Behavioral/Cognitive

Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment

Matthew R. Ginther,[1,2] Richard J. Bonnie,[3] Morris B. Hoffman,[4] Francis X. Shen,[5] Kenneth W. Simons,[6] Owen D. Jones,[2,7,8,9] and René Marois[9,10]

[1]Neuroscience Graduate Program, Vanderbilt University, Nashville, Tennessee 37203, [2]Vanderbilt Law School, Nashville, Tennessee 37203, [3]Institute of Law, Psychiatry and Public Policy, University of Virginia School of Law, Charlottesville, Virginia 22903, [4]District Judge, Second Judicial District (Denver), State of Colorado, Denver, Colorado 80202, [5]Department of Law, University of Minnesota, Minneapolis, Minnesota 55455, [6]Department of Law, University of California, Irvine School of Law, Irvine, California 92697, [7]Departments of Law and Biological Sciences, Vanderbilt University, Nashville, Tennessee 37203, [8]Director, MacArthur Foundation Research Network on Law and Neuroscience, [9]Center for Integrative and Cognitive Neuroscience, and [10]Department of Psychology, Vanderbilt University, Nashville, Tennessee 37240

The evolved capacity for third-party punishment is considered crucial to the emergence and maintenance of elaborate human social organization and is central to the modern provision of fairness and justice within society. Although it is well established that the mental state of the offender and the severity of the harm he caused are the two primary predictors of punishment decisions, the precise cognitive and brain mechanisms by which these distinct components are evaluated and integrated into a punishment decision are poorly understood. Using fMRI, here we implement a novel experimental design to functionally dissociate the mechanisms underlying evaluation, integration, and decision that were conflated in previous studies of third-party punishment. Behaviorally, the punishment decision is primarily defined by a superadditive interaction between harm and mental state, with subjects weighing the interaction factor more than the single factors of harm and mental state. On a neural level, evaluation of harms engaged brain areas associated with affective and somatosensory processing, whereas mental state evaluation primarily recruited circuitry involved in mentalization. Harm and mental state evaluations are integrated in medial prefrontal and posterior cingulate structures, with the amygdala acting as a pivotal hub of the interaction between harm and mental state. This integrated information is used by the right dorsolateral prefrontal cortex at the time of the decision to assign an appropriate punishment through a distributed coding system. Together, these findings provide a blueprint of the brain mechanisms by which neutral third parties render punishment decisions.

Key words: decision-making; fMRI; harm; law; mental state; punishment

Significance Statement

Punishment undergirds large-scale cooperation and helps dispense criminal justice. Yet it is currently unknown precisely how people assess the mental states of offenders, evaluate the harms they caused, and integrate those two components into a single punishment decision. Using a new design, we isolated these three processes, identifying the distinct brain systems and activities that enable each. Additional findings suggest that the amygdala plays a crucial role in mediating the interaction of mental state and harm information, whereas the dorsolateral prefrontal cortex plays a crucial, final-stage role, both in integrating mental state and harm information and in selecting a suitable punishment amount. These findings deepen our understanding of how punishment decisions are made, which may someday help to improve them.

Introduction

Punishment undergirds cooperation. Although forms of cooperation can occur without it, the potential for third-party punishment (i.e., punishment administered by a neutral party) helps counteract temptations to defect (free riding) (Fehr and Gächter, 2002). This, in turn, enabled our species, with uniquely extensive cooperation among nonkin, to flourish at massive scales reflect-

Received Dec. 11, 2015; revised July 20, 2016; accepted July 23, 2016.

Author contributions: M.R.G., R.J.B., M.B.H., F.X.S., K.W.S., O.D.J., and R.M. designed research; M.R.G. and R.M. performed research; M.R.G. and R.M. analyzed data; M.R.G., R.J.B., M.B.H., F.X.S., K.W.S., O.D.J., and R.M. wrote the paper.

This work was supported by a grant from the John D. and Catherine T. MacArthur Foundation to O.D.J. and by Vanderbilt's Center for Integrative and Cognitive Neuroscience. Its contents reflect the views of the authors, and do not necessarily represent the official views of either the John D. and Catherine T. MacArthur Foundation or the MacArthur Foundation Research Network on Law and Neuroscience (www.lawneuro.org). Legal analysis was contributed by M.R.G., R.J.B., M.B.H, F.X.S., K.W.S., and O.D.J. We thank B.J. Casey, Terry Lohrenz, Marc Raichle, and Anthony Wagner for helpful comments on the manuscript; and Benjamin Tamber-Rosenau for being a helpful resource from start to finish.

ing unparalleled social, technological, and economic achievement (Fehr and Rockenbach, 2004; Bowles and Gintis, 2011; Mathew and Boyd, 2011). Nonetheless, punishment decisions are costly to those punished and to society. Thus, efforts at criminal justice reform often center on improving and debiasing punishment decisions themselves, which are central to the fates of so many, and crucial to a just society. Yet despite its importance, little is known about the precise linkage between brain mechanisms and punishment decisions.

Behavioral studies have identified the primary factors that influence punishment decisions: (1) the mental state of the offender; and (2) the severity of harm he caused (Carlsmith et al., 2002; Cushman, 2008). Although this comports with real-world legal norms and practices (LaFave, 1986; Shen et al., 2011), the process by which these two distinct components are integrated into a single punishment decision has not been well characterized. Similarly, brain mechanisms underlying this integrative process remain poorly understood. Prior research of punishment decision-making has suggested that these two different components are neurally dissociable, with mental state evaluation primarily recruiting temporoparietal junction (TPJ), superior temporal sulcus (STS), and dorsomedial prefrontal cortex (DMPFC) (Corradi-Dell'Acqua et al., 2014), and the evaluation of harmful events predominantly engaging affective circuitry, such as the amygdala and the insula (Jackson et al., 2005; Buckholtz et al., 2008; Shenhav and Greene, 2014). However, these studies did not elucidate the functional contribution(s) of each brain region to harm or mental state evaluation, and it remains unclear how and where these components integrate. Prior studies have pinpointed activation in the dorsolateral prefrontal cortex (DLPFC), medial prefrontal cortex (MPFC), and posterior cingulate cortex (PCC) at the time of decision-making, suggesting that these regions may support the integration of mental state and harm (Buckholtz and Marois, 2012; Buckholtz et al., 2015), an argument buttressed by reports that MPFC and PCC may act as cortical "hubs" of information processing (Sporns et al., 2007; Buckner et al., 2009), although these studies could not dissociate integration from other task components. Finally, a debate persists about the specific role of the DLPFC in human punishment behavior. Although some studies have associated DLPFC with implementation of cognitive control (Sanfey et al., 2003; Knoch et al., 2006; Haushofer and Fehr, 2008; Tassy et al., 2012), we have claimed that the region acts as a superordinate node that supports the integration of signals to select the appropriate punishment decision (Buckholtz et al., 2008, 2015; Treadway et al., 2014).

The present study addresses these open questions by means of a novel experimental design. Specifically, the present design (1) independently and objectively parameterizes both the mental state and harm factors while (2) controlling information presentation in a manner allowing segregation of the evaluative, integrative, and response decision components of third-party punishment decision-making. We achieved the first element of the design by using harm levels based on independent metrics and mental state levels based on the Model Penal Code's hierarchy of mental state culpability (spanning blameless, negligent, reckless, knowing, and purposeful) (Simons, 2003; Shen et al., 2011). To achieve the second element, trials were divided into distinct sequential segments (context presentation, followed by harm and mental state evaluations, followed by response decision), each separated from the others by an arithmetic task to limit cognitive processes to their respective stimulus presentations. Together, these manipulations permit the isolation of brain mechanisms involved in the harm and mental state evaluative processes, in the integration of these evaluative processes, and in the use of this information in selecting an appropriate punishment.

Materials and Methods

Subjects. Twenty-eight right-handed individuals (13 females, ages 18–35 years) with normal or corrected-to-normal vision consented to participate for financial compensation. The Vanderbilt University Institutional Review Board approved the experimental protocol, and subjects provided their informed consent. Five subjects were not included in the analysis: two did not complete the scan due to discomfort with the MRI pulse sequences; two had excessive motion (>3 mm translation or 3 degrees of rotation) during the MRI scanning; and one failed to follow task instructions. That left 23 subjects (11 females, ages 18–35 years) for the analysis.

Paradigm. In this fMRI experiment, subjects participated in a simulated third-party legal decision-making task in which they determined the appropriate level of hypothetical punishment for the actions of a fictional protagonist ("John") described in short written scenarios. Participants were instructed to treat each scenario independently. The study improved on prior work in two principal ways: (1) by separating in time the cognitive processes of evaluating the harm and mental state components of the scenarios, the integration of these components, and the rendering of a punishment decision; and (2) by independently and objectively manipulating both the mental state of the actor and the resulting harm of the actor's conduct in a parametric fashion.

With regard to the first objective of the experimental design, in contrast to prior studies (Buckholtz et al., 2008, 2015; Treadway et al., 2014) in which all components of each scenario were presented at once, components of each scenario were presented in distinct temporal stages, with each of the four stages separated from the others by a variable interstage interval (ISI) drawn from an exponentially decaying distribution of ~3–10 s (Fig. 1). Stage A contained an introductory sentence describing the context in which the protagonist acted. Stages B and C each presented a sentence with either the harm or the mental state, respectively. The order in which Stages B and C appeared (harm then mental state, or mental state then harm) varied by trial within subject. Finally, Stage D presented the punishment scale on which subjects based their punishment decision and selected a punishment response by button press. Participants were instructed to make their response as soon as they had made a decision but instructed not to rush (they had up to 16 s to make their response).

Several details of the experiment were designed to optimize the likelihood that a given cognitive process occurred at a specific stage. First, to constrain the subjects' cognitive processing of each sentence to its presentation time and to preclude subjects from using the ISIs to ponder the appropriate response, the ISIs were filled with a secondary math task that lasted the duration of each ISI. Each math problem started 200 ms after a stage's end and included a series of addition or subtraction operations on integers between 1 and 9, with a solution between 0 and 9. The number of operations scaled with the ISI length. All integers and operations were individually presented at the center of the screen, changing at a rate of 1 item per 750 ms and followed at the end by "=," indicating that the subject should provide a response within 2 s. If no response was provided,

The authors declare no competing financial interests.

This article is freely available online through the *J Neurosci* Author Open Choice option.

