Neuron
Article

The Neural Correlates of Third-Party Punishment

Joshua W. Buckholtz,[1,2] Christopher L. Asplund,[1,2] Paul E. Dux,[1] David H. Zald,[1,5] John C. Gore,[3,5,8,9] Owen D. Jones,[5,6,7,*] and René Marois[1,4,5,*]
[1]Department of Psychology
[2]Neuroscience Graduate Program
[3]Institute of Imaging Science
[4]Vision Research Center
[5]Center for Integrative and Cognitive Neurosciences
[6]Law School
[7]Department of Biological Sciences
[8]Departments of Radiology and Radiological Sciences
[9]Department of Biomedical Engineering
Vanderbilt University, Nashville, TN 37240, USA
*Correspondence: owen.jones@vanderbilt.edu (O.D.J.), rene.marois@vanderbilt.edu (R.M.)
DOI 10.1016/j.neuron.2008.10.016

SUMMARY

Legal decision-making in criminal contexts includes two essential functions performed by impartial "third parties:" assessing responsibility and determining an appropriate punishment. To explore the neural underpinnings of these processes, we scanned subjects with fMRI while they determined the appropriate punishment for crimes that varied in perpetrator responsibility and crime severity. Activity within regions linked to affective processing (amygdala, medial prefrontal and posterior cingulate cortex) predicted punishment magnitude for a range of criminal scenarios. By contrast, activity in right dorsolateral prefrontal cortex distinguished between scenarios on the basis of criminal responsibility, suggesting that it plays a key role in third-party punishment. The same prefrontal region has previously been shown to be involved in punishing unfair economic behavior in two-party interactions, raising the possibility that the cognitive processes supporting third-party legal decision-making and second-party economic norm enforcement may be supported by a common neural mechanism in human prefrontal cortex.

INTRODUCTION

Though rare in the rest of the animal kingdom, large-scale cooperation among genetically unrelated individuals is the rule, rather than the exception, in *Homo sapiens* (Henrich, 2003). Ultrasociality and cooperation in humans is made possible by our ability to establish social norms—widely shared sentiments about appropriate behaviors that foster both social peace and economic prosperity (Fehr and Fischbacher, 2004a; Spitzer et al., 2007). In turn, norm compliance relies not only on the economic self-interest often served by cooperation and fair exchange, but also on the credible threat of unwelcome consequences for defection (Spitzer et al., 2007). Social order therefore depends on punishment, which modern societies administer through a system of state-empowered enforcers, guided by state-governed, impartial, third-party decision-makers, who are not directly affected by the norm violation and have no personal stake in its enforcement.

The role of legal decision-makers is twofold: determining responsibility and assigning an appropriate punishment. In determining responsibility, a legal decision-maker must assess whether the accused has committed a wrongful act and, if so, whether he did it with one of several culpable states of mind (so-called "mens rea") (Robinson, 2002). For many of the most recognizable crimes, the defendant must have engaged in the proscribed conduct with intent in order to merit punishment. Moreover, in sentencing an individual for whom criminal responsibility has been determined, a legal decision-maker must choose a punishment that fits the crime. This sentence must ordinarily be such that the combined nature and extent of punishment is proportional to the combined harmfulness of the offense and blameworthiness of the offender (Farahany and Coleman, 2006; LaFave, 2003).

Despite its critical utility in facilitating prosocial behavior and maintaining social order, little is known about the origins of, and neural mechanisms underlying, our ability to make third-party legal decisions (Garland, 2004; Garland and Glimcher, 2006; Zeki and Goodenough, 2004). The cognitive ability to make social norm-related judgments likely arose from the demands of social living faced by our hominid ancestors (Henrich, 2003; Richerson et al., 2003). These demands may have promoted the emergence of mechanisms for assessing fairness in interpersonal exchanges and enacting personal retaliations against individuals who behaved unfairly (second-party punishment) (Fehr and Fischbacher, 2004a). Recent work has greatly advanced our understanding of how the brain evaluates fairness and makes decisions based on the cooperative status and intentions of others during two-party economic exchanges (de Quervain et al., 2004; Delgado et al., 2005; King-Casas et al., 2005; Knoch et al., 2006; Sanfey et al., 2003; Singer et al., 2004, 2006; Spitzer et al., 2007). Notably, these studies have elucidated the neural dynamics that underlie human altruistic punishment, in which the victim of a social norm transgression, typically unfairness in an economic

exchange, punishes the transgressor at some significant additional cost to himself. These findings have specifically highlighted the importance of reward and emotion-related processes in fueling cooperative behavior (Seymour et al., 2007). However, how—or even whether—neural models of economic exchange in dyadic interactions apply to impartial, third-party legal decision-making is currently unknown (Fehr and Fischbacher, 2004a). Furthermore, the importance of uncovering neural mechanisms underlying third-party punishment is underscored by the proposal that the development of stable social norms in human societies specifically required the evolution of third-party sanction systems (Bendor and Swistak, 2001).

Given that, in great measure, criminal law strives toward the stabilization and codification of social norms, including moral norms, in legal rules of conduct (Robinson and Darley, 1995), moral decision-making is inherently embedded into the legal decision-making process. The relevance of moral decision-making to an investigation of legal reasoning is highlighted by experimental findings which suggest that individuals punish according to so-called "just deserts" motives; i.e., in proportion to the moral wrongfulness of an offender's actions (Alter et al., 2007; Carlsmith et al., 2002; Darley and Pittman, 2003). As such, the seminal work of Greene and others—which has demonstrated distinct contributions of emotion-related and cognitive control-related brain regions to moral decision-making (Greene et al., 2001, 2004; Heekeren et al., 2003, 2005; Moll et al., 2002a, 2002b)—is germane to the study of legal decision-making. However, despite the conceptual overlap between moral and legal reasoning, the latter process is not entirely reducible to the former (Hart, 1958; Holmes, 1991; Posner, 1998; Robinson, 1997; Robinson and Darley, 1995). Indeed, whereas determining blameworthiness may in many cases fall under the rubric of moral decision-making, the distinctive core and distinguishing feature of legal decision-making is the computation and implementation of a punishment that is appropriate both to the relative moral blameworthiness of an accused criminal offender, and to the relative severity of that criminal offense (Robinson, 1997; Robinson and Darley, 1995). The present study is focused on elucidating the neural mechanisms underlying this third-party, legal decision-making process.

In this study, we used event-related fMRI to reveal the neural circuitry supporting third-party decision-making about criminal responsibility and punishment. Given that these two legally distinct judgments are rendered on the basis of differing information and considerations (LaFave et al., 2007), we were particularly interested in determining whether these two decision-making processes may rely on at least partly distinct neural systems. To address this issue, we scanned 16 participants while they determined the appropriate punishment for actions committed by the protagonist (named "John") in a series of 50 written scenarios. Each of these scenarios belonged to one of three categories: Responsibility (R), Diminished-Responsibility (DR), and No-Crime (NC). Scenarios in the Responsibility set (n = 20) described John intentionally committing a criminal action ranging from simple theft to rape and murder. The Diminished-Responsibility set (n = 20) included actions of comparable gravity to those described in the Responsibility set but also contained mitigating circumstances that may have excused or justified the otherwise criminal behavior of the protagonist by calling his blameworthiness into question. The No-Crime set (n = 10) depicted John engaged in noncriminal actions that were otherwise structured similarly to the Responsibility and Diminished-Responsibility scenarios (scenarios available as Supplemental Experimental Procedures). Participants rated each scenario on a scale from 0–9, according to how much punishment they thought John deserved, with "0" indicating no punishment and "9" indicating extreme punishment. Two groups of 50 scenarios (equated for word length between conditions and between groups) were constructed and their presentation counterbalanced across the 16 participants. The Responsibility set of group 2 consisted of group 1 Diminished-Responsibility scenarios for which the mitigating circumstances had been removed, while the Diminished-Responsibility set of group 2 consisted of group 1 Responsibility scenarios with mitigating circumstances added. Thus, each criminal scenario (e.g., depicting theft, assault or murder) in the Responsibility and Diminished-Responsibility condition was created by modifying identical "stem" stories, with salient details such as magnitude of harm matched between conditions.

