doi:10.1093/scan/nst092 SCAN (2014) 9, 1143–1149

# An fMRI investigation of the effects of belief in free will on third-party punishment

**Frank Krueger,[1,2] Morris Hoffman,[3,4] Henrik Walter,[5] and Jordan Grafman[6]**

[1]Molecular Neuroscience Department, George Mason University, Fairfax, VA, USA, [2]Department of Psychology, George Mason University, Fairfax, VA, USA, [3]District Judge, Second Judicial District, State of Colorado, Denver, CO, USA, [4]John D. and Catherine T. MacArthur Foundation Research Network on Law and Neuroscience, Nashville, TN, USA,[5]Division of Mind and Brain Research, Department of Psychiatry and Psychotherapy, Charité Universitaetsmedizin Berlin, Germany, and [6]Brain Injury Research Program, Rehabilitation Institute of Chicago, Chicago, IL, USA

**The relationship between belief in free will (BFW) and third-party punishment (TPP) of criminal norm violations has been the subject of great debate among philosophers, criminologists and neuroscientists. We combined a TPP task with functional magnetic resonance imaging to investigate how lay people's BFW might affect their punishment of hypothetical criminal offenses varying in affective content. Our results revealed that people with strong BFW punished more harshly than people with weak BFW, but only in low affective cases, likely driven by a more robust commitment to moral responsibility. This effect was mirrored by a stronger activation in the right temporo-parietal junction, a region presumably involved in attentional selection to salient stimuli and attribution of temporary intentions and beliefs of others. But, for high affective cases, the BFW-based behavioral and neural differences disappeared. Both groups similarly punished high affective cases and showed higher activation in the right insula. The right insula is typically activated during aversive interoceptive-emotional processing for extreme norm violations. Our results demonstrated that the impact of BFW on TPP is context-dependent; perhaps explaining in part why the philosophical debate between free will and determinism is so stubbornly persistent.**

**Keywords:** criminal law; free will; neurolaw; social cognition

## INTRODUCTION

Third-party punishment (TPP) as a means of enforcing cooperation in response to social norm violations is probably an evolved behavior unique to humans (Riedl *et al.*, 2012). It was likely selected because it enabled large-scale and long-term cooperation among genetically unrelated individuals by deterring free-riding and cheating (Bowles and Gintis, 2004; Fehr and Fischbacher, 2004). As a result, large-scale human societies universally expect that criminal behavior will be punished, usually by impartial third-party decision makers (i.e. state-empowered enforcers such as jurors and judges), who will assess moral responsibility and determine the appropriate legal punishment. Legal TTP has been institutionalized in our criminal justice systems. With a few exceptions, in order to be held responsible under criminal law, offenders must have committed their prohibited actions (*actus reus*) with a bad or guilty intent (*mens rea*) and those actions must have caused actual harm (Shen *et al.*, 2011).

Free will is the often unspoken centerpiece of the criminal law, which presumes humans are responsible agents, who are free to choose to comply with social norms or violate them. We punish wrongdoers only because we believe they had the capacity to resist the wrongful act. Attacking notions of free will has been a favorite tactic of defense lawyers through the ages. The famous American defense lawyer Clarence Darrow (1857–1938) saved several murderers from the gallows, including Leopold and Loeb—two wealthy law students who were motivated to murder an innocent victim simply by their desire to commit a perfect crime—by making determinist arguments to jurors and judges, convincing them that in some deep way none of us is truly responsible for our actions. Darrow's defense strategy takes up one of the oldest and most controversial questions of moral philosophy and criminal law: the relationship between free will and moral responsibility. Can society hold a wrongdoer morally responsible, if his or her actions are completely determined?

In this study, we do not address whether free will exists or what its attributes might be—questions that have stimulated great debates among philosophers, criminologists and recently, neuroscientists (Nichols, 2011; Smith, 2011; Walter, 2011). Instead, we focus on the empirical question of whether a belief in free will (BFW) affects the punishment of moral transgressions. Philosophers have described two extremes of beliefs regarding free will: 'Libertarians' have strong BFW and believe that humans have the capacity to resist doing wrong, and therefore, we are morally responsible for our actions. 'Determinists' believe that all events have material antecedent causes that determine what happens next and that free will is a mirage and therefore, at the extreme, that humans are no more morally responsible for the harms they cause than a falling tree.

Several behavioral studies have begun to examine the question of the impact of BFW on some of our moral behaviors. They have shown that BFW decreases antisocial behaviors, including cheating, stealing, aggression and defection (Vohs and Schooler, 2008; Baumeister *et al.*, 2009). However, previous behavioral studies have been inconclusive as to the effects that BFW has on TPP. Some of those studies have found that libertarians punish more harshly than determinists (Viney *et al.*, 1982), some less (Nettler, 1959) and some observed no difference (Viney *et al.*, 1988). We surmised that this behavioral evidence might be mixed because previous researchers failed to consider whether any relationship between BFW and TPP might be influenced by the affective content of the offenses. A growing body of evidence reveals that affect plays an important role in many kinds of moral judgment (Greene *et al.*, 2001; Nichols, 2002). For example, a recent behavioral study shows that when criminal offenses have a high affective content, lay people are more likely to attribute free will and moral responsibility to the wrongdoer, even when they believed that the wrongdoer's behavior was determined (Nichols and Knobe, 2007).