Correspondence should be addressed to the following: Legal inquiries should be addressed to either of the following: Matthew R. Ginther, Neuroscience Graduate Program, Vanderbilt University, 111 21st Avenue South, Nashville, TN 37203, E-mail: ginther@gmail.com; or Prof. Owen D. Jones, Departments of Law and Biological Sciences, Vanderbilt University, 121 21st Avenue South, Nashville, TN 37203, E-mail: owen.jones@vanderbilt.edu. Neuroscientific inquiries should be addressed to either of the following: Matthew R. Ginther, Neuroscience Graduate Program, Vanderbilt University, 111 21st Avenue South, Nashville, TN 37203, E-mail: ginther@gmail.com; or Dr. René Marois, Department of Psychology, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37240. E-mail: rene.marois@vanderbilt.edu.

DOI:10.1523/JNEUROSCI.4499-15.2016

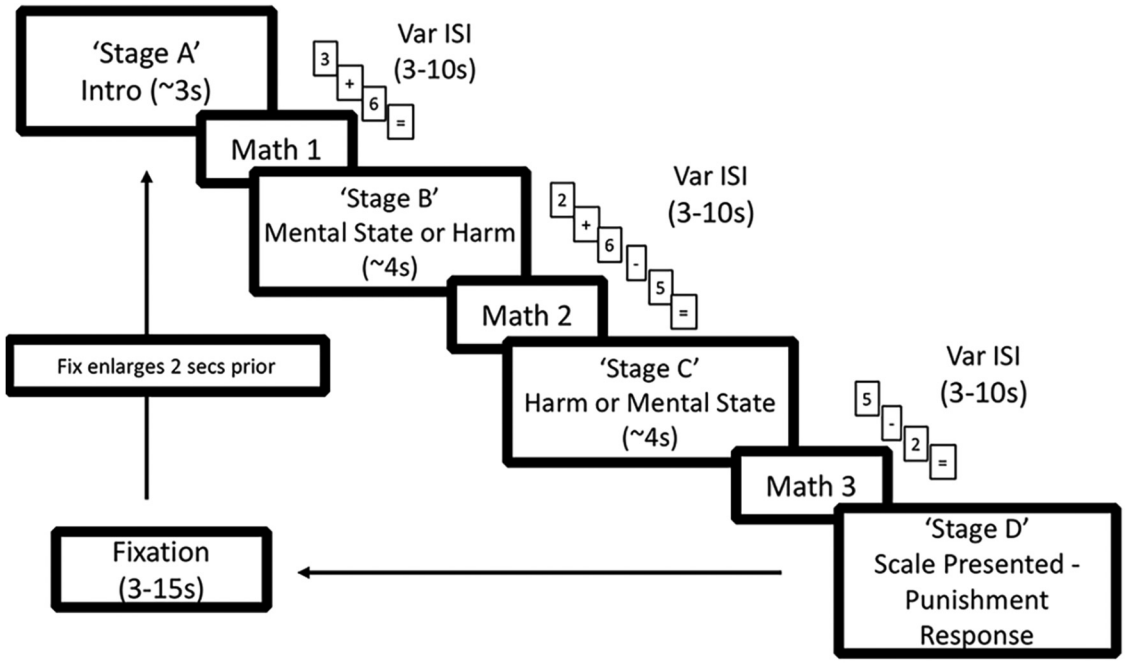

Figure 1. Timeline of a single trial. Each round began with the presentation of Stage A as an RSVP of words, which only contained introductory information about the scenario. Following Stage A, subjects were presented with the first of three intervening math tasks, which spanned the durations of each ISI. Subjects were then presented Stage B and Stage C, which contained the harm and mental state information, respectively, in randomized order. In the last stage (Stage D), subjects were probed for their punishment response. The variable ITI lasted for a duration of 3–15 s, with the last 2 s accompanied by a larger fixation square. Var, Variable.

the task continued as if a response had been made. For example, a 3 s ISI would consist of two integers and one operation (e.g., 3, +, 5) for 2 s plus an average of a 1-s-long response time. A small white fixation square (0.25° of visual angle) appeared following the subject's response.

Second, to help ensure that subjects were only processing the information presented during each of Stages A–C (and not using some of that time cogitating about a previous stage's information), we presented the scenarios as a rapid serial visual presentation (RSVP), wherein a given stage's words were presented sequentially at the center of the screen at the rate of 6 words per second (rather than being presented simultaneously in a full sentence) and followed immediately by the ISI. This rate of word presentation was selected because it does not reduce subjects' reading comprehension (Castelhano and Muter, 2001). We controlled for word length across harm and mental state sentences, as well as across the different harm levels, with an average scenario length of 77 words (SD of 4). Because the rate of word presentation was fixed and the duration of each stage was a function of the word length, stimulus duration was thus controlled for as well.

Third, to delay the time of the subject's punishment judgment until the decision stage, on each trial we randomly presented subjects at the decision stage with one of several available punishment scales. Overall, there were 10 different punishment scales: one "master" scale and nine derivative scales. The master scale, which spanned the entire range of possible punishments, was anchored such that 0 = no punishment, 3 = 1 day in jail, 6 = 1 year in jail, and 9 = most severe punishment that the subject personally endorsed. The nine derivative scales essentially "zoomed in" on a part of the master scale and remapped the 0–9 response space accordingly. As an example, a derivative scale may look as follows: 0 = 1 day in jail, 6 = 1 year in jail, and 9 = most severe punishment. For any given scenario, 6 of the 10 scales were available as possible options, with 1 of the 6 randomly selected for any given trial. The 6 scales per scenario were selected so as to ensure, based on pilot data, that the mean punishment response ± 2 SDs fell within the confines of the scale. Thus, we nearly guaranteed the available scale included the desired punishment response for each scenario. For analysis purposes, we algebraically converted the responses provided on the derivative scales to the equivalent response on the master scale (e.g., if a subject responded 0 on the derivative scale presented above, it was coded as a 3).

The data indicate that our efforts were largely successful in delaying subjects' punishment decisions to Stage D. First, pilot data showed a substantial increase in the amount of time subjects spent at the final stage (mean ± SD, 4.02 ± 1.84) compared with when that stage was not preceded by the ISI math task and RSVP format and did not include shifting scales, but did segregate the task stages (2.45 ± 2.09). Second, at the time of the decision, the distribution in reaction times (RTs) was not uniform across levels of harm or mental state, as one would expect if subjects had made their decisions before Stage D. Instead, there is a significant effect of both mental state and harm level on subject RT (Fig. 2*B*,*C*).

Following the subjects' response, an intertrial interval (ITI) drawn from a decaying exponential distribution from 3 to 15 s began. The small white fixation square was presented for the duration of the ITI, except that it was enlarged (to 0.49° of visual angle) for the last 2 s of the ITI to signal to the participants the imminent start of the next trial (for trial design, see Fig. 1).

To achieve the second principal experimental objective (independent and objective manipulation of the mental state and harm components in a parametric fashion), our scenarios parametrically manipulated the mental state of the actor using 4 of the 5 Model Penal Code categories. These are (in descending order of intentionality) purposeful (P), reckless (R), negligent (N), and blameless (B) (knowing was not included here because of subjects' difficulty in distinguishing this category from reckless in behavioral studies) (Shen et al., 2011; Ginther et al., 2014). The harm resulting from the actor's actions also varied parametrically in four categories, ranging from *de minimis* (no or insubstantial harm), to substantial (but impermanent), permanently life altering, and, finally, death. In figures, we categorize these as Harm 1–4.

Some of the scenarios were based upon scenarios used in previous research (Shen et al., 2011), whereas others were crafted for this study. The complete scenario set is available from the authors. Individual scenarios were derived from 64 distinct "themes." Each theme contained a unique set of contextual facts and the eventual harm. The severity of each harm fell into one of the four distinct categories described earlier, and there were 16 themes for each level of harm. In a pilot experiment, we had 23 online subjects rate the severity of the harm sentences alone (on a 0–9 scale) to fine-tune and verify our categorization of the scenarios along the