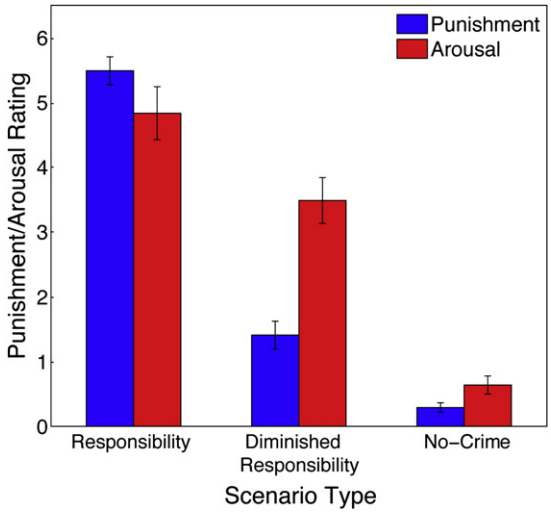

Figure 1. Punishment and Arousal Ratings for Each Scenario Type
While punishment and arousal scores were similar in the Responsibility condition, punishment scores were significantly lower than arousal scores in the Diminished-Responsibility condition. Error bars = SEM.

RESULTS

Behavioral Data

Behavioral data showed a significant effect of scenario category on punishment ratings [F (1,15) = 358.61, $p < 0.001$] (Figure 1), with higher mean ratings for the Responsibility (mean = 5.50, SE = 0.22) as compared with the Diminished-Responsibility scenarios (Mean = 1.45, SE = 0.21) ($p < 0.001$, paired t test), indicating that assessed punishment was strongly modulated by the protagonist's criminal responsibility. However, the fact that the mean punishment rating for the Diminished-Responsibility condition was greater than 0 suggests that some participants still attributed some blameworthiness to the protagonist despite the

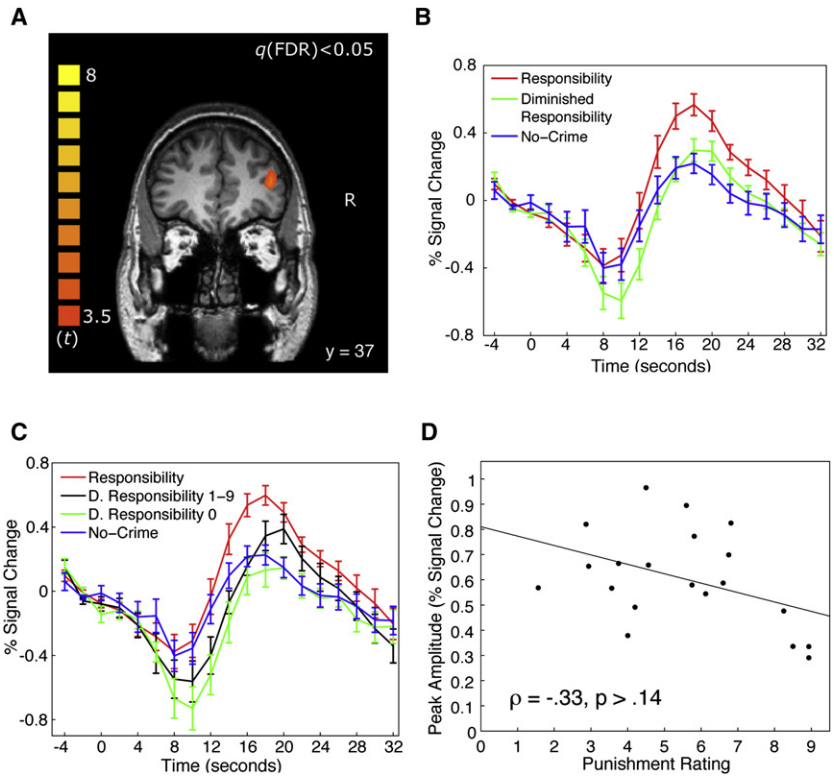

Figure 2. Relationship between Responsibility Assessment and rDLPFC Activity

(A) SPM displaying the rDLPFC VOI (rendered on a single-subject T1-weighted image), based on the contrast of BOLD activity between the Responsibility and Diminished-Responsibility conditions. $t(15) > 3.5$, $q < 0.05$, random effects analysis. R = Right Hemisphere.

(B) BOLD activity time courses in rDLPFC for the Responsibility, Diminished-Responsibility, and No-Crime conditions. BOLD peak amplitude was significantly greater in the Responsibility condition compared with both the Diminished-Responsibility and No-Crime conditions ($p = 0.002$, $p = 0.0004$, respectively). Peak was defined as the single TR with maximal signal change from baseline within the first 13 volumes after scenario presentation onset. t tests were performed on these peak volumes, which were defined separately for each condition and each subject.

(C) BOLD activity time courses in rDLPFC for Responsibility, "nonpunished" Diminished-Responsibility (Diminished-Responsibility 0), "punished" Responsibility (Diminished-Responsibility 1–9), and No-Crime scenarios. BOLD peak amplitude was significantly greater in punished compared with nonpunished Diminished-Responsibility scenarios ($p = 0.04$), while no difference was observed between nonpunished Diminished-Responsibility and No-Crime scenarios ($p = 0.98$).

(D) Relationship between BOLD peak amplitude in rDLPFC and punishment ratings in the Responsibility condition. These two variables were not significantly correlated ($p > 0.15$).

Error bars = SEM.

extenuating circumstances. To examine the subjective emotional experience elicited by the scenarios, all participants completed postscan ratings of emotional arousal for each scenario. These ratings also demonstrated an effect of condition [$F(1,15) = 94.61$, $p < 0.001$] (Figure 1), with greater mean arousal scores for the Responsibility (Mean = 4.83, SE = 0.41) compared to the Diminished-Responsibility scenarios (Mean = 3.48, SE = 0.35) ($p < 0.001$, paired t test). Additionally, we found a significant interaction between rating type (punishment versus arousal) and condition (Responsibility versus Diminished-Responsibility) [$F(1,15) = 68.8$, $p < 0.001$] such that, while the punishment and arousal ratings were not significantly different for the Responsibility scenarios ($p > 0.05$, paired t test), punishment ratings were significantly lower than the arousal ratings for the Diminished-Responsibility scenarios ($p < 0.001$, paired t test) (Figure 1). Lastly, we found a main effect of scenario condition on reaction times (RTs) [$F(1,15) = 21.87$, $p < 0.001$], such that RTs were shortest for the No-Crime condition and longest for the Diminished-Responsibility condition (mean, SE for: Responsibility = 12.69 s, 0.46; Diminished-Responsibility = 13.76 s, 0.46; No-Crime = 11.12 s, 0.44; respectively) (all paired comparisons $p < 0.01$).

fMRI Data: Criminal Responsibility

To identify brain regions that were sensitive to information about criminal responsibility, we contrasted brain activity between Responsibility and Diminished-Responsibility scenarios. The resulting statistical parametric map (SPM) revealed an area of activation in the right dorsolateral prefrontal cortex (rDLPFC, Brodmann Area 46, peak at Talaraich coordinates 39, 37, 22 [x,y,z]; Figure 2A) that was significantly more activated in the Responsibility as compared with the Diminished-Responsibility condition. Time course analyses of peak activation differences confirmed that there was greater rDLPFC activity in Responsibility compared with Diminished-Responsibility or No-Crime conditions ($R > DR$, $p = 0.002$; $R > NC$, $p = 0.0004$; paired t tests; see Figure 2B) and no difference between the Diminished-Responsibility and No-Crime conditions ($p = 0.19$). No effect of condition was found in the left DLPFC ($p > 0.2$ for all paired comparisons; see Experimental Procedures), and the rDLPFC was significantly more engaged than the left DLPFC in the Responsibility condition ($p = 0.04$, paired t test), suggesting that punishment-related prefrontal activation is confined to the right hemisphere. Bilateral anterior intraparietal sulcus (aIPS) demonstrated a pattern of responsibility-related activity that was similar to rDLPFC (Table S1 and Figure S1 available online, Supplemental Results), whereas the temporo-parietal junction (TPJ) showed the reverse pattern, with more activity in the Diminished-Responsibility as compared with the Responsibility condition (Table S1, Figure 3, see below).