Received 22 January 2013; Revised 10 May 2013; Accepted 4 June 2013
Advance Access publication 24 July 2013

The authors are grateful to E. Wassermann for performing the neurological exams and O. Dal Monte, A. Kumar and K. Knutson for helping to carry out the fMRI study. This work was supported by the Intramural Research Program of the CNS/NINDS/NIH.

Correspondence should be addressed to Frank Krueger, Department of Molecular Neuroscience, George Mason University, 4400 University Drive, Mail Stop 2A1, Fairfax, VA 22030, USA. E-Mail: FKrueger@gmu.edu

In this study we investigated whether, why and how libertarians and determinists punish differently depending on varying affective content, with the aim of identifying the underlying brain regions responsible for these differences. Recent investigations have begun to uncover the neural underpinnings of TPP, including brain regions of the mentalizing network (e.g. medial prefrontal cortex and temporo-parietal junction), salience network (e.g. amygdala, insula) and central-executive network (e.g. dorsolateral prefrontal cortex, posterior parietal cortex), which are involved in determining moral responsibility and assigning appropriate punishment (Buckholtz *et al.*, 2008; Schleim *et al.*, 2011; Buckholtz and Marois, 2012; Yamada *et al.*, 2012). We hypothesized that determining moral responsibility and punishment might vary dramatically depending on the subjective affective reactions caused by the criminal offenses. We predicted that libertarians would punish low affective offense more than determinists, because libertarians have more robust beliefs in moral responsibility. We therefore expected differential activation between libertarians and determinists in regions of the mentalizing network, in particular, the medial prefrontal cortex and/or the right temporo-parietal junction (R TPJ), regions known to be involved in inferring social mental states (intention, beliefs) of others (Van Overwalle and Baetens, 2009; Young *et al.*, 2010). We also hypothesized that this difference in punishment behaviors between libertarians and determinists would disappear for high affective offenses because of a more prominent influence of affect on decision making. We predicted a differential activation in the amygdala and/or the anterior insula, regions known to be associated with aversive emotional processing (Moll *et al.*, 2007; Buckholtz *et al.*, 2008), when comparing activations for the punishment of high *vs* low affective offenses, regardless of the subjects' BFW.

To test our hypotheses, we combined functional magnetic resonance imaging (fMRI) with a TPP task, asking healthy subjects to estimate how much punishment a hypothetical offender deserved for a set of prototypical offenses ranging across severity of crime from property destruction and theft to rape and murder. An example of these well-studied vignettes was: 'John knows the address of a woman who has highly offended him. As he had planned the day before, he waits there for the woman to return from work and, when she appears, John shoots her to death'. (Robinson and Kurzban, 2007; Krueger *et al.*, 2012). After scanning, criminal scenarios were divided into low and high affective offenses based on subjects' affective experiences elicited by the criminal vignettes. Further, subjects were assigned into two matched groups, libertarians or determinists, based on a validated psychological instrument measuring belief in free will and scientific determinism (Paulhus and Carey, 2011). Our results indicate that when punishing low affective criminal offenses, subjects with strong BFW punished more than subjects with weak BFW and showed greater activation in the R TPJ, a region probably involved in attentional selection and attribution of intentions and beliefs of others. However, this effect disappeared in high affective cases, mirrored in an activation of the right insula, a region presumably involved in aversive interoceptive-emotional processing.

## METHODS

### Subjects

We recruited 26 normal healthy volunteers (13 females, 13 males, age in years: $26.0 \pm 5.7$, years of education: $16.9 \pm 2.6$) for the fMRI study. Subjects were native English speakers and right-handed as determined from the Edinburgh Handedness Inventory (Oldfield, 1971) (mean $\pm$ s.d.: $93.0 \pm 10.4$), and had normal or corrected-to-normal vision. They underwent a neurological examination by a board-certified neurologist during the previous 12 months, had no history of medical, psychiatric or neurological diagnoses and were not taking medication. Based on a brief psychological survey, subjects were excluded from our studies if they met one of the following two criteria: (i) experienced any trauma that involved either injury or threat of injury to themselves or a close family or friend member or (ii) were the victim of, or witnessed, a violent crime. All subjects participated for financial compensation, understood the study procedures and gave written informed consent approved by the Institutional Review Board at the National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA.

### Stimuli and experimental task

Stimuli consisted of criminal vignettes ($n = 44$); each describing an event during which a hypothetical offender named 'John' engages intentionally in criminal offenses (Robinson and Kurzban, 2007) (Supplementary Table S1). The vignettes represent the typical crimes committed in the USA, including theft by taking, theft by fraud, property destruction, assault, burglary, robbery, kidnapping, rape and murder.