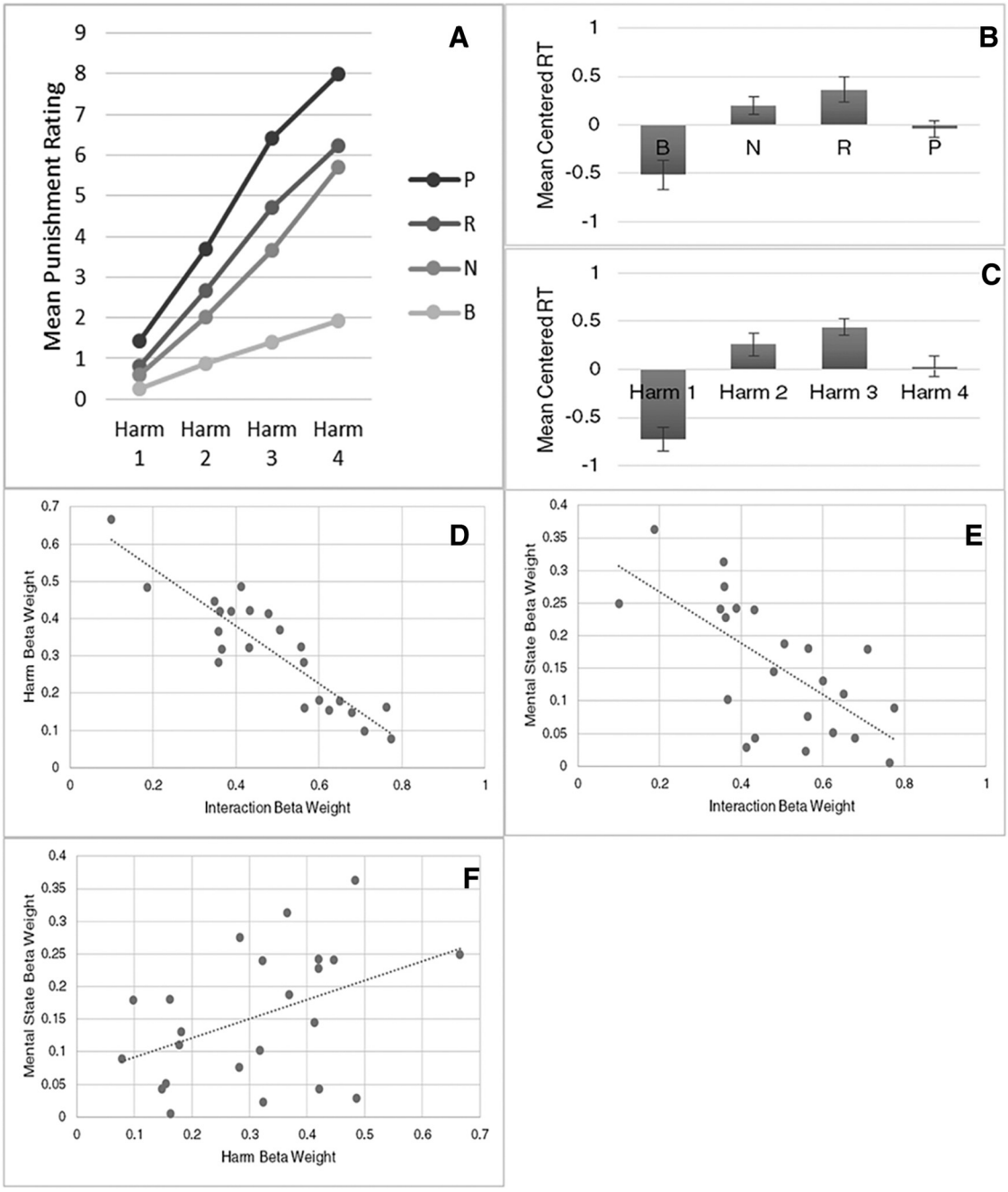

Figure 2. ***A***, Mean punishment ratings as a function of mental state and harm level. ***B***, ***C***, Mean centered RT as a function of mental state and harm level. Error bars indicate ± SEM. ***D***, Subjects' punishment ratings are primarily determined by the product of the harm × MS interaction term and the harm term. Subjects' weightings of these two terms show a strong negative correlation. ***E***, There is a negative correlation between subjects' weightings of the MS × harm interaction and the mental state term. P, Purposeful; R, reckless; N, negligent; B, blameless. ***F***, There is a positive correlation between subjects' weighting of the mental state and harm terms.

four harm levels (mean ± SD: Harm 1, 1.49 ± 0.29; Harm 2, 3.67 ± 0.50; Harm 3, 6.13 ± 0.37; Harm 4, 8.64 ± 0.24). These subjects were recruited using Amazon's Mechanical Turk, which provides a sample of high-quality participants largely representative of the population (Rand, 2012). Within each theme, the scenarios also varied the mental state of the protagonist across four possible levels of mental state (Table 1).

The levels of the 2 factors were orthogonal to one another such that, on any given trial, the harm level did not predict the mental state level, or vice versa. The 64 different themes, 4 levels of mental state, and 2 possible orderings (harm first or mental state first) yielded a total of 512 different possible scenarios (64 × 4 × 2), 64 of which were presented to each subject in pseudorandomized fashion. Each subject saw a single scenario from each theme, and all scenario conditions were balanced within each subject: that is, subjects saw 4 scenarios in each mental state (4 levels) × harm (4 levels) cell in the factorial design. An example of a single theme and the 8 derivative scenarios is presented in Table 1. Details of the text could change for a given cell (e.g., see reckless mental state) depending on its order of presentation to increase both its believability and comprehensibility.

Because of the complexity and novelty of the current paradigm, we first assessed whether it would yield similar punishment responses to those acquired when each scenario was presented in its entirety in the same frame (Buckholtz et al., 2008; Treadway et al., 2014). This possibility was tested by recruiting 20 subjects to complete the third-party punishment task online by means of Amazon's Mechanical Turk. These subjects were presented with scenarios in their complete paragraph form

Table 1. Example of one theme and the multiple derivative scenarios[a]

Illustrative theme (planks and bikes): four "mental state first" variations			
Introductory sentence: John is hauling planks to his cabin because he is in the middle of doing carpentry work on his house, which abuts a public mountain bike trail			
Purposeful mental state: Angry with the mountain bikers for making too much noise when biking past his house, John desires to injure some bikers by dropping planks on their trail so that they would hit them	*Reckless mental state*: John drops some planks onto the trail without retrieving them because he's in a rush, even though he is aware there is a substantial risk bikers will hit them and be injured	*Negligent mental state*: While John is carrying planks to his workshop in order to begin building new steps for his house, he drops some of the wood planks onto the bike trail without even noticing	*Blameless mental state*: While John is carefully carrying some planks from his shed to the backyard, an unexpectedly strong gust of wind causes John to inadvertently drop several planks, despite his best efforts not to
Harm sentence: Soon after John drops the planks, two bikers pass by and they hit the planks, which causes them to flip over their handlebars and one of the bikers suffers serious injuries as a result			
Illustrative theme (planks and bikes): four "harm first" variations			
Introductory sentence: John is hauling planks to his cabin because he is in the middle of doing carpentry work on his house, which abuts a public mountain bike trail			
Purposeful mental state: Angry with the mountain bikers for making too much noise when biking past his house, John had desired to injure some bikers by dropping planks on the trail so that they would hit them	*Reckless mental state*: John had dropped some planks onto the trail without retrieving them because he was in a rush, even though he was aware there was a substantial risk some bikers would hit them and be injured	*Negligent mental state*: While John was carrying planks to his workshop in order to begin building new steps for his house, he had dropped some of the wood planks onto the bike trail without even noticing	*Blameless mental state*: While John was carefully carrying planks from his shed to the backyard, he slipped on some mud, which caused him to unknowingly drop several planks, despite his best efforts not to
Harm sentence: Soon after John crosses the trail, two bikers pass by and they hit planks that John dropped onto the trail, which causes them to flip over their handlebars and one of the bikers suffers serious injuries as a result			

[a]Subjects evaluated only 1 of the possible 8 scenarios for each theme.

Table 2. Performance of seven different models of subjects punishment decisions: behavioral modeling for the fMRI experiment

Model	AIC	Model components	Beta	SE	*p*
1	7962	Mental state	0.45	0.02	0.000
2	7842	Harm	0.60	0.02	0.000
3	7659	Mental state × harm	0.75	0.03	0.000
4	7673	Mental state +	0.45	0.02	0.000
		Harm	0.60	0.02	0.000
5	7637	Harm +	0.20	0.03	0.000
		Mental state × harm	0.63	0.02	0.000
6	7660	Mental state +	−0.04	0.03	1.000
		Mental state × harm	0.78	0.02	0.000
7[a]	7631	Mental state +	0.15	0.03	0.005
		Harm +	0.30	0.03	0.000
		Mental state × harm	0.47	0.04	0.000

[a]Model 7 selected as the best model by means of AIC. All beta coefficients standardized.

in a single frame and subjects read at their own pace. There was no statistical difference in punishment ratings between these subjects and the participants who completed the present experiment ($F_{(1,41)} = 1.41$, $p = 0.241$).

Scenarios were presented in pseudorandomized fashion, ensuring that, in each 16 trial fMRI run, subjects rated the punishment for one scenario in each cell of the 4 mental state × 4 harm-level design. The runs varied in duration given the variable response times but never lasted >11.5 min. Each subject completed 4 of these fMRI runs. The experiment was programmed in MATLAB (MATLAB, RRID:SCR_001622) (The MathWorks) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997) (Psychophysics Toolbox, RRID:SCR_002881). Subjects were positioned supine in the scanner to be able to view the projector display using a mirror mounted on the head coil. Manual responses were recorded using two 5-button keypads (Rowland Institute of Science).

Statistical analysis: behavioral data. We analyzed trial-wise punishment responses by testing a family of multiple linear regression models by means of a mixed-effects model, treating subject as a random factor. We analyzed 7 models, consisting of all combinations of the mental state (0 = blameless, 1 = negligent, 2 = reckless, 3 = purposeful), harm (0 = *de minimis*, 1 = substantial, 2 = life altering, and 3 = death), and interaction components (Table 2). Models were assessed using the Akaike Information Criterion (AIC), which quantifies both model fit and simplicity. Although AIC scores constitute a unitless measure, a relatively lower AIC score reflects a more accurate and generalizable model. Subject parameters used below are estimated using the best model as identified by AIC score.

fMRI data acquisition. Our imaging pulse sequences and image acquisition followed conventional methods. All fMRI scans were acquired using a 3T Philips Achieva scanner at the Vanderbilt University Institute of Imaging Science. Low- and high-resolution structural scans were first acquired using conventional parameters. Functional BOLD images were acquired using a gradient-EPI pulse sequence with the following parameters: TR 2000 ms, TE 35 ms, flip angle 79°, FOV 192 × 112 × 192 mm, with 34 axial slices (3.0 mm, 0.3 mm gap) oriented parallel to the AC-PC line and collected in an ascending interleaved pattern (T2*-weighted).

Statistical analysis: fMRI data. Image analysis was conducted using Brain Voyager QX 2.8 (BrainVoyager QX, RRID:SCR_013057) (Brain Innovation) in conjunction with custom MATLAB software (The MathWorks). All images were preprocessed using slice timing correction, 3D motion correction, linear trend removal (1/128 Hz), temporal high pass filtering, and spatial smoothing with a 6 mm Gaussian kernel (FWHM) as implemented through Brain Voyager software. Spatial smoothing was omitted for data analyzed using multivariate techniques. Subjects' functional data were aligned with their T1-weighted anatomical volumes and transformed into standardized Talairach space.

We created design matrices for each subject by convolving the task events with a canonical hemodynamic response function (double gamma, including a positive γ function and a smaller, negative γ function to reflect the BOLD undershoot). For the task events, the presentation of each stage of a scenario was modeled as a boxcar function spanning the duration of the stage's RSVP. The punishment decision phase of the task was modeled from the display of the punishment scale to the time of response. The interstimulus math task was modeled from the start of the ISI to the time of subject response. We also inserted 6 estimated motion parameters (X, Y, and Z translation and rotation) as nuisance regressors into each design matrix.