Greater rDLPFC activation in the Responsibility condition did not simply result from longer time on task: RTs to Responsibility scenarios were shorter than those of Diminished-Responsibility scenarios ($p = 0.005$, paired t test), and the effect of condition on rDLPFC activity was still significant when response time was used as a covariate in an analysis of covariance [ANCOVA, $F(1,37) = 10.15$, $p = 0.003$] or when response times were equated

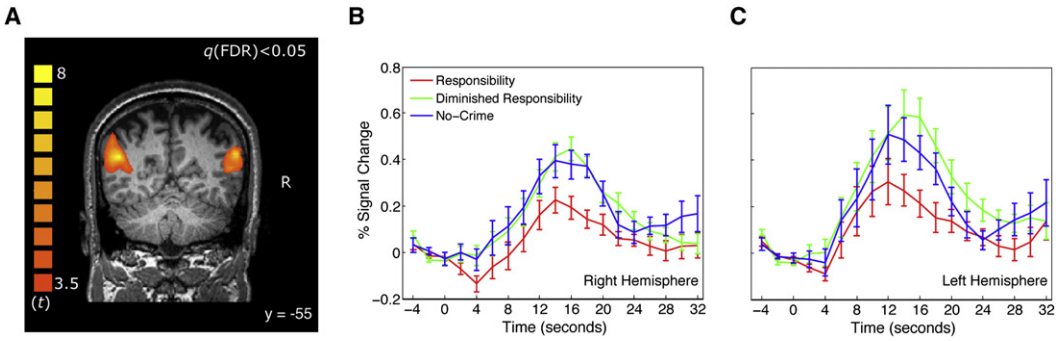

Figure 3. Relationship between Responsibility Assessment and Bilateral Temporo-Parietal Junction Activity

(A) SPM displaying the right and left temporo-parietal junction (TPJ) VOIs (rendered on a single-subject T1-weighted image), based on the contrast of BOLD activity in the Diminished-Responsibility condition and that of the Responsibility condition. $t(15) > 3.5$, $q < 0.05$; random effects analysis. R = Right Hemisphere. BOLD activity time courses in right (B) and left (C) TPJ for the Responsibility, Diminished-Responsibility, and No-Crime conditions are also shown. BOLD peak amplitude was significantly greater in the Diminished-Responsibility condition compared with the Responsibility condition for right ($p = 0.0005$) and left ($p = 0.001$) TPJ. Peak was defined as the single TR with maximal signal change from baseline within the first 13 volumes after scenario presentation onset. t tests were performed on these peak volumes, which were defined separately for each condition and each subject. Error bars = SEM.

between conditions (see Experimental Procedures; R > DR, $p = 0.006$; R > NC, $p = 0.002$; Figure S2). In addition, rDLPFC activity was not correlated with RT ($p = 0.09$ in Responsibility scenarios, $p = .12$ in Diminished-Responsibility scenarios). We also assessed whether the activity pattern in rDLPFC might have been driven by between-condition differences in emotional arousal rather than by differences in criminal responsibility. To this end, we performed a peak activation difference analysis between the Responsibility and Diminished-Responsibility conditions after equating their mean arousal ratings (Responsibility = 3.62, Diminished-Responsibility = 3.50; $p > 0.10$, paired t test; see Experimental Procedures). The results still revealed greater rDLPFC activity in the Responsibility condition as compared with the Diminished-Responsibility condition, even in the absence of arousal differences ($p = 0.0005$, paired t test).

If rDLPFC is involved in the decision-making process to punish blameworthy behavior, then this brain region should be more activated during Diminished-Responsibility scenarios in which subjects still decided to punish (punishment ratings of 1 or greater) as compared with Diminished-Responsibility scenarios in which they did not (punishment ratings of 0). Consistent with this hypothesis, rDLPFC activity was higher in "punished" Diminished-Responsibility trials than in "nonpunished" Diminished-Responsibility trials ($p = 0.04$, paired t test, Figure 2). In turn, rDLPFC activity during nonpunished Diminished-Responsibility trials was not greater than that in No-Crime trials ($p = 0.98$, Figure 2). These results, as well as those for aIPS (Supplemental Results, Figure S1), strongly support the notion that prefrontal and parietal activity is modulated by a punishment-related decisional process.

In addition to the peak activation differences, the time course of rDLPFC activity revealed an early deactivation (negative percent-signal change [PSC] from baseline) around 8 s poststimulus onset. Importantly, this early deactivation ("dip") does not account for the peak activation results outlined above: the activation differences between conditions at the dip do not predict corresponding activation differences at the peak (correlation of subjects' activity differences between the Responsibility and Diminished-Responsibility conditions at the dip and at the peak: $\rho = -0.19$, $p = 0.49$; Figure S3; see Experimental Procedures). Furthermore, rDLPFC activity during nonpunished Diminished-Responsibility and No-Crime trials strongly differed at the dip ($p = 0.008$) but not at the peak ($p = 0.97$), indicating that peak activation differences are not simply carryover effects from differences during the dip.

fMRI Data: Punishment Magnitude

The finding that rDLPFC activity was higher when subjects decided to punish, in either Responsibility scenarios or in punished Diminished-Responsibility trials, raised the possibility that this brain region might track the amount of assessed punishment for a given criminal scenario. However, rDLPFC signal amplitude was not correlated with punishment ratings ($\rho = -0.33$, $p = 0.15$; Figure 2D) in the Responsibility condition. This finding suggests that the magnitude of punishment is not simply coded by a linear increase in rDLPFC activity.

Although rDLPFC activity was not proportional to punishment amount, a linear relationship between peak BOLD amplitude and punishment magnitude was found in a set of brain regions that have been extensively linked to social and affective processing. To isolate such effects, we compared Responsibility scenarios with high punishment ratings to those with low ratings (median split by scenario across subjects; see Experimental Procedures). The resulting SPM revealed activation in the right amygdala (peak Talairach coordinates 29, −7, −13; Figure 4; Figure S5) as well as in other brain regions commonly associated with social and affective processing (LeDoux, 2000; Phelps, 2006; Phillips et al., 2003; Price, 2005), including the posterior cingulate, temporal pole, dorsomedial and ventromedial prefrontal cortex, and inferior frontal gyrus (Table S2; Figures S4 and S5). The association between amygdala activity and punishment magnitude was further demonstrated by a strong correlation between amygdala BOLD signal and punishment ratings across Responsibility scenarios ($\rho = 0.70$, $p = 0.001$; Figure 4). However, punishment rating was not the only variable that correlated with amygdala function, as participants' arousal ratings yielded a similar correlation with

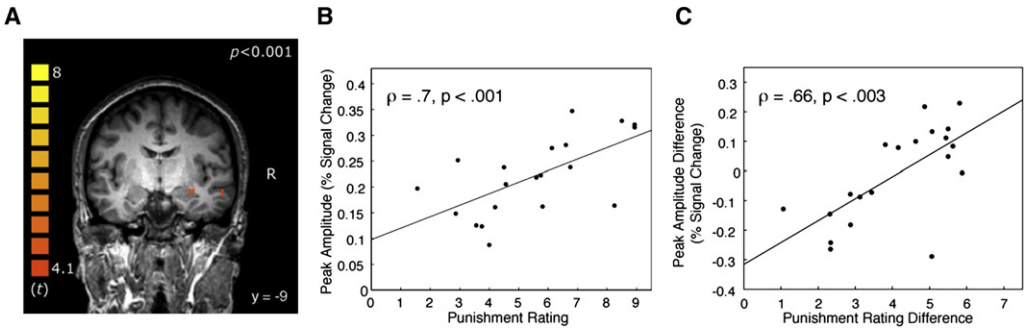

Figure 4. Relationship between Punishment and Right Amygdala Activity

(A) SPM displaying the right amygdala VOI (rendered on a single-subject T1-weighted image), based on the contrast of BOLD activity between high and low punishment (computed from the median split for Responsibility scenarios), thresholded at $t(15) > 4.1$, $p < 0.001$ (uncorrected) for visualization. This amygdala activation survives correction for multiple comparisons. $q < 0.05$; random effects analysis. R = Right Hemisphere.

(B) Relationship between BOLD peak amplitude in the right amygdala and punishment ratings in the Responsibility condition. These two variables were significantly positively correlated ($p = 0.001$).

(C) Relationship between condition differences in right amygdala BOLD peak amplitude (Responsibility minus Diminished-Responsibility) and condition differences in punishment score (Responsibility minus Diminished-Responsibility); these two variables are significantly correlated ($p = 0.001$).

amygdala activity ($\rho = 0.67$, $p = 0.001$), and punishment and arousal ratings were themselves highly correlated ($\rho = 0.98$, $p = 0.000001$). Correlations between peak BOLD signal and punishment ratings (and between peak BOLD signal and arousal ratings) also held for a number of the other affective regions, including ventromedial prefrontal cortex and posterior cingulate cortex (Table S2; Figures S4 and S5), indicating that the relationship between affective processing and punishment involved a distributed neural circuit.