The experiment consisted of two judgment tasks: for the Experimental condition, subjects were asked to estimate the punishment that John deserved for each vignette on a Likert scale (no punishment: 0 to extreme punishment: 100). For the Control condition, subjects were asked to estimate (but not to count) the number of syllables for each vignette on a Likert scale ($\sim$30 to $\sim$95). During the fMRI experiment, each trial started with a fixation cross (+; 0.5 s) followed by a vignette (i.e. header and scenario) in the middle of the screen (12 s). Afterwards, the task (Legal Task or Syllables Task) and the Likert scale were presented und the marker was always placed in the middle of the Likert scale for each trial. Using their response pads, subjects gave their answers by pressing the index (to move left) or middle finger (to move right) of their right hand until the marker reached the desired value and then by pressing the index finger of their left hand to mark their final decision (time limit: 6 s). A blank screen with two fixation crosses (++) was displayed for a jittered interstimulus interval (mean 4 s, range: 2–6 s). Before entering the scanner, subjects were familiarized with the task using a separate set of stimuli (to anchor the punishment scale). During each of the three experimental runs, subjects had to respond as quickly and accurately as possible and response times and punishment ratings were recorded for each of the randomly assigned trials. Note that each vignette was presented twice during the fMRI experiment (once in the Experimental condition and once in the Control condition) but never within the same run. The experiment lasted for $\sim$2 h (1 h for scanning, 1 h for the post-scan questionnaires).

Immediately after scanning, subjects completed several psychological surveys in a randomized order. First, subjects completed the Free Will and Determinism (FAD–Plus) questionnaire and were asked to rate their BFW, scientific determinism and closely related constructs (fatalistic determinism and unpredictability) (5-point Likert scale: strongly disagree: 1 to strongly agree: 5) (Paulhus and Carey, 2011). Second, subjects were asked to list what kinds of punishment they were imagining for punishment scores of 1, 25, 50, 75 and 100 to examine the internal scale of punishment applied during the experiment. Third, subjects were asked to rate their subjective affective experience elicited by each of the legal vignettes using a 5-point rating scale (Valence: 1, positive to negative, 5; Arousal: low, 1 to high, 5) version of the Self-Assessment Manikin (SAM) (Lang *et al.*, 1993). Fourth, subjects completed the Interpersonal Reactivity Index (IRI) as a control measure to ensure that predicted group effects were not driven by group differences in empathy (subscales: perspective taking, empathic concern, personal distress and fantasy; 5-point Likert scale: strongly disagree, 1 to strongly agree, 5) (Davis, 1983). Finally, subjects completed the

Toronto Alexithymia Scale (TAS-20) as a control measure to assure that the hypothesized group effects were not due to group differences in identifying and describing emotions (subscales: difficulty identifying feelings, difficulty describing feelings and externally oriented thinking; 5-point Likert scale: strongly disagree, 1 to strongly agree, 5) (Bagby *et al.*, 1994).

### Data acquisition

Neuroimaging was performed on a 1.5 Tesla GE MRI scanner (General Electric, Milwaukee, WI, USA) equipped with an eight-channel receiver head coil located at the NMR Research Center at the National Institutes of Health, Bethesda, MD, USA. Anatomical images (T1-weighted 3D MP-RAGE sequence: time of repetition (TR), 8.9 ms; flip angle, 12°; number of slices, 124; field of view (FOV), 240 mm; matrix size, $256 \times 256$; voxel size, $1 \times 1 \times 1\,mm^3$) and functional images (2D gradient echo planar imaging sequence: TR, 2000 ms; TE, 28 ms; flip angle, 90°; thickness, 3.5 mm; number of slices, 31; FOV, 240 mm; matrix size, $64 \times 64$) were acquired. Functional images were taken parallel to the anterior commissure–posterior commissure line, where the first five volumes were discarded to allow for T1 equilibration effects.

### Behavioral data analysis

The behavioral data analyses were carried out using SPSS 19.0 (SPSS Inc., Chicago, USA) with $\alpha$ set to $P < 0.05$ (two-tailed). Data were normally distributed (Kolmogorov–Smirnov test) and assumptions for analysis of variance (ANOVA; Bartlett's test) were not violated. First, we averaged subjects' post-scanning affective ratings (Valence and Arousal) regarding criminal scenarios from the SAM instrument (Lang *et al.*, 1993) and divided them based on a median split approach into low and high affective offenses. Second, we assigned subjects into two groups—one group believing more in free will ('libertarians') and another group believing more in scientific determinism ('determinists')—applying a median split on the difference score between subjects' BFW and scientific determinism from the FAD–Plus questionnaire (Paulhus and Carey, 2011) (Supplementary Table S2). Third, we ran independent samples *t*-tests to determine whether libertarians and determinists were matched on other types of beliefs (fatalistic determinism and unpredictability), psychological control measures (empathy and emotional awareness) and demographics (age, education and handedness). Fourth, we computed bivariate Pearson correlation coefficients between subjects' punishment ratings and post-scanning affective ratings (Valence and Arousal) to investigate the relationship between subjects' TPP and their negative affective reactions toward criminal scenarios. Finally, we ran mixed $2 \times 2$ ANOVAs on Experimental measures (punishment ratings and response times) and SAM ratings (Valence and Arousal) with Affect (low and high) as a within-subjects factor and Group (libertarians and determinists) as a between-subjects factor and performed planned follow-up independent sample *t*-tests to identify group effects of BFW on TPP.

### fMRI data analysis

The fMRI data analyses were performed using BrainVoyager QX 2.0 (Brain Innovation, Maastricht, The Netherlands). Preprocessing of the functional data included slice-scan time correction (sinc interpolation), small head movements correction by spatially aligning all volumes to the first volume (rigid body transformation), removal of linear trends and low frequency nonlinear drifts of three or fewer cycles for the time series (temporal high-pass filtering) and spatial smoothing of the functional images [Gaussian filter of 8 mm full width at half maximum]. Preprocessing of the anatomical data included reassembling into 1 mm resolution and normalizing into Talairach space using a piecewise linear transformation. Functional data were co-registered with the individual's 3D anatomical images and then reassembled into $3 \times 3 \times 3\,mm^3$ isotropic voxels.