For our first-level analysis of the functional imaging data, we created 6 distinct GLMs for each subject's data, with each GLM created to address a different question and avoid colinearity issues between regressors. Specifically, to assess the evaluative process for harm and mental state separately, the first GLM (GLM1) modeled each stage of the task as well as the interstimulus math task, with the identification of Stage B and Stage C classified as either mental state or harm based on which occurred at that stage on that trial. To model the cognitive systems recruited by the different task stages, regardless of the information presented at the stage, we created GLM2, which was the same as GLM1, except that we did not reclassify Stage B and Stage C into mental state and harm. To identify regions sensitive to the different harm levels, the third GLM (GLM3)

modeled only the harm component, but with different regressors for each level of harm in the sentence. The fourth GLM (GLM4) did the same level-based regressor analysis for mental state. To identify regions that are sensitive to the integration of harm and mental state, the fifth GLM (GLM5) modeled Stage C only, categorizing the stage both in terms of whether the scenario had a culpable (P, R, or N) or blameless (B) mental state and whether the harm contained was high (life altering/death) or low (*de minimis*/substantial). We designed GLM5 to contain 4 cells to maximize the number of trials per cell so as to assure a more reliable estimate of the condition parameter for each subject. We divided the mental state conditions into blameless and culpable (the latter of which combines the purposeful, reckless, and negligent mental states) because that reflects the most meaningful legal demarcation in our conditions. For the harm condition, we performed a median split such that we had high- and low-harm conditions. We achieved qualitatively similar results if we demarcated the mental state using a median split of conditions as well. We modeled only Stage C for GLM5 because this is the first stage at which the integration of harm and mental state could occur.

All GLMs were created using *z*-transformed time course data. Second-order random-effects analyses were conducted on the β weights calculated for each subject. To control for multiple comparisons when performing whole-brain analyses, we applied a False Discovery Rate (FDR) threshold of $q < 0.05$ (with $c(V) = 1$) and a 10 functional voxel cluster size minimum. In the case a conjunction analysis was used, we applied a minimum test statistic (Nichols et al., 2005). For visualization purposes, some analyses display BOLD signal time courses extracted using a deconvolution analysis. For this analysis, we defined a set of 10 finite impulse response (FIR) regressors for each condition and ran first-level region of interest (ROI) GLMs using the FIR regressors. Although we display SEs of the mean for these time courses, these are strictly for the purpose of visualizing the variance and shape of the hemodynamic responses. To avoid nonindependent selective analysis of the data (the "double-dipping" problem), these time course data were not subjected to inferential statistical analyses. When we perform *post hoc* analyses on regions identified in the whole-brain analyses, we control for multiple comparisons again using a FDR threshold of $q < 0.05$.

For the multivoxel pattern analysis (MVPA), *z*-transformed BOLD signals at each time point for each condition were extracted and activity was centered as a function of condition such that there was no longer a mean univariate difference between event types. Independently for each ROI, subject, and time point, we performed a leave-one-run-out procedure: all but one run of data were used to train a linear support vector machine (Chang and Lin, 2001) (LIBSVM, RRID:SCR_010243) that was then tested on the held-out run; this process was iterated until all runs had served as the test data once (4-fold cross-validation). Classifier proportion correct was aggregated to determine an ROI-, subject-, and time point-specific MVPA result. Within an ROI, MVPA results across time points were concatenated to form an ROI- and subject-specific event-related MVPA (er-MVPA) time course (Tamber-Rosenau et al., 2013) with perfect performance at 1.0. The set of subject er-MVPA time courses was compared with chance at the mean peak time point across ROIs via a one-tailed *t* test (because below-chance classification is not interpretable). The peak time point occurred 12 s after the decision prompt or 10 s after the start of the stage RSVP, which corresponds, on average, to 6 s following the mean decision time and the end of the stage RSVP, respectively. Whole-brain searchlight analysis was performed only at the peak time points due to practical computation limitations. For the searchlight analysis, we defined a spherical 3 mm region extending from every cortical voxel and performed the same MVPA procedure described above in each subject and in each of these spherical regions across the brain. As with the whole-brain univariate inquiries, we performed an FDR ($q < 0.05$) correction for multiple comparisons. Chance MVPA performance was empirically estimated for each analysis to rule out artifactual above-chance performance (as a result of, for instance, imperfect balance of number of correct trials of each type per run). We achieved this by running 200 iterations of the classifier on data using randomly shuffled condition labels for the training set. Because of practical limitations, we used the mean chance performance calculated on the ROI-based MVPA as chance for the searchlight analysis.

Results

Behavioral results

Figure 2*A* shows subjects' punishment ratings as a function of both harm and mental state levels. Using a repeated-measures ANOVA, the results indicate main effects of both the actor's mental state ($F_{(3,66)} = 199.46$, $p < 0.001$) and the resulting harm ($F_{(3,66)} = 414.90$, $p < 0.001$) on punishment ratings. There was also an interaction between the levels of harm and mental state ($F_{(9,198)} = 22.096$, $p < 0.001$), such that the increase in punishment ratings with higher harm levels is greater under more culpable states of mind. This interaction is present even when the blameless condition is excluded from the analysis ($F_{(6,144)} = 3.84$, $p < 0.005$).

Figure 2*B*, *C* shows subjects' mean RTs at the decision phase as a function of mental state and harm levels, respectively. Both mental state and harm level display a quadratic relationship with RT, wherein the intermediate levels of mental state and harm are more time-consuming for subjects at the decision stage than the extreme levels of mental state and harm (Fig. 2*B*,*C*). We explicitly tested this relationship by means of a repeated-measures ANOVA with within-subjects quadratic contrasts for both mental state ($F_{(1,22)} = 19.87$, $p < 0.001$) and harm ($F_{(1,22)} = 26.65$, $p < 0.001$).

To understand the contributions of harm and mental state and the interaction of these two factors in punishment decision-making, we compared behavioral models that could ostensibly account for how individuals weigh and integrate these factors in their decisions. As displayed in Table 2, the model with harm, mental state, and interaction components was identified as the best model using AIC. The standardized model parameters indicate that, by a large margin, subjects weight the interaction component most heavily in their punishment response, followed by harm and then mental state. As seen in Figure 2*A*, the nature of this interaction is a superadditive effect between mental state and harm. Mean r^2 across subjects using the selected model was 0.66. The importance of the interaction of harm and mental state in punishment decisions is also illustrated by a regression analysis of individual subjects' weighing of each of the three components. Specifically, the most heavily weighted component, the interaction, displayed a strong negative correlation with both harm ($r = -0.90$, $p < 0.0001$; Fig. 2*D*) and mental state ($r = -0.67$, $p = 0.0005$; Fig. 2*E*), whereas harm and mental state showed a positive correlation ($r = 0.43$, $p = 0.041$; Fig. 2*F*). These results suggest that subjects who tend to weigh heavily the interaction term in their punishment decisions do not put much weight on the harm or mental state components alone.

fMRI data

The analysis of the imaging data was directed at addressing three primary questions. First, to what extent do mental state and harm evaluation engage separable or common neural processes? Second, what regions support the integration of these two components? Third, is the punishment decision neurally separable from harm/mental state evaluations and, to the extent that it is, what brain regions are associated with it?

fMRI data: evaluation of mental state and harm information

Identified here are those regions that show preferential engagement for the evaluation of the mental state component and, subsequently, those regions that show preferential engagement for

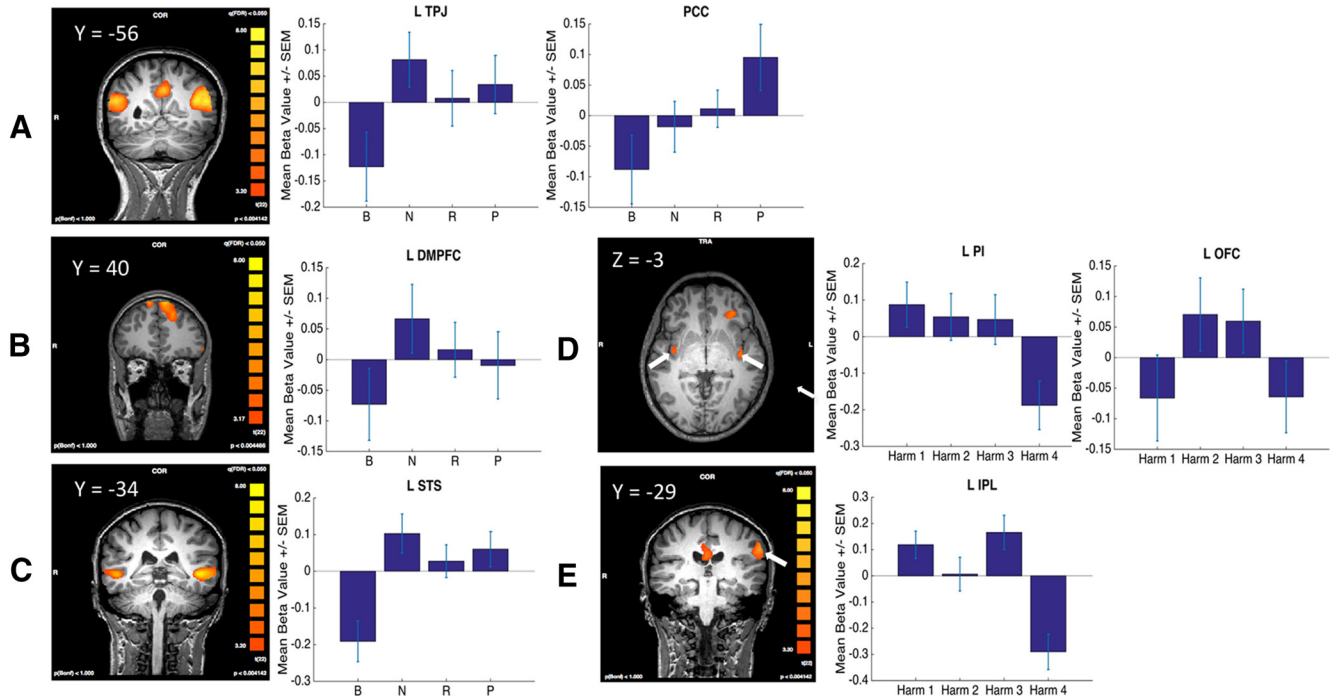

Figure 3. ***A–C***, Left, SPM results of the contrast mental state − harm, highlighting. TPJ and PCC (***A***), DMPFC (***B***), and STS (***C***). Right, Activity in the respective ROIs (when the ROI is bilateral, we only show the left) as a function of mental state level. ***D***, ***E***, Left, SPM results of the contrast harm − mental state illustrating PI and left OFC (***D***) and left IPL (***E***). Right, Activity in the respective ROIs as a function of harm level.