Although the correlation between amygdala activity and punishment scores could be interpreted as evidence for a role of emotional arousal in the assignment of deserved punishment, it is also possible that such activity simply reflected subjects' emotional reaction to the graphic content of the scenarios rather than its involvement in the decision-making process per se. To avoid the potential arousal confound inherent to an examination of criminal scenarios that differ in graphic content (as was the case for our comparison of high versus low punishment scores within the Responsibility condition), we examined the relationship between punishment ratings and amygdala activity after controlling for the possible confounding effect of graphic arousal. Because Responsibility and Diminished-Responsibility scenarios were equated for graphic content and differed only by the presence of mitigating circumstances (see Experimental Procedures), the potentially confounding contribution of graphic arousal to amygdala activity in the Responsibility scenarios can be controlled for by subtracting amygdala activity in the Diminished-Responsibility scenarios from that in the corresponding Responsibility scenarios. If amygdala activity appertains to punishment magnitude rather than, or in addition to, emotional arousal related to the graphic content of the scenarios, it should still track punishment ratings even after subtracting out graphic content differences in the scenarios. To this end, we created, for each pair of Responsibility and Diminished-Responsibility scenarios, punishment rating difference scores (Responsibility minus Diminished-Responsibility) and assessed whether these scores were correlated with the corresponding difference scores for peak amygdala BOLD signal. That correlation was significant ($\rho = 0.62$, $p = 0.001$; Figure 4), indicating that the magnitude of amygdala BOLD signal difference between Responsibility and Diminished-Responsibility conditions for a given scenario predicted a corresponding change in punishment rating for that scenario. Similar correlations were found in posterior cingulate and ventromedial prefrontal cortex (Table S2). These findings suggest that activity within brain regions previously implicated in social and affective processing reflects third-party decisions about how much to punish, even after controlling for the potentially confounding arousal associated with the graphic content of the criminal scenarios.

DISCUSSION

The present findings suggest that the two fundamental components of third-party legal decision-making—determining responsibility and assigning an appropriate punishment magnitude—are not supported by a single neural system. In particular, the results reveal a key role for the right dorsolateral prefrontal cortex in third-party punishment. This brain region appears to be involved in deciding whether or not to punish based on an assessment of criminal responsibility. The only other brain region that demonstrated a comparable pattern of responsibility-related activity (R > DR, R > NC, DR = NC) to rDLPFC was the aIPS (Table S1, Figure S1, Supplemental Results). This parietal region has been associated with a number of diverse cognitive functions including general response selection (Gobel et al., 2004) and quantitative numerical comparisons (Dehaene et al., 1999, 2003; Feigenson et al., 2004), which may hint at a role for this area in associating a specific action (i.e. selecting a punishment outcome) with a given scenario.

Our results also implicate neural substrates for social and affective processing (including amygdala, medial prefrontal cortex, and posterior cingulate cortex) in third-party punishment, albeit in ways distinct from the rDLPFC. Specifically, while prefrontal activity was linked to a categorical aspect of legal

decision-making (deciding whether or not to punish on the basis of criminal responsibility), the magnitude of assigned punishments for criminal transgressions parametrically modulated activity in affective brain regions, even after controlling for the potentially confounding arousal-related activity associated with the graphic content of the criminal scenarios. Our findings suggest that a set of brain regions (e.g., amygdala, medial prefrontal cortex, and posterior cingulated cortex) consistently linked to social and emotional processing (Adolphs, 2002; Amodio and Frith, 2006; Barrett et al., 2007; Lieberman, 2007; Phelps, 2006; Phillips et al., 2003; Zald, 2003) is associated with the amount of assigned punishment during legal decision-making. As such, these results accord well with prior work pointing to social and emotional influences on economic decision-making and moral reasoning (De Martino et al., 2006; Delgado et al., 2005; Koenigs and Tranel, 2007; Greene and Haidt, 2002; Greene et al., 2001, 2004; Haidt, 2001; Heekeren et al., 2003; Koenigs et al., 2007; Moll et al., 2002b, 2005), and provide preliminary neuroscientific support for a proposed role of emotions in legal decision-making (Arkush, 2008; Maroney, 2006). Our data concur with behavioral studies that have proposed a link between affect and punishment motivation in both second- and third-party contexts, and are consistent with the hypothesis that third-party sanctions are fueled by negative emotions toward norm violators (Darley and Pittman, 2003; Fehr and Fischbacher, 2004a, 2004b; Seymour et al., 2007). However, it must be acknowledged that the present conclusions rest exclusively on correlational data. Thus, additional research will be required to confidently determine the contributions of socio-affective brain regions to third-party punishment in the absence of any graphic arousal confound. In particular, it will be important in future experiments to fully dissociate the factors of crime severity and arousal by employing task conditions that manipulate arousal without affecting crime severity. Furthermore, future research should also focus on determining how these affective brain regions interact with DLPFC during third-party punishment decisions.

An additional concern in interpreting our findings, or any others based on simulated judgments, is whether they are relevant to real-world decision-making. After all, the punishment decisions made by our participants did not have direct, real-world consequences for real criminal defendants. Thus, it remains to be seen if our findings, generated by examining brain activation patterns during hypothetical judgments, will generalize to circumstances in which real punishments are made. However, there is some evidence suggesting that the hypothetical judgments made by our subjects may be a good proxy measure for real-world legal judgments. For example, postscan debriefing of our subjects indicated that their punishment assessments were implicitly legal, with lower numbers corresponding to low prison sentences and higher numbers corresponding to high prison sentences (see Table S3). Thus, participants appeared to adopt an internal punishment scale based on incarceration duration—a legal metric—when making their judgments, even in the absence of explicit instructions to do so. Further, we found that participants' decisions about punishment amount for each of the crimes depicted in the Responsibility scenarios were strongly correlated with the recommended prison sentences for those crimes, according to the benchmark sentencing guidelines of North Carolina, a model state penal code ($\rho = 0.8$, $p < .0001$; Figure S6; see Experimental Procedures). Thus, although our subjects were not literally applying a criminal statute to an accused individual, these data suggest that subjects' punishment decisions were consistent with statutory legal reasoning. However, despite these suggestions, further empirical studies are required to confirm our supposition that neuroimaging studies of simulated third-party legal decision-making can validly model real-world legal reasoning.

Relative Contributions of TPJ and rDPLFC to Third-Party Punishment Decisions

The neural mechanisms of third-party punishment are undoubtedly complex, involving a dynamic regional interplay that unfolds in a temporally specific manner. In particular, the decision to punish a person for his blameworthy act is generally preceded by an evaluation of that person's intention in committing that act (Alter et al., 2007; Carlsmith et al., 2002; Darley and Pittman, 2003; Darley and Shultz, 1990; Robinson and Darley, 1995; Robinson et al., 2007; Shultz et al., 1986). Such an evaluation ought therefore to activate brain regions that underlie the attribution of goals, desires, and beliefs to others, referred to as theory of mind (TOM) (Gallagher and Frith, 2003). One such region, the TPJ—a key node in the distributed TOM network (Decety and Lamm, 2007; Gallagher and Frith, 2003; Saxe and Kanwisher, 2003; Vollm et al., 2006)—might be predicted to serve this function during legal decision-making given recent evidence of its role in attributing mental beliefs in moral judgments (Young et al., 2007) and its involvement in dyadic economic exchange games (Rilling et al., 2004). Given this context, it is noteworthy that the TPJ was activated in all of our conditions (Figure 3). Furthermore, TPJ came online during the period when rDLPFC was deactivated (see Figure 2B), a result that is consistent with the suggestion that temporo-parietal cortex and DLPFC operate within largely distinct and at times functionally opposed networks (Fox et al., 2005). Given this proposed antagonistic response pattern in the TPJ and DLPFC, we speculate that the early rDLPFC deactivation may reflect a perspective-taking-based evaluation of the beliefs and intentions of the scenarios' protagonist, which is followed by a robust rDLPFC activation as subjects go on to make a decision to punish based on assessed responsibility and blameworthiness. However, the conclusion that rDLPFC's biphasic time course reflects an initial socio-evaluative process followed by a decisional process must be viewed as tentative because the present experiment did not constrain the temporal sequence of evaluative and decisional processes involved in this task.