A general linear model (GLM) corrected for first-order serial correlation was applied. Random effect analyses were performed on the multisubject level to explore brain regions that were associated with the decision phase of the Experimental condition. The GLM consisted of a set of 12 regressors: two categorical regressors for the main scenario reading phase modulation (Experimental and Control), four categorical regressors for the main decision phase modulation based on the affective levels of scenarios (Experimental: Low, E_L; High, E_H; Control: Low, C_L; High, C_H) and six parametric regressors of no interest for the 3D motion correction (translation in X, Y, Z direction and rotation around X, Y, Z axis). The regressor time courses were adjusted for the hemodynamic response delay by convolution with a dual-gamma hemodynamic response function (Buchel *et al.*, 1998).

After computing the coefficients ($\beta$ parameters) for all regressors, one statistical model was fit on the multisubject level for the decision phase in the experiment. To reveal brain activations associated with the effects of BFW on TPP, a mixed $2 \times 2$ ANOVA on $\beta$ parameters was applied with Affect (E_L and E_H) as a within-subjects factor and Group (libertarian and determinist) as a between-subjects factor. Activations were reported using a threshold of $q(\text{FDR}) < 0.05$ by applying a whole brain analysis approach (Genovese *et al.*, 2002). For display purposes, statistical images were overlaid onto the mean anatomical image from the group of subjects in Talairach space and were reversed left to right according to radiological convention. Brodmann areas were determined by using the Talairach Daemon Client software (Research Imaging Center, San Antonio, TX, USA) and the coplanar stereotaxic atlas of the human brain (Talairach and Tournoux, 1988).

## RESULTS

### Behavioral results

To confirm that subjects were following the instructions defining the 100-point punishment scale (no punishment: 0 to extreme punishment: 100), subjects were asked to list what kind of punishment they imagined during the experiment for selected punishment ratings after scanning. They demonstrated a strong agreement about their internal scale of justice: financial or social penalties justified low punishment ratings (1, 25), higher punishment ratings were associated with longer jail times (50, 75) and life imprisonment or death penalty led to the highest punishment ratings (100) (Supplementary Table S3). The TPP ratings were significantly positively correlated with subjects' post-scanning affective ratings (Valence: $r = 0.89$, $P < 0.001$; Arousal: $r = 0.92$, $P < 0.001$), indicating that subjects' degree of punishment was strongly associated with their degree of self-reported negative affective reactions toward criminal scenarios.

Next, criminal scenarios were divided into two groups based on a median split (mean ± s.d., $3.39 \pm 0.70$) for subjects' averaged post-scanning affective ratings (Valence and Arousal) using the SAM instrument (Lang *et al.*, 1993) (Supplementary Figure S1): low affective offenses ($2.83 \pm 0.35$) and high affective offenses ($3.95 \pm 0.49$) [$t(42) = 8.65$, $P < 0.0001$]. Moreover, subjects were assigned into two BFW groups based on a median split of the difference score (mean ± s.d., $0.81 \pm 0.66$) between subjects' ratings of BFW and scientific determinism using the FAD–Plus questionnaire (Paulhus and Carey, 2011): one group believing more in free will ('libertarians') ($1.31 \pm 0.42$) and another group believing more in scientific determinism ('determinists') ($0.31 \pm 0.44$) [$t(24) = -5.94$, $P < 0.0001$]. Groups were matched on other types of beliefs (fatalistic determinism and unpredictability), psychological control measures (empathy and

**Table 1** Descriptive (mean ± s.d.) and inferential statistics for demographic and psychological control measures collected for subjects classified as either determinists or libertarians

| Category | Determinists | Libertarians | Statistics |
|---|---|---|---|
| Demographics | | | |
| Age | 24.31 ± 3.07 | 27.69 ± 7.19 | $t(24) = -1.56$, $P = 0.132$ |
| Education | 17.08 ± 2.10 | 16.69 ± 3.15 | $t(24) = 0.37$, $P = 0.717$ |
| Handedness | 92.00 ± 11.29 | 93.00 ± 10.02 | $t(24) = -0.24$, $P = 0.813$ |
| Gender (male/female) | 6/7 | 7/6 | $\chi^2(1) = 1.54$, $P = 0.695$ |
| Belief ratings (free will and determinism questionnaire) | | | |
| Free will | 3.42 ± 0.69 | 3.98 ± 0.42 | $t(24) = -2.49$, $P < 0.020$ |
| Scientific determinism | 3.10 ± 0.53 | 2.65 ± 0.41 | $t(24) = 2.39$, $P < 0.024$ |
| Fatalistic determinism | 1.78 ± 0.41 | 1.98 ± 0.56 | $t(24) = -1.03$, $P = 0.311$ |
| Unpredictability | 3.01 ± 0.54 | 2.82 ± 0.40 | $t(24) = 1.04$, $P = 0.311$ |
| Emotional awareness (TAS-20) | | | |
| Difficulty identifying feelings | 11.46 ± 5.28 | 9.77 ± 2.85 | $t(24) = 1.00$, $P = 0.323$ |
| Difficulty describing feelings | 11.62 ± 4.68 | 9.46 ± 3.38 | $t(24) = 1.34$, $P = 0.191$ |
| Externally oriented thinking | 18.77 ± 5.37 | 17.46 ± 4.08 | $t(24) = 0.69$, $P = 0.491$ |
| Empathy (IRI) | | | |
| Perspective taking | 25.45 ± 4.46 | 25.62 ± 4.33 | $t(24) = -0.09$, $P = 0.930$ |
| Fantasy scale | 24.00 ± 4.49 | 22.69 ± 5.17 | $t(24) = 0.69$, $P = 0.498$ |
| Empathic concern | 26.15 ± 4.72 | 26.08 ± 4.41 | $t(24) = 0.04$, $P = 0.966$ |
| Personal distress | 16.08 ± 4.73 | 14.54 ± 3.33 | $t(24) = 0.96$, $P = 0.348$ |
| Response times (punishment) | | | |
| Offenses (low affect) | 3745 ± 578 | 3635 ± 552 | $t(24) = 0.03$, $P = 0.973$ |
| Offenses (high affect) | 3753 ± 610 | 3670 ± 517 | $t(24) = 0.17$, $P = 0.869$ |