Table 3. Regions showing significant activation for mental state evaluation as contrasted with harm evaluation[a]

	Talairach coordinates						Linear contrast		Contrast with MS difficulty		MS decoding	
Region	*X*	*Y*	*Z*	*t*	*p*	Size	*F*	*p*	*F*	*p*	*t*	*p*
R middle temporal gyrus	50	−35	−3	6.60	1.0E-6	81	0.00	1.00	0.21	0.47	−1.83	0.21
R TPJ	50	−53	18	8.10	<1.0E-6	275	0.69	0.34	2.12[c]	0.08[c]	1.71	0.21
R STS	53	−32	−1	6.59	<1.0E-6	77	0.01	1.00	0.29	0.64	0.24	0.94
PCC	−4	−56	30	7.01	<1.0E-6	221	7.14[c]	4.8E-3[c]	1.73	0.10	0.12	0.94
R caudate	8	4	18	4.47	1.9E-4	13	0.09	1.00	0.12	0.53	−0.49	0.93
R DMPFC	11	37	51	5.84	7.0E-6	17	0.44	0.48	3.39[c]	0.05[c]	1.82	0.21
L DMPFC	−7	41	51	7.03	<1.0E-6	620	0.30	0.62	2.30[c]	0.08[c]	−3.06	0.08
L medial frontal gyrus	−4	−17	54	4.21	3.6E-4	20	1.50	0.15	0.71	0.22	−0.39	0.93
L caudate	−16	4	15	5.01	5.2E-5	52	0.35	0.56	0.16	0.51	−2.63	0.10
L IFG	−46	28	−3	6.98	1.0E-6	50	7.19[b]	4.6E-3[b]	8.34[c]	7.6E-3[c]	−1.66	0.21
L STS	−52	7	−22	11.47	<1.0E-6	266	8.20[b]	2.7E-3[b]	13.09[c]	1.5E-3[c]	−1.61	0.21
L TPJ	−43	−59	21	9.13	<1.0E-6	473	2.17[b]	0.09[b]	4.16[c]	0.04[c]	−0.08	0.94

[a]Whole-brain contrast corrected at q(FDR) = 0.05. Linear contrast column presents results of repeated-measures ANOVA with a linear contrast. Contrast with MS difficulty column presents the results of a repeated-measures ANOVA with a contrast based on mental state difficulty (Ginther et al., 2014; Shen et al., 2011). MS decoding column presents the results of a *t* test compared with chance level decoding of mental state level in each region. All sizes are in units of functional voxels. All ROI analyses corrected for multiple comparisons.

[b]Significance at $p < 0.1$.

[c]If both contrasts account for the data, significantly more consistent with the data than the other contrast (Rosnow and Rosenthal, 1996).

the harm component. In both cases, the initial region identification is followed by analyses that seek to provide supporting evidence for the involvement of the identified brain regions in the evaluation of that component and to characterize the nature of that region's involvement.

To identify regions preferentially involved in mental state evaluation, we performed a contrast of mental state evaluation > harm evaluation using GLM1 (which modeled all stages, with Stage B and Stage C collapsed across either mental state or harm, although we achieved qualitatively similar results when mental state or harm activity was solely derived from Stage B). The resulting statistical parametric map (SPM) revealed areas of differential activation in regions associated with a Theory of Mind (ToM) network thought to be involved in interpreting others' minds (Gallagher and Frith, 2003; Carrington and Bailey, 2009), including bilateral TPJ, bilateral dorsomedial prefrontal cortex (dmPFC), and bilateral STS (Fig. 3*A–C*, left; Table 3), as well as PCC (Fig. 3*A–C*, left; Table 3). We also observed activations in a number of other regions not commonly associated with a ToM network, including bilateral caudate, right middle temporal gyrus, left medial frontal gyrus, and left inferior frontal gyrus (IFG) (Table 3).

In each identified ROI, the relationship between the level of mental state and brain activity was further characterized by considering three possibilities: (1) activity in the region is linearly related to the level of mental state, consistent with the commensurate increase in punishment amount seen with increases in the level of mental state; (2) activity in the region is related to the

Table 4. Regions showing significant activation for harm evaluation as contrasted with mental state evaluation[a]

	Talairach coordinates						Linear contrast		Difficulty effect		Death condition significantly lower		Harm decoding	
Region	*X*	*Y*	*Z*	*t*	*p*	Size	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
R LPFC	41	34	18	5.71	1.0E-5	146	20.02[c]	8.7E-5[c]	0.95	0.25	18.74[b]	4.9E-5[b]	1.29	0.37
R PI	38	−8	−6	5.53	1.5E-5	15	7.55[b]	5.4E-3[b]	1.10	0.25	8.68[b]	3.0E-3[b]	2.21	0.26
Corpus callosum	−1	−32	24	5.10	4.2E-5	99	0.22	0.90	1.51	0.21	0.01	1.00	−0.03	0.98
L OFC	−28	34	−4	6.06	4.0E-6	15	0.00	1.00	4.66[c]	0.04[c]	1.51	0.18	−1.76	0.26
L PI	−40	−11	−3	5.17	3.5E-5	24	11.90[b]	1.0E-3[b]	3.46[b]	0.07[b]	16.14[c]	1.1E-4[c]	−0.90	0.53
L fusiform gyrus	−52	−53	−6	5.72	9.0E-6	30	10.79[b]	1.3E-3[b]	7.69[b]	0.01[b]	23.44[c]	1.1E-5[c]	−0.37	0.83
L IPL	−62	−29	33	5.61	1.2E-5	64	18.09[b]	9.8E-5[b]	9.41[b]	0.01[b]	35.74[c]	<1.0E-6[c]	1.67	0.26

[a]Whole-brain contrast corrected at q(FDR) = 0.05. Linear contrast column presents results of repeated-measures ANOVA with a linear contrast. Difficulty effect column presents the results of a repeated-measures ANOVA with a quadratic contrast as a proxy of harm evaluation difficulty. Death condition significantly lower column presents the results of a repeated-measures ANOVA with the contrast [−1, −1, −1, 3]. Harm decoding column presents the results of a *t* test compared with chance level decoding of harm level in each region. All ROI analyses corrected for multiple comparisons.

[b]Significance at $p < 0.1$.

[c]If more than one contrast accounts for the data, contrast accounts for significantly more of the variance in the data than the other two contrasts (Rosnow and Rosenthal, 1996).

difficulty subjects have in evaluating the offender's state of mind, reflecting demand or time-on-task effects; and (3) each mental state is coded by a distinct pattern of neural ensembles within a given brain region rather than by the overall level of activation of that region.

To examine the extent to which the mental state activations were consistent with the linear and/or difficulty-based models, we ran a repeated-measures ANOVA on β parameters extracted using GLM4 (which modeled the different mental state levels, collapsed across Stage B and Stage C), using both a simple linear contrast and a contrast based on mental state evaluation difficulty. The latter was based on subjects' difficulty in classifying different mental states as belonging to each P, R, N, and B categories as assessed in prior studies from our group (Shen et al., 2011; Ginther et al., 2014). Specifically, we defined difficulty as 1-classification accuracy to arrive at the following difficulty values: P: 0.22, R: 0.60, N: 0.52, B: 0.12. (The quadratic fit of the classification accuracy data is similar to the RT data at response time for mental states; Fig. 2*B*). We chose to use the former fit for the fMRI data because it more likely reflects the process that is taking place at the evaluative than at the decisional stages. However, the results are similar if RTs are used. This pair of analyses tested whether either model significantly accounted for the data. If a region was sensitive to both contrasts, we examined whether one of the contrasts accounted for significantly more of the variance in the data (Rosnow and Rosenthal, 1996). In a final analysis, MVPA was used to assess whether distinct neural ensembles in the identified ROIs encoded the different mental state levels by training and testing a support vector machine on brain activity during the period of evaluation. For all MVPA analyses, univariate differences were first subtracted out (see Materials and Methods) so that the analysis was specific for multivariate patterns.

As displayed in Table 3 and visualized in Figure 3*A–C*, TPJ, STS, and DMPFC, the regions comprising the putative ToM network (TPJ, STS, DMPFC), are accounted for by the difficulty model with the exception of right STS. Other than left IFG, no other region showed activity consistent with the mentalization difficulty model. By contrast, the linear model better accounted for the activation profile in the PCC (Table 3; Fig. 3*A*). Finally, we did not find above-chance levels of classification accuracy in any of the identified ROIs (Table 3). Together, these results suggest that regions engaged by the evaluation of mental state show patterns of activations consistent with both an effect of mentalization difficulty in the case of TPJ, STS, and DMPFC, and with the amount of culpability in the case of the PCC.

The same set of analyses was performed to identify regions that may be implicated in the evaluation of harm. We again used GLM1 to identify regions displaying greater activity for the harm evaluation compared with the mental state evaluation by means of the reverse contrast from the prior analysis (harm evaluation > mental state evaluation). This analysis identified bilateral posterior insula (PI), the left inferior parietal lobule (IPL), the left orbitofrontal cortex (OFC), left fusiform gyrus, and left lateral prefrontal cortex (LPFC) as showing preferential engagement for evaluation of harm statements (Fig. 3*D*,*E*, left; Table 3).

In each of these regions, we next characterized the relationship between the different categories of harm and neural activity. As with mental state, both a linear and quadratic relationship were considered, consistent with the commensurate increase in punishment and evaluation difficulty, respectively, as well as the possibility that MVPA would reveal distinct patterns of neural ensembles for each harm level. Because we did not have an independent measure of evaluation difficulty as a function of harm level, we used a quadratic ([1, −1, 1, −1]) pattern under the premise that intermediate harms are more difficult to evaluate than harms at the boundary, a pattern that is consistent with the RT distribution at the time of decision. As with mental state, we achieve qualitatively similar results if we use a contrast based on decision RT.

We compared how well these three potential relationships explained the pattern of activation in each harm ROI. Activity in the OFC was best accounted for by the quadratic relationship ("Difficulty effect") such that there was greater activation for the intermediate harms than the extreme harms (Fig. 3*D*; Table 4), whereas right lateral prefrontal cortex activity was best accounted for by a negative linear contrast (Table 4). As with mental state, we used MVPA to examine whether the identified regions displayed distinct patterns of activation as a function of the level of harm and found no evidence that they did (Table 4). Thus, only two of the harm ROIs exhibited any of the predicted functional relationships. Most of the other ROIs, namely bilateral PI, left IPL, and left fusiform gyrus, showed an unexpected activity pattern in which the highest category of harm, death, exhibited less activity than the three other harm levels (Fig. 3*D*,*E*; Table 4). We speculate that this pattern may reflect vicarious somatosensation of pain (Rozzi et al., 2008; Singer et al., 2009; Keysers et al., 2010) in which representations of others' pain or bodily harm can be imagined in all harm levels except death.