Moral versus Legal Decision-Making

The results of the present neuroimaging study underscore the conceptual relationship between moral and legal decision-making. Indeed, the general involvement of both the prefrontal cortex and affective brain regions in legal reasoning is reminiscent of their roles in moral judgment (Greene et al., 2001, 2004). Specifically, moral decision-making studies have indicated that regions of lateral prefrontal cortex and inferior parietal lobe are preferentially involved in impersonal moral judgments, whereas socio-affective areas (e.g., amygdala, medial prefrontal cortex, and posterior cingulate cortex) may be primarily engaged during

personal moral decision-making (Greene et al., 2001, 2004). Thus, both legal and moral decision-making may rely on "cold" deliberate computations supported by the prefrontal cortex and "hot" emotional processes represented in socio-affective brain networks, although the extent to which these two decision-making processes rely on the same brain circuitry remains to be determined.

While these findings serve to highlight an important conceptual overlap between moral reasoning and legal reasoning in criminal contexts, they do not imply that third-party punishment decisions are reducible to moral judgment. Indeed, while legal decision-making may in most (but not all) criminal cases have an essential moral component, there are crucial distinctions between morality and law (Hart, 1958; Holmes, 1991; Posner, 1998). Perhaps the most critical distinguishing feature of legal decision-making, compared with moral decision-making, is the action of punishment—intrinsic to the former and secondary to the latter (Robinson, 1997). Although our participants likely evaluated the moral blameworthiness of the scenarios' protagonist, our study was designed to investigate the neural substrates of a fundamental legal decision—assigning punishment for a crime—that is not a defining characteristic of moral judgment. Indeed, while moral decision-making studies to date have focused on assessing brain function during decisions about the moral rightness or wrongness of actions depicted in written scenarios, they have not specifically addressed the issue of punishment (Borg et al., 2006; Greene et al., 2001, 2004; Heekeren et al., 2003, 2005; Kedia et al., 2008; Luo et al., 2006; Moll et al., 2001, 2002a, 2002b; Young et al., 2007; Young and Saxe, 2008).

Neural Convergence of Second-Party and Third-Party Punishment Systems

The prefrontal cortex area activated in the present third-party legal decision-making study corresponds well to an area that is involved in the implementation of norm enforcement behavior in two-party economic exchanges (peak Talairach coordinates of 39, 37, 22 [x,y,z] for Knoch et al., 2006; Sanfey et al., 2003; versus 39, 38, 18 [x,y,z] for the present study), raising the possibility that rDLPFC serves a function common to both third-party legal and second-party economic decision-making. In this respect, it is noteworthy that this region of rDLPFC is recruited when participants decide whether or not to punish a partner by rejecting an unfair economic deal proposed by that partner (Sanfey et al., 2003); this result is analogous to our finding that rDLPFC is activated by the decision to punish the perpetrator of a criminal act. Furthermore, while disruptive magnetic stimulation of this region impairs the ability to punish economic norm violations in dyadic exchanges (Knoch et al., 2006; van't Wout et al., 2005), this manipulation has no effect on norm enforcement behavior when the unfair economic exchanges are randomly generated by a computer instead of a human agent (Knoch et al., 2006). This result accords well with our finding that rDLPFC was much less activated when the scenario protagonist was not criminally responsible for his behavior, and supports the notion that this prefrontal cortex area is primarily recruited when punishment can be assigned to a responsible agent (Knoch et al., 2006). Finally, we still observed greater rDLPFC activity in the Responsibility condition (as compared with Diminished-Responsibility scenarios) when we restricted our analysis to scenarios that only contained physical harms ($p < 0.005$, paired t test), suggesting that the overlap of rDLPFC activity between studies of economic decision-making and the present examination of legal decision-making is not solely driven by scenarios describing economic transgressions.

The parallels between these previous findings and our current results lead us to suggest that the rDLPFC is strongly activated by the decision to punish norm violations based on an evaluation of the blameworthiness of the transgressor. This proposed function of rDLPFC appears to apply equally to situations where the motive for punishment is unfair behavior in a dyadic economic exchange or when responding to the violation of an institutionalized social norm in a disinterested third-party context. Of course, confirmation of this hypothesis will require further experimental evidence that legal and economic decision-making (and perhaps moral decision-making as well) rely on the same neural substrates. That said, this apparent overlap illustrates an important point: that the brain regions identified in our study are not specifically devoted to legal decision-making. Rather, a more parsimonious explanation is that third-party punishment decisions draw on elementary and domain-general computations supported by the rDLPFC. In particular, on the basis of the convergence between neural circuitry mediating second-party norm enforcement and impartial third-party punishment, we conjecture that our modern legal system may have evolved by building on preexisting cognitive mechanisms that support fairness-related behaviors in dyadic interactions. Though speculative and subject to experimental confirmation, this hypothesis is nevertheless consistent with the relatively recent development of state-administered law enforcement institutions, compared to the much longer existence of human cooperation (Richerson et al., 2003); for thousands of years before the advent of state-implemented norm compliance, humans relied on personal sanctions to enforce social norms (Fehr et al., 2002; Fehr and Gachter, 2002).

EXPERIMENTAL PROCEDURES

Subjects

Sixteen right-handed individuals (eight males, ages 18–42) with normal or corrected-to-normal vision participated for financial compensation. The Vanderbilt University Institutional Review Board approved the experimental protocol, and informed consent was obtained from each subject after they were briefed on the nature and possible consequences of the study. A brief psychological survey was also administered to exclude individuals who may react adversely to the content of the criminal scenarios. Exclusion criteria included history of psychiatric illness, being the victim of or having witnessed a violent crime (including sexual abuse), and having experienced any trauma involving injury or threat of injury to the subject or a close friend/family member.

Paradigm

In this experiment, subjects participated in a simulated third-party legal decision-making task in which they determined the appropriate level of punishment for the actions of a fictional protagonist described in short written scenarios. The principal goal of our study was to isolate the neural processes associated with the two fundamental processes of legal decision-making: deciding whether or not an accused individual is culpable for a given criminal act, and determining the appropriate punishment for that act (a parametric process based on the ordinal severity of a crime). Correspondingly, our design manipulated responsibility in a dichotomous fashion and crime severity in a continuous

fashion. Each participant viewed 50 scenarios (some inspired by prior behavioral studies of relative blameworthiness; Robinson and Darley, 1995; Robinson and Kurzban, 2007) depicting the actions of the protagonist named "John." The 50 scenarios were subdivided into three sets (complete scenario list is available as Supplemental Experimental Procedures). In the Responsibility set (n = 20), the scenarios described John intentionally committing a criminal action ranging from simple theft to rape and murder. The Diminished-Responsibility set (n = 20) included similar actions comparable in gravity to those in the Responsibility set, but contained circumstances that would often legally excuse or justify the otherwise criminal behavior of the protagonist. The No-Crime set (n = 10) depicted John engaged in noncriminal actions that were otherwise structured similarly to the Responsibility and Diminished-Responsibility scenarios. The No-Crime scenarios were included to assist in interpreting activity differences between Responsibility and Diminished-Responsibility scenarios (e.g. Figure 2).

Two groups of 50 scenarios were constructed and their presentation counterbalanced across the 16 participants (8 subjects received group 1 scenarios, and 8 others received group 2 scenarios) and across gender (equal numbers of men and women received scenarios from each group). The Responsibility set of group 2 consisted of group 1 Diminished-Responsibility scenarios from which the mitigating circumstances had been excised, while the Diminished-Responsibility set of group 2 consisted of group 1 Responsibility scenarios with mitigating circumstances added. As a result, the Responsibility and Diminished-Responsibility scenarios were counterbalanced across subjects, and differed only by the presence of mitigating circumstances. Thus, exactly the same scenario premises were used in constructing the Responsibility and Non-Responsibility conditions. Finally, the No-Crime set was identical in both groups of scenarios, and all scenario sets were equated for word length.

Participants rated each scenario on a scale from 0–9, according to how much punishment they thought John deserved, with "0" indicating no punishment and "9" indicating extreme punishment. Punishment was defined for participants as "deserved penalty." Participants were asked to consider each scenario (and thus, each "John") independently of the others and were encouraged to use the full scale (0–9) for their ratings. In the scanner but prior to the functional scans, subjects were shown five practice scenarios that were designed to span the punishment scale. Scenarios were presented as white text (Times New Roman font) on a black background (14.2° [width] × 9.9° [height] of visual angle). Below each scenario, text reminded participants of the task instructions: "How much punishment do you think John deserves, on a scale from 0 to 9 where 0 = No punishment and 9 = Extreme punishment? By punishment, we mean deserved penalty." Participants were instructed to make a response as soon as they had reached their decision.