emotional awareness) and demographics (age, gender, education and handedness) (Table 1).

To test our first prediction, libertarians, compared with determinists would only punish more harshly for low affective offenses, we ran mixed 2 Affect (low and high) × 2 Group (libertarians and determinists) ANOVAs on punishment ratings and response times recorded during the fMRI experiment. The ANOVA on punishment ratings revealed no significant main effect of Group [$F(1, 24) = 4.18$, $P = 0.054$], but a significant main effect of Affect [$F(1, 24) = 394.24$, $P < 0.0001$] and a significant interaction effect of Affect × Group [$F(1, 24) = 7.47$, $P < 0.012$]. Planned follow-up independent samples *t*-tests revealed that libertarians punished more than determinists for low affective offenses [$t(24) = -2.52$, $P < 0.019$] but not high affective offenses [$t(24) = -1.04$, $P = 0.310$] (Figure 1). In contrast, the ANOVA on response times demonstrated no significant main effects of Affect [$F(1, 24) = 1.38$, $P = 0.252$] and Group [$F(1, 24) = 0.01$, $P = 0.918$] and no significant interaction effect of Affect × Group [$F(1, 24) = 0.03$, $P = 0.871$] (Table 1).

To test our second prediction, differences in punishment behaviors in both groups would disappear for high affective offenses, we ran mixed 2 Affect (low and high) × 2 Group (libertarians and determinists) ANOVAs on SAM ratings (Valence and Arousal). The ANOVAs demonstrated expected significant main effects of Affect [Valence: $F(1, 24) = 224.42$, $P < 0.0001$; Arousal: $F(1, 24) = 107.82$, $P < 0.0001$], but no significant main effects of Group [Valence: $F(1, 24) = 0.01$, $P = 0.931$; Arousal: $F(1, 24) = 0.18$, $P = 0.676$] and no significant interaction effects of Affect × Group [Valence: $F(1, 24) = 1.26$, $P = 0.273$; Arousal: $F(1, 24) = 1.38$, $P = 0.251$] (Figure 2).

### Neuroimaging results

A GLM analysis was applied and a random effect analysis on $\beta$ parameters was performed for the decision phase of the experiment. We ran a 2 Affect (low and high) × 2 Group (libertarian and determinist) ANOVA on $\beta$ parameters to identify those brain regions whose blood oxygen level dependent (BOLD) responses were associated with the impact of BFW on TPP. The ANOVA revealed an interaction effect of Affect × Group in the R TPJ (Talairach peak x, y, z: 50, −50, 15) [$F(1, 24) = 22.46$, $P < 0.0001$] (Figure 3). Planned follow-up independent samples *t*-tests showed that libertarians had higher activations in the R TPJ than determinists [$t(24) = -2.86$, $P < 0.009$] for low affective offenses, whereas activations in this region was the same for both groups for high affective offenses [$t(24) = 0.45$, $P = 0.655$]. Moreover, the ANOVA showed no main effect of Group, but revealed a significant main effect of Affect in the right anterior insula (R AI, Talairach peak x, y, z: 30, 14, 12) [$F(1, 24) = 31.10$, $P < 0.001$], indicating that both groups activated the R AI more when punishing high affective compared to low affective offenses (Figure 4).

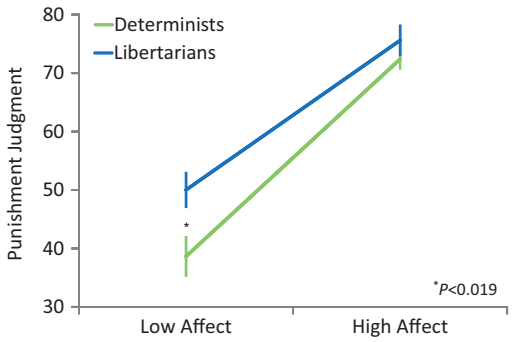


**Fig. 1** Punishment ratings (mean ± s.e.m.). Libertarians punished differently than determinists depending on the affective content of criminal offenses: libertarians punished more than determinists for low affective offenses, whereas both groups punished the same for high affective offenses.