Directly contrasting harm and mental state does not identify brain regions that may be commonly activated by the evaluation of the two components. To identify commonly recruited regions, we performed a conjunction analysis of contrasts that removed activity related to reading and comprehending text (by subtracting Stage A) and any potential decision-related activity (by subtracting the decision stage): 1, mental state > Stage A; 2, harm >

Table 5. Regions sensitive to a conjunction contrast of mental state compared with Stage A and Stage D as well as harm compared with Stage A and Stage D[a]

	Talairach coordinates						MS versus harm decoding	
Region	*X*	*Y*	*Z*	*t*	*p*	Size	*t*	*p*
R STS	51	−19	−5	7.50	<1.0E-6	96	4.95[b]	1.4E-4[b]
R TPJ	48	−46	19	4.84	7.7E-5	35	5.54[b]	5.1E-5[b]
R STS2	45	5	−17	5.75	9.0E-6	29	2.63[b]	0.02[b]
R insula	36	5	10	−4.59	1.4E-4	15	0.73	0.47
R motor	12	5	37	−4.04	5.5E-4	17	1.74	0.11
L STS	−51	−19	−5	6.63	1.0E-6	52	3.95[b]	1.2E-3[b]
L TPJ	−48	−52	13	6.21	1.0E-6	110	8.03[b]	7.0E-7[b]

[a]Whole-brain contrast corrected at q(FDR) = 0.05. Right two columns present results of analysis testing whether across-subject classification accuracy between harm and mental state was significantly greater than chance.
[b]Statistically significant declassification (corrected for multiple comparisons).

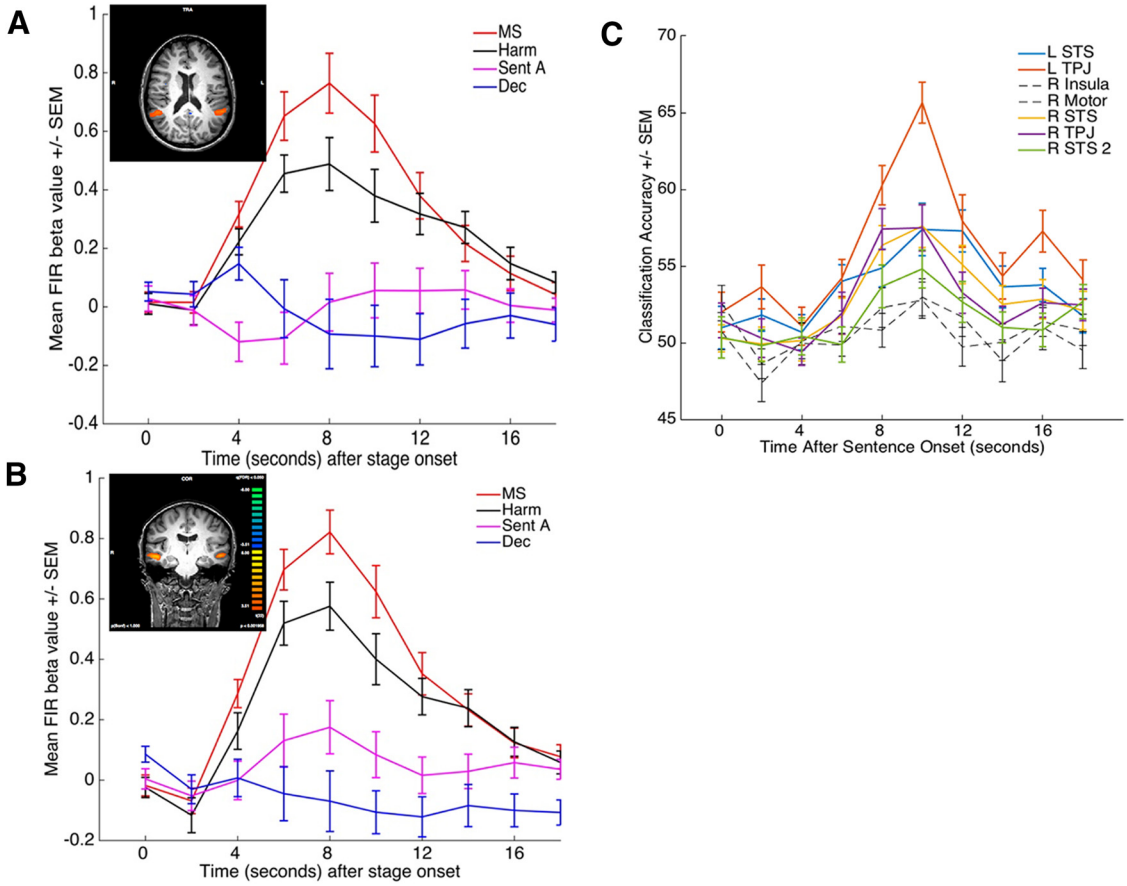

Figure 4. ***A***, ***B***, Deconvolution time courses of activity in TPJ (***A***) and STS (***B***). Insets, Locations of the relevant regions. ***C***, Event-related MVPA time courses illustrating mean classification accuracy as a function of time and ROI. Colored time courses represent above chance classification. MS, Mental State; Sent A, Sentence A; Dec, decision stage.

Stage A; 3, mental state > decision; 4, harm > decision. This conjunction of contrasts revealed shared positive activations in bilateral STS and bilateral TPJ (Table 5; Fig. 4*A*,*B*). Both STS and TPJ regions overlap substantially or entirely with the regions identified in the mental state > harm analysis (compare Tables 3, 5; Figs. 3*A*,*C*, 4*A*,*B*). As the time courses in Figure 4*A*, *B* reveal, in each of these regions, mental state evaluation shows greater activation than harm evaluation, but there is also pronounced activation associated with harm evaluation. To test whether these common activations represent recruitment of shared resources or instead reflect the recruitment of distinct neural ensembles, we performed MVPA in the identified regions to determine whether a pattern classifier could decode whether subjects were evaluating harm or mental state at the time of the evaluation. We observed marked decoding in both TPJ and STS (Fig. 4*C*), providing evidence for the conclusion that harm and mental state evaluation engage overlapping regions but use largely distinct neural ensembles.

Table 6. Regions displaying a linear relationship between level of mental state and brain activity in a whole-brain contrast: linear whole-brain contrast of mental state[a]

	Talairach coordinates					
Region	*X*	*Y*	*Z*	*t*	*p*	Size
PCC	−3	−49	25	4.00	1.6E-4	19
L MPFC	−6	56	34	5.00	4.0E-6	38
L STG	−46	17	−14	5.52	1.0E-6	62

[a]Whole-brain contrast corrected at q(FDR) = 0.05.

To assess whether the ROI analysis may have missed brain regions involved in processing mental state or harm evaluation, we also tested for such regions using whole-brain analyses that looked

Table 7. Regions showing evidence of supporting mental state and harm integration by means of the contrast (Stage C > Stage B) > (Stage B > Stage A)[a]

	Talairach coordinates						Superadditive harm × MS interaction		Punishment decoding (C)	
Region	*X*	*Y*	*Z*	*t*	*p*	Size	*F*	*p*	*F*	*p*
R middle occipital gyrus	39	−70	1	4.46	<1.0E-6	34	0.00	1.00	−0.06	0.96
PCC	−3	−22	28	6.41	<1.0E-6	774	0.05	1.00	0.52	0.61
R DLPFC	30	32	40	4.10	<1.0E-6	26	3.09	0.10	0.76	0.45
R amygdala	24	−13	−14	5.53	<1.0E-6	72	12.46	<1.0E-6[b]	−0.49	0.63
MPFC	6	41	7	6.11	<1.0E-6	380	0.05	1.00	0.57	0.57
L amygdala	−21	−7	−20	6.53	<1.0E-6	52	7.84	0.01[b]	−0.41	0.69

[a]Whole-brain contrast corrected at q(FDR) = 0.05. Superadditive harm × MS interaction column shows statistics for an ROI-based analysis in each region identifying patterns consistent with a superadditive interaction similar to that displayed in the behavioral results and a nonspecific mental state × harm interaction, respectively. Punishment decoding (C) reports the significance of MVPA decoding of punishment amount during Stage C in each of these regions compared with chance. All ROI analyses corrected for multiple comparisons. The PCC region is rostral to and does not overlap with the region identified in the mental state > harm contrast (compare Figs. 3*A*, 5*A*; Tables 3, 5, 7), just as the present MPFC region does not overlap with the left MPFC region identified in the whole-brain linear effect of mental state analysis (compare Tables 6 and 7).

[b]Statistically significant interaction effect.

for patterns of activations consistent with the various processing patterns described in the above analysis. As such, this whole-brain analysis removes the antecedent step of requiring a significant difference in activations for mental state compared with harm, or vice versa. For mental state, in addition to the same PCC region identified in the mental state > harm analysis (compare Table 3 and Table 6), we identified positive linear relationships in left MPFC and left superior temporal gyrus (STG) (Table 6). The whole-brain approach did not reveal any areas using the quadratic or searchlight MVPA analyses. In the case of harm, no regions were observed with a whole-brain linear, quadratic, MVPA, or vicarious somatosensation-based [1, 1, 1, −3] analysis.

Together, these results not only reveal that the neural substrates processing harm and mental state evaluations are largely dissociable, they also indicate that brain regions involved in each of these two factors may code distinct properties of the factor, such as the difficulty of its evaluation or its amount of culpability or harm.

fMRI data: integration of the harm and mental state components

The above results indicate that separable neural systems are recruited to evaluate harm and mental state information. Even regions showing common activations for harm and mental state, specifically the STS and TPJ, display evidence that distinct neural ensembles are recruited for the evaluation of the two components. This raises the question of what regions may support the real-time neural integration of these two components. To answer this question, we isolated regions that were preferentially recruited at Stage C compared with Stage B (Stage C − Stage B) because Stage C is the first stage at which integration can happen as subjects have access to both the mental state and the harm. However, given that Stage C also involves greater working memory demand than Stage B, it is likely that at least some of the regions isolated may be related to working memory per se rather than the integration of harm and mental state. We can address this issue with the following contrast ((Stage C − Stage B) − (Stage B − Stage A)), as the Stage B − A component of this contrast should also compare two stages with similarly different working memory demands. The resulting SPM of this contrast revealed activation indicative of integration in bilateral amygdala, MPFC, right DLPFC, PCC, and right middle occipital gyrus (Table 7; Fig. 5*A–C*), with most of these regions previously identified as putative sites of integration of information (Buckholtz and Marois, 2012; Buckholtz et al., 2015; Yu et al., 2015).

