pvt Ltd* (1925).