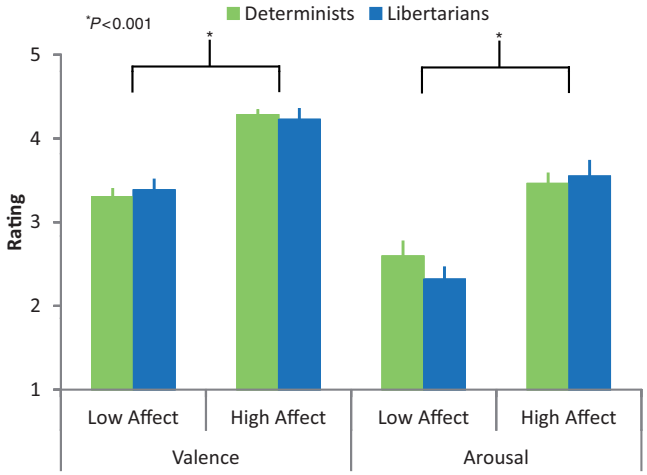


**Fig. 2** Affective ratings (mean ± s.e.m.). Subjects were shown the same criminal scenarios again after scanning to rate their affective experiences using a 5-point rating scale (Valence: 1, positive to negative, 5; Arousal: low, 1 to high, 5) version of the SAM instrument. Libertarians and determinists rated their affective reactions for high affective offenses significantly higher compared with low affective offenses.

## DISCUSSION

In this study, we addressed the question whether libertarians and determinists punish criminal offenses differently as impartial third-party decision makers and investigated the underlying neural signatures reflecting these differences. We predicted that the existing inconsistent evidence regarding the impact of BFW on TPP might be a result of studies failing to distinguish between high and low affective content of criminal offenses. Our findings revealed that libertarians punished low affective cases more harshly than determinists, initiated by a stronger

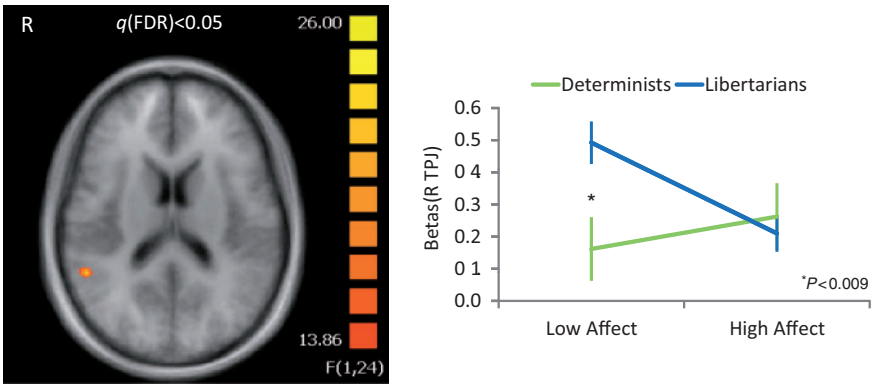


**Fig. 3** Brain activation (mean ± s.e.m.) for belief and punishment. Libertarians and determinists showed differential activations in their R TPJ depending on affective content. Libertarians had higher activations in the R TPJ than determinists for low affective offenses, whereas activation in this region was the same for both groups for high affective offenses. For display purposes, statistical images were overlaid onto the mean anatomical image from the group of subjects in Talairach space and were reversed left to right according to radiologic convention.

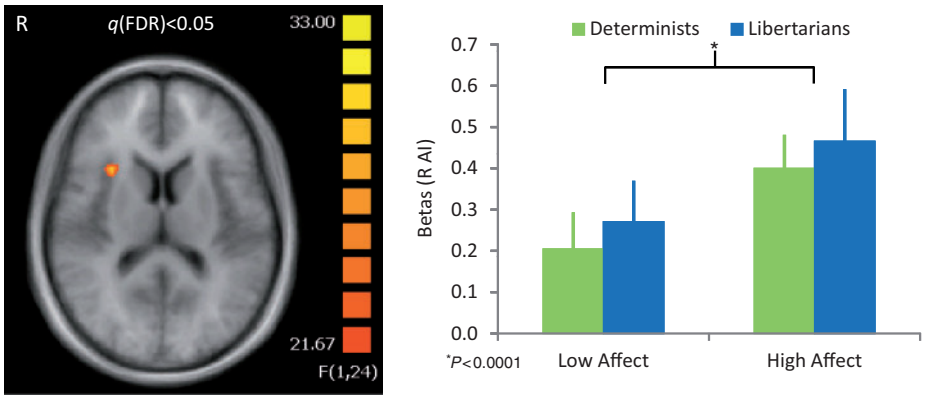


**Fig. 4** Brain activation (mean ± s.e.m.) for affect and punishment. Punishing high affective offenses was associated with significantly higher activation in R AI compared with punishing low affective offenses for both libertarians and determinists. For display purposes, statistical images were overlaid onto the mean anatomical image from the group of subjects in Talairach space and were reversed left to right according to radiological convention.

activation in the R TPJ, a region probably involved in attentional selection to salient stimuli and attribution of temporary intentions and beliefs of others. This effect disappeared for high affective cases, presumably due to activation of the R AI, a region of the salience network probably involved in aversive interoceptive-emotional processing.