To more precisely characterize the role these regions play in integrating harm and mental state, we sought evidence of differential activation as a function of an interaction between level of harm and mental state that parallels the behavioral results (i.e., a superadditive effect of culpable mental state and severe harm). Specifically, using GLM5 (see Materials and Methods), we modeled conditions based on a 2 × 2 factorial design of mental state (blameless, culpable) and harm (low, high) at Stage C. As displayed in Table 7 and Figure 5*D*, both left and right amygdala display a robust interaction mirroring the superadditive behavioral effect of mental state and harm integration (Fig. 2*A*). No other regions were observed when performing this interaction analysis on whole brains.

That the pattern of amygdalae activity mirrors subjects' punishment behavior is evidence for a relationship between the amygdalae and the ultimate punishment decision. To further explore this potential brain-behavior relationship, we examined how subjects' individual differences in amygdalae response correlated with their differences in weighting the interaction factor in their punishment decisions. Specifically, for each subject, we calculated an index of the strength of the interaction in subjects' amygdalae activity ((culpable high harm − blameless high harm)) − (culpable low harm − blameless low harm)) and compared it with the interaction β weights calculated for each subject. If the interaction effect observed in the amygdalae were associated with the interaction effect observed in the behavior, we would expect that the strength of the interaction displayed in subjects' amygdalae to predict the strength of the interaction displayed in subjects' behavior. Consistent with this hypothesis, we found that subjects' interaction indices in the amygdalae were positively correlated with the interaction term ($r = 0.42$, $p = 0.044$; Fig. 5*E*).

fMRI data: the punishment decision stage

Brain regions involved in the decisional stage of a punishment judgment should display at least the two following characteristics: (1) preferential activation during the punishment decision stage of the task and (2) a functional relationship between brain activity during the time of the punishment decision and the outcome of the decision.

To search for such regions, we first identified those meeting the first criterion and then limited our analysis for the second criterion to the regions identified in the first step. To test the first criterion, we extracted subjects' β values for each task stage and used GLM2 (which modeled each of the different task stages) to perform a conjunction analysis of the decision stage of the task compared with each of the other task conditions, namely, Stage A, mental state and harm evaluation, and the ISI math task. We included the ISI task in the conjunction

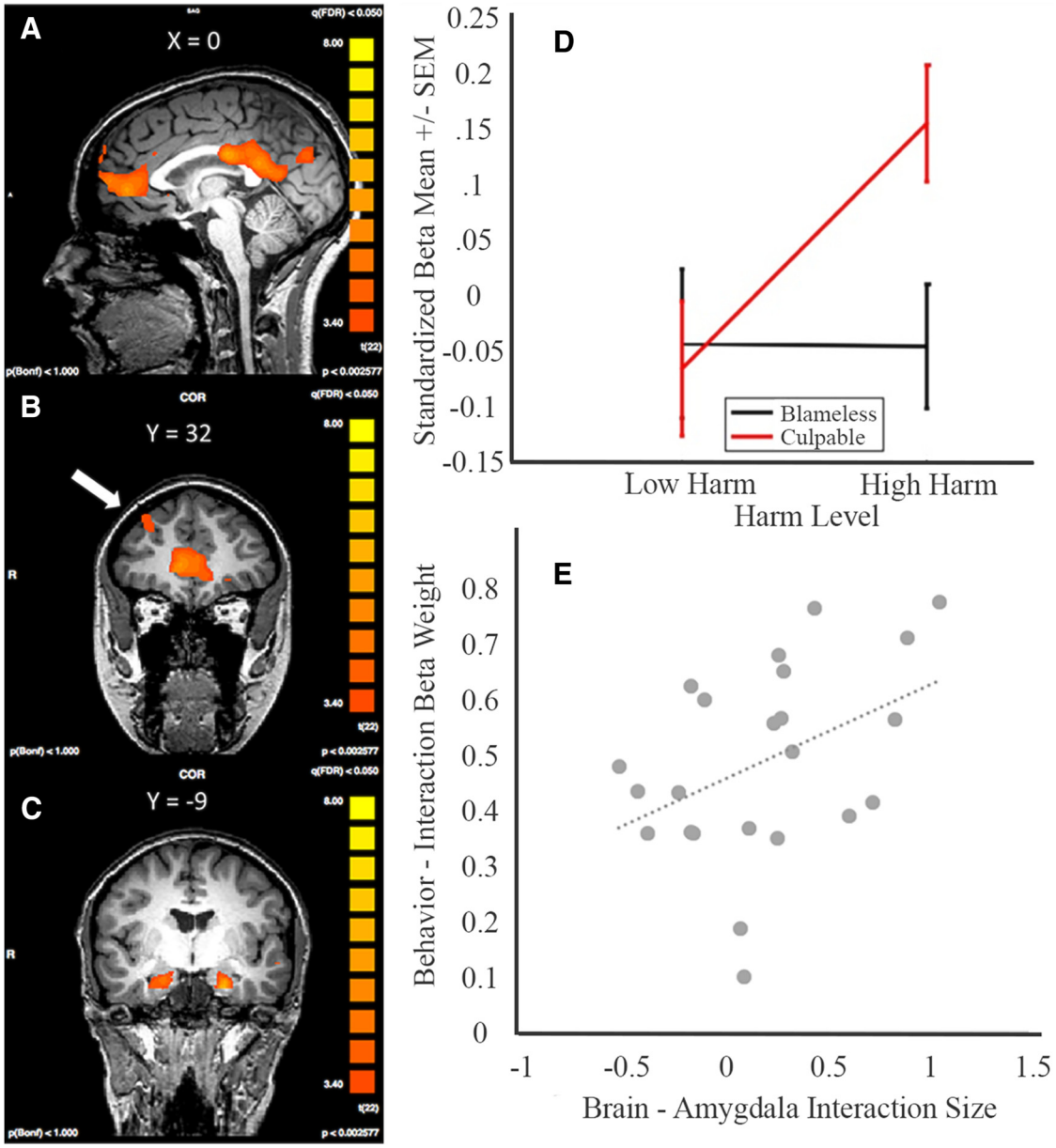

Figure 5. ***A***, MPFC, PCC. ***B***, DLPFC. ***C***, Bilateral amygdala display activity consistent with integration using the following contrast: (Stage C − Stage B) − (Stage B − Stage A). ***D***, The amygdala (left) displays an interaction activation profile in which there is an effect of harm level when the actor has a culpable mental state. ***E***, There is a positive correlation between the strength of the interaction in the amygdala and how much subjects weighted the interaction term in their punishment decisions ($r = 0.4195$, $p = 0.046$).

as it is the only other task condition that involves response selection. Given the unique demands of Stage D compared with other task components, this analysis expectedly revealed preferential activity in a number of regions, including right DLPFC, left ventrolateral prefrontal cortex, bilateral IFG, and visual and motor areas (Fig. 6*A*; Table 8). Each of these regions displayed activity that was significantly correlated with RT at the decision screen (Table 8).

To test the second criterion (i.e., to assess whether activity in any of the brain regions isolated above was linked to the decision of whether or how much to punish at the time of the decision), we sought to identify relationships between brain activity and decisional metrics using both univariate and multivariate approaches. First, we found no robust correlation between activity amplitude and level of punishment (Table 8), replicating Buckholtz et al. (2008). This may not be surprising given that subjects may engage in similar decisional reasoning across punishment ratings. Another possibility, assessed with MVPA, is that different neural ensembles in the DLPFC encode different punishment ratings. To address this issue, for each region, we divided subjects' punishment decisions into quartiles and trained and tested a classifier on the activity corresponding with punishment decisions falling into each of the quartiles. Of the regions identified by the first criterion, we observed significant decoding of the trial-by-trial punishment amount in only right DLPFC and visual cortex (Table 8; Fig. 6*B*). As some have cautioned that differences in subject-by-subject RT can induce false-positive decoding (Todd et al., 2013), we also performed the original analysis after regressing out differences in activity associated with differences in trial-by-trial RT and still observed significant decoding in the DLPFC ROI ($t = 1.74$, $p = 0.048$ one-tailed) and in the visual region ($t = 2.831$, $p = 0.005$ one-tailed). We hypothesize that decoding in the

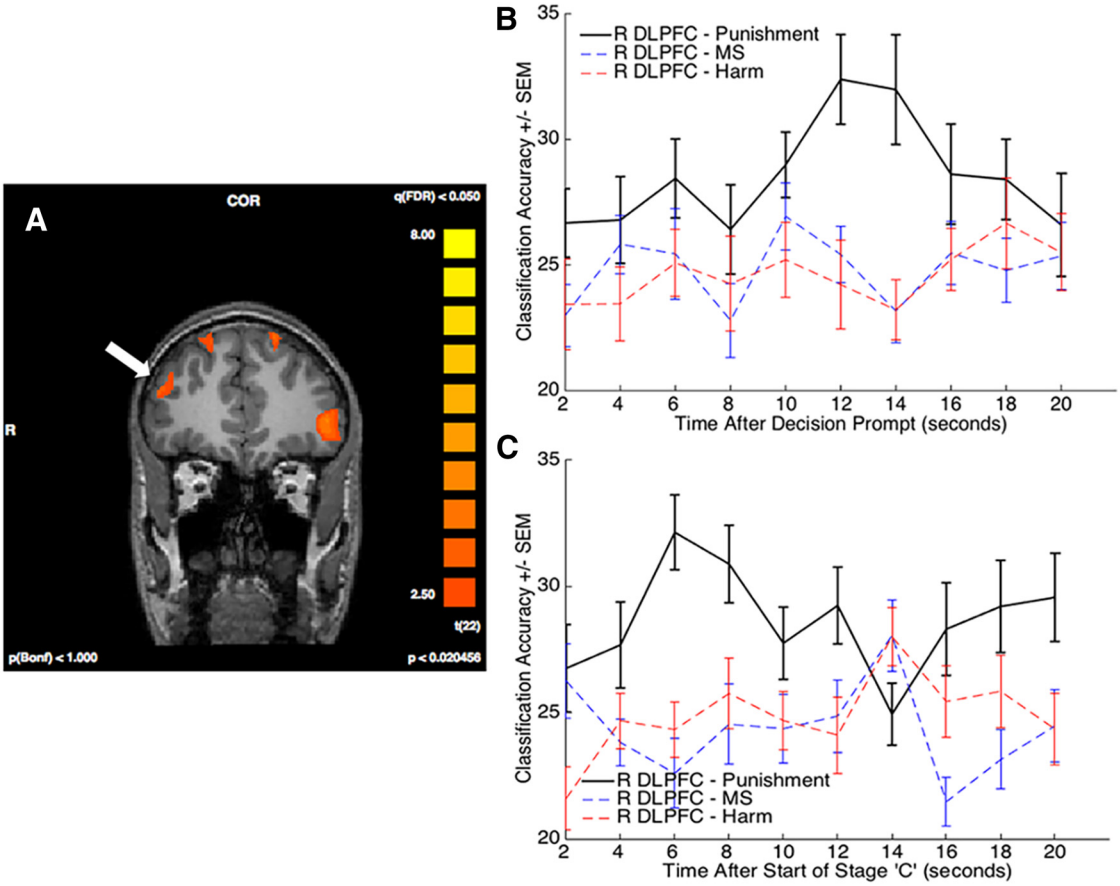

Figure 6. ***A***, SPM showing regions (arrow points to right DLPFC) with preferential engagement at the time of decision by means of a four-way conjunction between the time of decision and the other task components (see Results). ***B***, ***C***, Decoding of punishment rating in the right DLPFC region. The er-MVPA time courses plot classification accuracy of the voxels in the identified right DLPFC region on punishment rating as well on the level of mental state and harm at Stage B, the time of the decision, and Stage C. MS, Mental State.