Each trial began with the presentation of a scenario, which remained onscreen until participants made a button press response, or up to a maximum of 30 s. Participants then viewed a small white fixation square (0.25° of visual angle) for 12–14 s (as stimulus onset was synched to scan acquisition [TR = 2 s], while stimulus offset was synched to subject response), which was followed by a larger fixation square (0.49° of visual angle) for 2 s prior to the presentation of the next scenario. Ten scenarios (four Responsibility, four Diminished-Responsibility, and two No-Crime)—selected randomly without replacement from the fifty scenarios—were presented in each of the five fMRI runs. Scenario identity and condition order were randomized for each run. The duration of each fMRI run was variable, with a maximum length of 7.33 min. The experiment was programmed in Matlab (Mathworks, Natick MA) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997) and was presented using a Pentium IV PC.

Following the scanning session, participants rated the same scenarios along scales of emotional arousal and valence. They first rated each of the 50 scenarios (presented in random order on a computer screen outside the scanner) on the basis of how emotionally aroused they felt following its presentation (0 = calm, 9 = extremely excited). They then rated each of the scenarios, presented again in random order, on the basis of how positive or negative they felt following its presentation (0 = extremely positive, 9 = extremely negative). In these sessions, subjects rated the same scenarios they viewed in the scanner. The valence data were highly correlated with arousal ratings, and multiple regression analysis demonstrated that they did not account for any additional variance in punishment ratings that is unaccounted for by the arousal data. Therefore, the valence data are not further discussed in this manuscript.

Internal Scale Questionnaire

In a postscan debriefing, participants were questioned about the internal scale of punishment they used during the scan. Specifically, participants were asked "what kind of punishment did you imagine"? for punishment scores of 1, 3, 5, 8, and 9. There was strong agreement among participants about their internal scale of justice. While low punishment scores (1, 3) were generally associated with financial or social penalties, greater punishment scores (5, 8) included incarceration time, with higher scores associated with longer jail times and, at the extreme (9), life imprisonment or state execution.

Relationship between Punishment Ratings and Legal Statutes

To investigate the relationship between punishment ratings for Responsibility scenarios obtained in the present experiment and an existing, statutorily prescribed punishment for each of the crimes depicted in these scenarios, we coded each Responsibility scenario using the criminal law and criminal procedure statutes of the state of North Carolina. Among those states that have a sentencing statute, North Carolina's is widely considered to be both comprehensive and exemplary (Stanley, 1996; Wright, 2002).

For each Responsibility scenario, we determined the crime or crimes (such as larceny, involuntary manslaughter, or murder) with which John might reasonably be charged under the criminal code of North Carolina (2005 General Statutes of North Carolina, Chapter 14). We then determined, for each crime, the authorized presumptive sentencing range (such as 58 to 73 months in prison), assuming no aggravating or mitigating factors that could, under the statute, increase or decrease the authorized sentencing range (2005 General Statutes of North Carolina, Chapter 15A, Article 81). We then calculated and assigned to each scenario the mean for this range, in months. As the distribution of sentence values was highly right-skewed, we log-transformed (natural log) to create a normal distribution of sentence values (we verified that nontransformed data produced similar correlations as transformed data). For scenarios with multiple crimes, the averages for each respective crime were summed (whether this summed value or simply the mean value for the most severe crime depicted in a given scenario was used in the correlation analysis did not significantly affect the results). Where the upper limit of the sentencing range was life in prison, it was coded as 29 years (which has been estimated as the average time likely to be served by lifers newly admitted in 1997) (Mauer et al., 2004). Similarly, where the upper limit of the sentencing range was death, it was also quantified as life in prison (29 years). The log-transformed mean sentences for each of the 20 scenarios were then correlated with the group-averaged punishment ratings for these scenarios.

Statistical Analysis

Mean punishment and arousal scores and RTs were calculated for each subject for each condition (Responsibility, Diminished-Responsibility, and No-Crime) and entered into a repeated-measures analysis of variance (ANOVA) using SPSS 15 (SPSS Inc., Chicago, IL) to determine main effects and interactions. Data from 16 subjects were used for all analyses. Punishment, arousal scores, and RTs were compared between conditions and post hoc tests were performed using Fisher's Least Significant Difference (LSD) measure using an alpha level of 0.05. Two-tailed tests were used in all cases. For correlational analyses, data from Responsibility scenarios (n = 20) were averaged across all (n = 16) subjects. Examination of scatterplots for the correlation of rDLPFC signal and punishment suggested the presence of outliers. As nonparametric correlations tend to be more robust to outliers, we used Spearman's ρ to measure correlations between fMRI signal, behavioral measures, and recommended sentences. All correlations that were significant using Spearman's ρ were also significant ($p < 0.05$) when we employed Pearson's r.

fMRI Data Acquisition

High-resolution 2D and 3D anatomical images were acquired with conventional parameters on a 3T Philips Achieva scanner at the Vanderbilt University Institute of Imaging Science. The visual display was presented on an LCD panel and back-projected onto a screen positioned at the front of the magnet bore. Subjects lied supine in the scanner and viewed the display on a mirror positioned above them. Stimulus presentation was synchronized to fMRI volume acquisition. Manual responses were recorded using two five-button keypads (one for each hand; Rowland Institute of Science, Cambridge, MA).

Functional (T2* weighted) images were acquired using a gradient-echo echoplanar imaging (EPI) pulse sequence with the following parameters: TR 2000 ms, TE 25 ms, flip angle 70°, FOV 220 × 220 mm, 128 × 128 matrix with 34 axial slices (3 mm, 0.3 mm gap) oriented parallel to the gyrus rectus. These image parameters produced good T2* signal across the brain except in ventromedial frontal cortex, where some signal dropout was evident in all subjects (Brodmann area 11).

Each of the 16 participants performed five fMRI runs, except for 2 participants who could only complete four runs due to technical malfunctions.

fMRI Data Preprocessing

Image analysis was performed using Brain Voyager QX 1.4 (Brain Innovation, Maastricht, The Netherlands) with custom Matlab software (MathWorks, Natick, MA).

Prior to random effects analysis, images were preprocessed using 3D motion correction, slice timing correction, linear trend removal, and spatial smoothing with a 6 mm Gaussian kernel (full width at half maximum). Subjects' functional data were coregistered with their T1-weighted anatomical volumes and transformed into standardized Talairach space.

Responsibility Analysis

This analysis was performed to isolate brain regions that were sensitive to responsibility during punishment assessment. Signal values for each fMRI run were transformed into Z-scores representing a change from the signal mean for that run and corrected for serial autocorrelations. Design matrices for each run were constructed by convolving a model hemodynamic response function (double gamma, consisting of a positive γ function and a small, negative γ function reflecting the BOLD undershoot – SPM2, http://www.fil.ion.ucl.ac.uk/spm) with regressors specifying volumes acquired during the entire trial (stimulus onset to stimulus offset) for a given condition. These were entered into a general linear model (GLM) with separate regressors created for each condition per subject (random effects analysis). We then contrasted the beta-weights of regressors using a t test between conditions to create an SPM showing voxels that demonstrated significantly increased activation in the Responsibility condition as compared with the Diminished-Responsibility condition. Predictors for the No-Crime condition were weighted with a zero (i.e., not explicitly modeled). We applied a False-Discovery Rate (FDR) threshold of $q < 0.05$ (with $[c(V) = ln(V) + E]$) to correct for multiple comparisons. Only activations surviving this corrected threshold are reported.

Volumes of interest (VOIs) were created from the suprathreshold clusters isolated in the above SPM at the conservative FDR threshold. The boundary of these VOIs was drawn from SPMs thresholded using a less conservative implementation of FDR ($q < 0.05$, $c(V) = 1$). The signal for each trial (event) included the time course from 2 TRs (4 s) before stimulus onset to 13 TRs (26 s) after. Each event's signal was transformed to a PSC relative to the average of the first three TRs (0–4 s before stimulus onset). Event-related averages (ERAs) were created by averaging these PSC-adjusted event signals; separate ERAs were created for each combination of VOI, condition, and subject. These ERAs were then averaged across subjects for display purposes.