Although subjects' hypothetical legal decisions had no direct, real-world consequences for real criminal offenders, our post-scan debriefing results demonstrated that their punishment assessments were a good proxy measure for real-world legal criminal judgments. Subjects applied an implicit legal metric: associating lower punishment judgments with shorter jail times (1, 25), higher punishment judgments with longer jail times (50, 75) and highest punishment judgments with life imprisonment or the death penalty (100). Moreover, subject's punishment increased as a function of their self-reported affective reactions toward criminal scenarios, replicating previous evidence showing that sanctions are driven by affective reactions toward norm violations (Darley and Pittman, 2003; Fehr and Fischbacher, 2004).

As hypothesized, we found that libertarians punished low affective offenses significantly more harshly (an average of 10% more) than determinists and had a stronger activation in the R TPJ, but no difference between groups were observed when punishing high affective offenses. The R TPJ, at the border between the superior temporal and angular gyrus, is linked to a number of higher order cognitive functions, related to attentional selection such as reorienting attention to salient stimuli (Mesulam, 1981; Mitchell, 2008) and social cognition such as attribution of mental states (intentions, beliefs) of self or others (Van Overwalle and Baetens, 2009; Young and Saxe, 2009). This line of evidence suggests that determinists punished less because they dismissed the (bad) intention, because they have doubts about the extent to which any of us are truly in control of our intentional behaviors. In contrast, libertarians arguably punished more harshly because they focused more on the negative intention involved in the offenses, driving a more robust commitment to moral responsibility, and therefore leading to a stronger activation of the R TPJ. Indeed, a previous study has shown that disruption of the R TPJ with transcranial magnetic stimulation caused subjects to judge attempted harms as less morally forbidden and more morally permissible (Young *et al.*, 2010), providing additional evidence that interfering with activity in the R TPJ disrupts the capacity to use mental states in determining responsibility.

However, as hypothesized, the observed behavioral and neural differences between determinists and libertarians disappeared for high affective scenarios. When subjects of both groups punished high affective as compared to low affective offenses, they showed a differential activation in the right anterior insula. The anterior insula is part of the salience network (Seeley *et al.*, 2007; Bressler and Menon, 2010), which is involved in the orientation of attention to the most homestatically relevant (salient) of ongoing interpersonal and extrapersonal events (Kelly *et al.*, 2008; Eckert *et al.*, 2009). Recent evidence supports the

view that the anterior insula is linked to interoceptive awareness of body states as well as emotional processing via representations of signals of (especially aversive) internal states (Phan *et al.*, 2002; Craig, 2003). For example, greater right anterior insular gray matter volume is associated with increased accuracy in the subjective sense of the inner body and with negative emotional experience (Critchley *et al.*, 2004). Given that the anterior insula is also sensitive to emotions linked to sociality such as moral disgust toward violation of social and moral norms (Sanfey *et al.*, 2003; Greene *et al.*, 2004; Chapman and Anderson, 2012), the engagement of this region in aversive interoceptive-emotional processing for high affective scenarios presumably muted the contributions of the R TPJ. A recent study demonstrated that the R AI and R TPJ are structurally and functionally connected to each other (Mars *et al.*, 2012), belonging to a common ventral attentional network that mediates reorienting of attention in response to behaviorally relevant events (Corbetta and Shulman, 2002). This would potentially allow either structure, given the right context, to dominate processing. Taken together, our imaging and behavioral results suggest that people's negative affective reaction toward extreme norm violations causes determinists to act like libertarians when punishing high harms.

There are some limitations to our study that deserve discussion. First, we divided the criminal scenarios into low and high affective offenses based on a recent finding, showing that lay people are more likely to attribute free will and moral responsibility to the wrongdoer when criminal offenses have a high affective content, even when they believed that the wrongdoer's behavior was determined (Nichols and Knobe, 2007). Although in our study, subjects' assessment of punishment were strongly associated with their degree of self-reported negative affective reactions toward criminal scenarios, other differences between these low and high affective offenses (i.e. other than affect) might have generated the observed differences in behavior and brain activity. Whereas our criminal scenarios fulfilled the criminal law's central tenet of punishment, i.e. a hypothetical offender committed with a guilty intent a set of prototypical offenses ranging across severity of crime; however, future studies should investigate other potential factors that might modify the effects of BFW on TPP, such as types of harm to the victims (e.g. physical, financial, etc.) and types of mental states of the defendant (e.g. purposeful, knowing, reckless, negligent, etc.) at the time the crime was committed (Shen *et al.*, 2011).

Second, we divided subjects into determinists and libertarians based on the FAD–Plus questionnaire measuring different BFW (free will, scientific determinism, fatalistic determinism and unpredictability) (Paulhus and Carey, 2011). We created these two extreme groups by applying a median split on the difference score between subjects' belief in free will and scientific determinism. Although group differences were observed in both behavior and brain activity and the groups were matched on related beliefs (fatalistic determinism, unpredictability), future studies should broaden our findings by utilizing groups that differ on each of the FAD–Plus subscales. Moreover, future studies should investigate how manipulation of subjects' BFW, either by encouraging to shift beliefs toward greater free will or persuading to reject free will, modifies the observed patterns in punishment behaviors and brain activities.