As subjects were instructed to make a response as soon as they had reached a decision about punishment amount, and in keeping with other neuroimaging studies of decision-making (Aron and Poldrack, 2006; Coricelli et al., 2005; Dux et al., 2006; Ivanoff et al., 2008; Rahm et al., 2006), decision-related activity should correspond to the portion of the time course that follows subjects' response. Given that mean RTs hovered around 12 s (mean, SE for: Responsibility = 12.69 s, 0.46; Diminished-Responsibility = 13.76 s, 0.46; No-Crime = 11.12 s, 0.44; respectively) and accounting for a hemodynamic peak rise time of about 5 s poststimulus (Boynton et al., 1996; Friston et al., 1994; Heeger and Ress, 2002), peridecision activity should occur approximately 17 s after trial onset, which corresponds well with the time of peak hemodynamic response observed in rDLPFC (see Figure 2). We therefore used the peak hemodynamic response as a measure of decision-related activity. To determine condition effects on BOLD signal within a given brain region, we then contrasted each condition's activation averaged across subjects by using paired t tests applied on these peak estimates. The peak was experimentally defined as the single volume with maximal signal change from baseline between volumes 1 and 13 (2–26 s poststimulus onset). However, we ascertained that the same results were obtained when the peak was defined using a narrower volume range of 14 to 22 s poststimulus (R > DR, p = 0.00070; R > NC, p = 0.00025; DR > NC, p = 0.19), or even when using a single volume 16 s poststimulus (R > DR, p = 0.00023; R > NC, p = 0.00027; DR > NC, p = 0.84). Thus, our rDLPFC peak activation results are insensitive to the temporal width of the analysis window.

Arousal- and Reaction-Time-Equated Analyses

To determine whether activation differences between the Responsibility and Diminished-Responsibility conditions were driven by punishment assessment rather than any differences in arousal, these two conditions were compared after equating for arousal ratings. This was accomplished by deleting the six trials with the highest arousal ratings from the Responsibility condition for each subject. Time courses were extracted and peak differences were compared as above.

We also determined whether RT differences between the Responsibility, Diminished-Responsibility, and No-Crime conditions affected the brain activation results by comparing these conditions after equating for response times. This was accomplished by deleting, for each subject, the trials with the highest RTs for Diminished-Responsibility scenarios and the trials with the lowest RTs for the No-Crime scenarios until the RTs across conditions (for each subject) were approximately equal ($p > 0.1$ for all paired t tests between conditions). In addition, we compared rDLPFC activation between Responsibility and Diminished-Responsibility scenarios controlling for RT by performing a GLM ANCOVA using the extracted rDLPFC BOLD signal and punishment RTs for each Responsibility and Diminished-Responsibility scenario averaged across subjects.

Dissociation of Activation Peak and Deactivation Dip

To assess the relationship between early (~8 s) deactivation in the rDLPFC time course and later (~16 s) peak activation, we calculated peak and dip values for the Responsibility and Diminished-Responsibility conditions from each subject's ERA. Peak and dip were defined as the volume with the maximal positive and maximal negative change from baseline, respectively. For each subject, we subtracted the Diminished-Responsibility peak value from the Responsibility peak value, and the Diminished-Responsibility dip value from the Responsibility dip value. Per-subject peak and dip difference values were then correlated via Spearman bivariate correlation in SPSS 15.

Laterality Analyses

To confirm the lateral specificity of Responsibility-related activation in rDLPFC, we extracted BOLD signal from the corresponding left DLPFC VOI (i.e., "x-mirrored" VOI, centered on Talairach coordinate −39, 37, 22). We performed a two-way ANOVA with "Condition" (Responsibility, Diminished-Responsibility, and No-Crime) and "Side" (Left and Right) as independent variables and BOLD signal as the dependent variable. Post hoc comparisons between conditions in each hemisphere, and between hemispheres for the Responsibility condition, were performed using paired t tests.

Punishment Rating Analysis

To identify brain regions that tracked the degree of punishment subjects assigned to a scenario, we performed a median split for punishment scores given during Responsibility scenarios. Based on the median punishment value for each scenario in the Responsibility condition across subjects, scenarios were separated into two groups, high and low. Design matrices and GLMs were constructed as above, with predictors for high and low scores for each subject specifying volumes acquired during Responsibility trials on which a high or low punishment score was given, respectively. We contrasted the beta-weights of these predictors using a t test between high and low punishments to create an SPM showing voxels that demonstrated significantly increased activation during Responsibility trials in which subjects gave high (at or above the median) punishments relative to Responsibility trials in which subjects gave low (below the median) punishments. We applied a threshold of $q < 0.05$ FDR to correct for multiple comparisons. Using a conservative implementation of the FDR correction technique ($c(V) = ln(V) + E$), we did not find significant activation differences. We report activations significant at FDR $q < 0.05$, using a less conservative implementation of FDR ($c(V) = 1$). The differences between the two implementations relate to assumptions about the independence of tests being performed on the data; both are valid controls for multiple testing in functional imaging data (Genovese et al., 2002).

VOIs were created as described for the Responsibility analysis. The extracted peak activation values were used for a correlation analysis between

punishment rating and BOLD response. Specifically, for each of the 20 Responsibility scenarios, the peak amplitude of the group-averaged ERA was computed, and the resulting value was correlated with the corresponding group-averaged punishment rating for that scenario. These peak values were also used in the between-condition difference score analyses.

SUPPLEMENTAL DATA

The supplemental data for this article include Supplemental Results, six supplemental Figures, Experimental Scenarios, and three supplemental Tables and can be found at http://www.neuron.org/supplemental/S0896-6273(08)00889-1.

ACKNOWLEDGMENTS

This research was supported by grants from the John D. and Catherine T. MacArthur Foundation Law and Neuroscience Project, the Vanderbilt University Central Discovery Grant Program, the Vanderbilt Law and Human Behavior Program, and the Cecil D. Branstetter Litigation and Dispute Resolution Program of Vanderbilt University. The authors wish to thank Martha Presley for providing valuable background research and Jeffrey Schall, Nita Farahany, Terry Maroney, Michael Treadway, Eyal Aharoni, Terrence Chorvat, and Walter Sinnott-Armstrong for useful comments.

Accepted: October 14, 2008
Published: December 10, 2008

REFERENCES

Adolphs, R. (2002). Neural systems for recognizing emotion. Curr. Opin. Neurobiol. *12*, 169–177.

Alter, A.L., Kernochan, J., and Darley, J.M. (2007). Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. Law Hum. Behav. *31*, 319–335.

Amodio, D.M., and Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. *7*, 268–277.

Arkush, D.J. (2008). Situating Emotion: A Critical Realist View Of Emotion and Nonconscious Cognitive Processes for the Law. SSRN, http://ssrn.com/abstract=1003562.

Aron, A.R., and Poldrack, R.A. (2006). Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. J. Neurosci. *26*, 2424–2433.

Barrett, L.F., Mesquita, B., Ochsner, K.N., and Gross, J.J. (2007). The experience of emotion. Annu. Rev. Psychol. *58*, 373–403.

Bendor, J., and Swistak, P. (2001). The Evolution of Norms. AJS *106*, 1493–1545.

Borg, J.S., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an FMRI investigation. J. Cogn. Neurosci. *18*, 803–817.

Boynton, G.M., Engel, S.A., Glover, G.H., and Heeger, D.J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. *16*, 4207–4221.

Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

Carlsmith, K.M., Darley, J.M., and Robinson, P.H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. J. Pers. Soc. Psychol. *83*, 284–299.

Coricelli, G., Critchley, H.D., Joffily, M., O'Doherty, J.P., Sirigu, A., and Dolan, R.J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. Nat. Neurosci. *8*, 1255–1262.

Darley, J.M., and Shultz, T.R. (1990). Moral Rules: Their Content and Acquisition. Annu. Rev. Psychol. *41*, 525–556.

Darley, J.M., and Pittman, T.S. (2003). The psychology of compensatory and retributive justice. Pers. Soc. Psychol. Rev. *7*, 324–336.

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. Science *313*, 684–687.

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). The neural basis of altruistic punishment. Science *305*, 1254–1258.

Decety, J., and Lamm, C. (2007). The Role of the Right Temporoparietal Junction in Social Interaction: How Low-Level Computational Processes Contribute to Meta-Cognition. Neuroscientist *13*, 580–593.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., and Tsivkin, S. (1999). Sources of mathematical thinking: behavioral and brain-imaging evidence. Science *284*, 970–974.

Dehaene, S., Piazza, M., Pinel, P., and Cohen, L. (2003). Three Parietal Circuits for Number Processing. Cogn. Neuropsychol. *20*, 487–506.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nat. Neurosci. *8*, 1611–1618.

Dux, P.E., Ivanoff, J., Asplund, C.L., and Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved FMRI. Neuron *52*, 1109–1120.

Farahany, N., and Coleman, J., Jr. (2006). Genetics and Responsibility: To Know the Criminal From the Crime. Law Contemp. Probl. *69*, 115–164.