In conclusion, our study demonstrated that the impact of BFW on TPP is context-dependent. Our findings of the importance of context helps to resolve why there is a mixed literature on the effects of BFW on TPP, and suggests that without controlling for contextual variables such as affective content of the criminal offense, the philosophical debate between BFW and determinism will remain obstinately persistent. Our results may also lead to some practical legal applications. In low affective offenses, BFW might influence jurors' blaming decisions in ways to which lawyers, judges and the law as a whole might need to attend. In high affective offenses, judges and lawyers should be less concerned about jurors' BFW and more concerned about the emotional impact of the crime.

## SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

### Conflict of Interest

None declared.

## REFERENCES


Bagby, R.M., Taylor, G.J., Parker, J.D. (1994). The Twenty-item Toronto Alexithymia Scale–II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, *38*, 33–40.

Baumeister, R.F., Masicampo, E.J., Dewall, C.N. (2009). Prosocial benefits of feeling free: disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, *35*, 260–8.

Bowles, S., Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*, 17–28.

Bressler, S.L., Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences*, *14*, 277–90.

Buchel, C., Holmes, A.P., Rees, G., Friston, K.J. (1998). Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage*, *8*, 140–148.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., et al. (2008). The neural correlates of third-party punishment. *Neuron*, *60*, 930–40.

Buckholtz, J.W., Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*, 655–61.

Chapman, H.A., Anderson, A.K. (2012). Understanding disgust. *Annals of the New York Acad of Sciences*, *1251*, 62–76.

Corbetta, M., Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–15.

Craig, A.D. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*, 500–5.

Critchley, H.D., Wiens, S., Rotshtein, P., Ohman, A., Dolan, R.J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*, 189–95.

Darley, J.M., Pittman, T.S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*, 324–36.

Davis, M. (1983). Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–26.

Eckert, M.A., Menon, V., Walczak, A., et al. (2009). At the heart of the ventral attention system: the right anterior insula. *Human Brain Mapping*, *30*, 2530–41.

Fehr, E., Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*, 185–90.

Genovese, C.R., Lazar, N.A., Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*, 870–8.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

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.

Kelly, A.M., Uddin, L.Q., Biswal, B.B., Castellanos, F.X., Milham, M.P. (2008). Competition between functional brain networks mediates behavioral variability. *Neuroimage*, *39*, 527–37.

Krueger, F., Parasuraman, R., Moody, L., et al. (2012). Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social Cognitive and Affective Neuroscience*, *8*(5), 494–8.

Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O. (1993). Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*, 261–73.

Mars, R.B., Sallet, J., Schuffelgen, U., Jbabdi, S., Toni, I., Rushworth, M.F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral Cortex*, *22*, 1894–903.

Mesulam, M.M. (1981). A cortical network for directed attention and unilateral neglect. *Annals of Neurology*, *10*, 309–25.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, *18*, 262–71.

Moll, J., De Oliveira-Souza, R., Garrido, G.J., et al. (2007). The self as a moral agent: linking the neural bases of social agency and moral sensitivity. *Society for Neuroscience*, *2*, 336–52.

Nettler, G. (1959). Cruelty, dignity, and determinism. *American Sociological Review*, *24*, 375–384.

Nichols, S. (2002). Norms with feeling: towards a psychological account of moral judgment. *Cognition*, *84*, 221–36.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, *331*, 1401–3.
Nichols, S., Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Nous*, *41*, 663–85.
Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*, 97–113.
Paulhus, D.L., Carey, J.M. (2011). The FAD-Plus: measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, *93*, 96–104.
Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, *16*, 331–48.
Riedl, K., Jensen, K., Call, J., Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences USA*, *109*, 14824–9.
Robinson, P., Kurzban, R. (2007). Concordance and conflict in intuitions of justice. *Minnesota Law Review*, *91*, 1829–93.
Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–8.
Schleim, S., Spranger, T.M., Erk, S., Walter, H. (2011). From moral to legal judgment: the influence of normative context in lawyers and other academics. *Social Cognitive and Affective Neuroscience*, *6*, 48–57.
Seeley, W.W., Menon, V., Schatzberg, A.F., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*, 2349–56.
Shen, F.X., Hoffman, M.B., Jones, O.D., Greene, J.D., Marois, R. (2011). Sorting guilty minds. *New York University Law Review*, *86*, 1306–60.
Smith, K. (2011). Neuroscience vs philosophy: taking aim at free will. *Nature*, *477*, 23–5.
Talairach, J., Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. New York, NY: Thieme Medical Publishers.
Van Overwalle, F., Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*, 564–84.
Viney, W., Parker-Martin, P., Dotten, S. (1988). Belief in free will and determinism and lack of relation to punishment rationale and magnitude. *Journal of General Psychology*, *115*, 15–23.
Viney, W., Waldman, D.A., Barchilon, J. (1982). Attitudes toward punishment in relation to beliefs in free will and determinism. *Human Relations*, *35*, 939–49.
Vohs, K.D., Schooler, J.W. (2008). The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science*, *19*, 49–54.
Walter, H. (2011). Contributions of neuroscience to the free will debate: from random movement to intelligible action. In: Kane, R., editor. *Oxford Handbook of Free Will*. New York, NY: Oxford University Press, pp. 515–29.
Yamada, M., Camerer, C.F., Fujie, S., et al. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nature Communications*, *3*, 759.
Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences USA*, *107*, 6753–8.
Young, L., Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*, 1396–405.