Fehr, E., and Gachter, S. (2002). Altruistic punishment in humans. Nature *415*, 137–140.

Fehr, E., and Fischbacher, U. (2004a). Social norms and human cooperation. Trends Cogn. Sci. *8*, 185–190.

Fehr, E., and Fischbacher, U. (2004b). Third-party punishment and social norms. Evol. Hum. Behav. *25*, 63–87.

Fehr, E., Fischbacher, U., and Gachter, S. (2002). Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms. Hum. Nat. *13*, 1–25.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. Trends Cogn. Sci. *8*, 307–314.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., and Raichle, M.E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. USA *102*, 9673–9678.

Friston, K.J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. Hum. Brain Mapp. *1*, 153–171.

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind'. Trends Cogn. Sci. *7*, 77–83.

Garland, B. (2004). Neuroscience and the Law: Brain, Mind and the Scales of Justice (Washington, D.C.: Dana Press).

Garland, B., and Glimcher, P.W. (2006). Cognitive neuroscience and the law. Curr. Opin. Neurobiol. *16*, 130–134.

Genovese, C.R., Lazar, N.A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage *15*, 870–878.

Gobel, S.M., Johansen-Berg, H., Behrens, T., and Rushworth, M.F. (2004). Response-selection-related parietal activation during number comparison. J. Cogn. Neurosci. *16*, 1536–1551.

Greene, J., and Haidt, J. (2002). How (and where) does moral judgment work? Trends Cogn. Sci. *6*, 517–523.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., and Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science *293*, 2105–2108.

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

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. *108*, 814–834.

Hart, H.L.A. (1958). Positivism and the Separation of Law and Morals. Harv. Law Rev. *71*, 593–629.

Heeger, D.J., and Ress, D. (2002). What does fMRI tell us about neuronal activity? Nat. Rev. Neurosci. *3*, 142–151.

Heekeren, H.R., Wartenburger, I., Schmidt, H., Schwintowski, H.P., and Villringer, A. (2003). An fMRI study of simple ethical decision-making. Neuroreport *14*, 1215–1219.

Heekeren, H.R., Wartenburger, I., Schmidt, H., Prehn, K., Schwintowski, H.P., and Villringer, A. (2005). Influence of bodily harm on neural correlates of semantic and moral decision-making. Neuroimage *24*, 887–897.

Henrich, J. (2003). The cultural and genetic evolution of human cooperation. In Genetic and Cultural Evolution of Cooperation, P. Hammerstein, ed. (Cambridge, MA: MIT Press), pp. 445–468.

Holmes, O.W., Jr. (1991). The Common Law (New York: Dover Publications).

Ivanoff, J., Branning, P., and Marois, R. (2008). fMRI evidence for a dual process account of the speed-accuracy tradeoff in decision-making. PLoS ONE *3*, e2635.

Kedia, G., Berthoz, S., Wessa, M., Hilton, D., and Martinot, J.L. (2008). An Agent Harms a Victim: A Functional Magnetic Resonance Imaging Study on Specific Moral Emotions. J. Cogn. Neurosci. *20*, 1788–1798.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. Science *308*, 78–83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science *314*, 829–832.

Koenigs, M., and Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. J. Neurosci. *27*, 951–956.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. Nature *446*, 908–911.

LaFave, W. (2003). Criminal Law, Fourth edition (St. Paul, MN: West Group).

LaFave, W.R., Israel, J.H., King, N.J., and Kerr, O.S. (2007). Criminal Procedure, Volume 6, Third edition (St. Paul, MN: West Group).

LeDoux, J.E. (2000). Emotion circuits in the brain. Annu. Rev. Neurosci. *23*, 155–184.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. Annu. Rev. Psychol. *58*, 259–289.

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., and Blair, R.J. (2006). The neural basis of implicit moral attitude–an IAT study using event-related fMRI. Neuroimage *30*, 1449–1457.

Maroney, T.A. (2006). Law and emotion: a proposed taxonomy of an emerging field. Law Hum. Behav. *30*, 119–142.

Mauer, M., King, R.S., and Young, M.C. (2004). The Meaning of "Life": Long Prison Sentences in Context (Washington, D.C.: The Sentencing Project).

Moll, J., Eslinger, P.J., and Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task: preliminary functional MRI results in normal subjects. Arq. Neuropsiquiatr. *59*, 657–664.

Moll, J., de Oliveira-Souza, R., Bramati, I.E., and Grafman, J. (2002a). Functional networks in emotional moral and nonmoral social judgments. Neuroimage *16*, 696–703.

Moll, J., de Oliveira-Souza, R., Eslinger, P.J., Bramati, I.E., Mourao-Miranda, J., Andreiuolo, P.A., and Pessoa, L. (2002b). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. J. Neurosci. *22*, 2730–2736.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). Opinion: the neural basis of human moral cognition. Nat. Rev. Neurosci. *6*, 799–809.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. *10*, 437–442.

Phelps, E.A. (2006). Emotion and cognition: insights from studies of the human amygdala. Annu. Rev. Psychol. *57*, 27–53.

Phillips, M.L., Drevets, W.C., Rauch, S.L., and Lane, R. (2003). Neurobiology of emotion perception I: The neural basis of normal emotion perception. Biol. Psychiatry *54*, 504–514.

Posner, R. (1998). The Problematics of Moral and Legal Theory. Harv. Law Rev. *111*, 1657–1710.

Price, J.L. (2005). Free will versus survival: brain systems that underlie intrinsic constraints on behavior. J. Comp. Neurol. *493*, 132–139.

Rahm, B., Opwis, K., Kaller, C.P., Spreer, J., Schwarzwald, R., Seifritz, E., Halsband, U., and Unterrainer, J.M. (2006). Tracking the subprocesses of decision-based action in the human frontal lobes. Neuroimage *30*, 656–667.

Richerson, P.J., Boyd, R.T., and Henrich, J. (2003). Cultural Evolution of Human Cooperation. In Genetic and Cultural Evolution of Cooperation, P. Hammerstein, ed. (Cambridge, MA: MIT Press).

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2004). The neural correlates of theory of mind within interpersonal interactions. Neuroimage *22*, 1694–1703.

Robinson, P. (1997). Structure and Function in Criminal Law (Oxford: Clarendon Press).

Robinson, P. (2002). Mens Rea. In Encyclopedia of Crime and Justice, J. Dressler, ed. (New York: Macmillan Reference USA), pp. 995–1006.

Robinson, P., and Darley, J.M. (1995). Justice, Liability and Blame (San Francisco: Westview).

Robinson, P., and Kurzban, R. (2007). Concordance and conflict in intuitions of justice. Minnesota Law Review *91*, 1829–1907.

Robinson, P., Kurzban, R., and Jones, O.D. (2007). The Origins of Shared Intuitions of Justice. Vanderbilt Law Rev. *60*, 1633–1690.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. Science *300*, 1755–1758.

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage *19*, 1835–1842.

Seymour, B., Singer, T., and Dolan, R. (2007). The neurobiology of punishment. Nat. Rev. Neurosci. *8*, 300–311.

Shultz, T.R., Wright, K., and Schleifer, M. (1986). Assignment of Moral Responsibility and Punishment. Child Dev. *57*, 177–184.

Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., and Frith, C.D. (2004). Brain responses to the acquired moral status of faces. Neuron *41*, 653–662.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., and Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature *439*, 466–469.

Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G., and Fehr, E. (2007). The neural signature of social norm compliance. Neuron *56*, 185–196.

Stanley, L. (1996). Breaking Up Prison Gridlock. ABA J. *82*, 70–75.

van't Wout, M., Kahn, R.S., Sanfey, A.G., and Aleman, A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. Neuroreport *16*, 1849–1852.

Vollm, B.A., Taylor, A.N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J.F., and Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. Neuroimage *29*, 90–98.

Wright, R.F. (2002). Counting the Cost of Sentencing in North Carolina, 1980-2000. In Crime and Justice: A Review of Research, M. Tonry, ed. (Chicago: U. Chicago Press), pp. 39–112.

Young, L., and Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. NeuroImage *40*, 1912–1920.

Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. USA *104*, 8235–8240.

Zald, D.H. (2003). The human amygdala and the emotional evaluation of sensory stimuli. Brain Res. Brain Res. Rev. *41*, 88–123.

Zeki, S., and Goodenough, O. (2004). Law and the brain: introduction. Philosophical Transactions: Biological Sciences *359*, 1661–1665.