Table 8. Regions showing significant activation for the conjunction contrast between Stage D and all other task stages[a]

	Talairach coordinates						Correlation with decoding RT		Main effect of punishment amount		Punishment decoding (D)		Punishment decoding (C)	
Region	*X*	*Y*	*Z*	*t*	*p*	Size	*t*	*p*	*F*	*p*	*t*	*p*	*t*	*p*
L VLPFC	−48	42	1	6.42	2.0E-6	30	273.88	<1.0E-6[b]	0.73	0.68	0.42	0.34	−0.44	0.67
L IFG	−45	14	−7	6.07	4.0E-6	50	118.14	<1.0E-6[b]	2.16	0.30	−0.66	0.30	0.04	0.97
Medial frontal gyrus	3	18	53	5.46	1.8E-5	148	96.17	<1.0E-6[b]	1.67	0.39	1.47	0.11	0.16	0.88
Visual	6	−63	1	13.37	<1.0E-6	5510	175.23	<1.0E-6[b]	1.26	0.39	7.25[b]	2.8E-06[b]	1.96	0.06
R DLPFC	36	45	26	4.91	6.6E-5	95	192.78	<1.0E-6[b]	1.32	0.39	2.92[b]	0.02[b]	2.80[b]	0.01[b]
R IFG	39	21	−2	5.29	2.6E-5	22	163.83	<1.0E-6[b]	1.00	0.53	1.79	0.10	−1.58	0.13
R Motor	39	3	52	5.77	8.0E-6	126	89.55	<1.0E-6[b]	1.44	0.39	1.57	0.11	1.08	0.29

[a]Whole-brain contrast corrected at q(FDR) = 0.05. Correlation with decoding RT column tests whether there is a significant effect of Stage D RT on activity in the identified regions. Main effect of punishment amount column tests for a main effect on activity in each ROI as a function of subjects punishment quantities. Punishment decoding (D) column reports the significance of MVPA decoding of punishment amount during the decision stage in each of these regions compared with chance. Punishment decoding (C) column reports the same for Stage C. All ROI analyses corrected for multiple comparisons. VLPFC, Ventrolateral prefrontal cortex.

[b]Statistically significant correlation with decision RT, statistically significant main effect of punishment amount, or significant punishment amount classification accuracy.

visual ROI is associated with subjects' visual evaluation of the punishment scale and response.

Importantly, the involvement of the DLPFC ROI in punishment rating is relatively specific, as this ROI failed to decode either the different mental state or harm levels ($t = 0.69$, $p = 0.25$ and $t = 0.90$, $p = 0.19$ one-tailed, respectively; Fig. 6*B*). This right DLPFC ROI also overlaps with the right DLPFC ROI previously hypothesized to be involved in the decision to punish (Buckholtz et al., 2008; Buckholtz and Marois, 2012). Previous studies investigating second- and third-party punishment decision-making have frequently found punishment decision-making to selectively engage the right as opposed to the left DLPFC (Sanfey et al., 2003; Knoch et al., 2006; Buckholtz et al., 2008; Baumgartner et al., 2014). Here punishment classification accuracy was similarly right-lateralized, as we failed to find any decoding ($t = 0.94$, $p = 0.18$ one-tailed) in a region with the same y and z coordinates in the left hemisphere.

In a final analysis, we examined whether this same right DLPFC ROI encoded punishment levels during Stage C as well. While the task is designed to interfere with decision-making at Stage C, subjects most likely make their first approximations of the punishment decision at Stage C, after they have been pre-

sented with both harm and mental state information. Furthermore, analysis of the punishment decision at Stage C has the added benefit over Stage D of not having any potential motor response confound. Thus, using the same methodological approach previously applied to Stage D, we tested each of the regions identified by the integration and decision contrasts (Tables 7 and 8, respectively). Of the regions tested, the only one to decode punishment level was the right DLPFC region identified in the decision contrast (Fig. 6*C*; Tables 7, 8), thereby further implicating this brain region in assignment of punishment. And once again, this region does not seem to encode either mental state or harm level. It is also noteworthy that the visual area that survived MVPA at Stage D failed to decode at Stage C, a result that supports our hypothesis that its decoding at the decision stage is due to subjects' visual evaluation of the scale.

Discussion

Our behavioral results indicate that punishment decisions are primarily driven by the interaction between mental state and harm. This interaction is characterized by a superadditive relationship between the component factors. This is consistent with studies showing that intentionality augments the negative valence associated with the same harmful outcome (Gray and Wegner, 2008) and can even augment a person's quantification of the severity of a harmful outcome (Ames and Fiske, 2013, 2015). Using functional imaging, we sought to parse how these two components, mental state and harm, converge into a punishment response that is defined by their interaction.

The data indicate that mental state and harm evaluation are distinct processes that engage separable neural resources. In regards to mental state, a group of regions consisting of TPJ, DMPFC, and STS were preferentially engaged by the evaluation of the offender's intentions. These activations overlap with a network of regions sometimes described as a ToM network (Gallagher and Frith, 2003), although the regions also colocalize with elements of the Default Mode Network (DMN) (Decety and Lamm, 2007; Hacker et al., 2013). By implementing a parametric manipulation of mental states, we were able to reveal a relationship between the difficulty of the mentalization task and the amount of activity in ToM regions. The parametric manipulation also provides insight into the function of the PCC. Although the PCC is a hallmark feature of the DMN (Hacker et al., 2013), it is sometimes, but not consistently, linked with ToM processes (Carrington and Bailey, 2009). The present results indicate that, while the PCC shows activation for mental state evaluation, it displays a linear correlation with level of culpability instead of a relationship with mentalization difficulty. We hypothesize that PCC activity, perhaps in concert with the mPFC and STG, reflects the negative valence associated with the evaluation of the offender's culpable mental state (Maddock et al., 2003; Leech and Sharp, 2014) rather than ToM processing per se. That we do not see a similar activation profile for harm evaluation is consistent with prior studies showing that the PCC does not show augmented activity in trials containing bodily harms (Heekeren et al., 2005). Finally, it is interesting to note that we failed to decode in the brain the different mental states with MVPA despite marked univariate amplitude differences. While we acknowledge that a null result could reflect low power, robust decoding in other analyses (e.g., at the decision stage) provides some confidence that absence of decoding here is not an intrinsic lack of power. Based on these findings, we conclude that the distinct mental states are not encoded by distinct neural ensembles. Rather, the univariate results suggest that differences in mental state evaluations result from differential activations of the same neural ensembles.

In regards to harm evaluation, bilateral PI, left IPL, and left OFC show heightened activation. The functional profiles of the PI and IPL are consistent with studies linking them with perceptions of others' bodily pain, perhaps co-opting the same mechanisms used to process the subject's individual interoceptive signals (Singer et al., 2004, 2009; Lamm et al., 2011). Consistent with this interpretation, these regions were far less activated when the outcome was death, which may be expected if the region is engaged in evaluation of another party's pain. Preferential activation in OFC, on the other hand, may reflect its role in evaluations of relative value or cost (Wallis, 2007; Janowski et al., 2013). Its quadratic activity pattern is consistent with this hypothesis on the premise that determining the magnitude (i.e., negative value) of the offense is most challenging in the intermediate categories.

That harm and mental state evaluation deploy distinct neural systems raises the question of how these processes are cortically integrated. Buckholtz and Marois (2012) proposed that activity in mPFC and PCC in legal decision-making tasks were potentially related to their role in integrating these component processes, and this prediction was borne out by the present experiment; both mPFC and PCC are sites of integration of harm and mental state evaluation. This is consistent with studies indicating that these two brain regions act as cortical hubs interconnecting distinct and functionally specialized systems (Sporns et al., 2007; Buckner et al., 2009; Bullmore and Sporns, 2012; Liang et al., 2013), such as those engaged by the evaluation of an offender's mental state and the resulting harm. Our results also provide evidence that the right DLPFC supports integration, a finding consistent with recent work showing that disruption of activity in the DLPFC alters how harm and mental state are integrated into a punishment decision (Buckholtz et al., 2015).

A role of the amygdalae in punishment decision-making has long been proposed (Buckholtz et al., 2008), although their specific function in that context has been debated. While Buckholtz et al. (2008) showed that harmful outcomes but not culpable mental states engaged the amygdalae, Yu et al. (2015) found the opposite in a second-party punishment task. Yu et al. (2015) further observed effective connectivity between the amygdalae and brain regions associated with integration of intention and harm, although they did not observe an interaction effect in the amygdalae. What the present results suggest is that the role of the amygdalae in punishment decision-making is more complex; it is less responsive to either of the simple factors of harm or mental state than it is to the interaction of these factors. Specifically, we found that activation in the amygdalae are defined by a superadditive interaction wherein the amygdalae display robust activation only in the case of a culpable mental state and substantial harm. Most strikingly, the activation profiles of the amygdalae mimic the pattern of subjects' punishment decisions, as evidenced by the relationship between the strength of the interaction activity in individuals' amygdalae and the weight that they attribute to the interaction between harm and mental state in rendering their decisions. These behavioral and neurobiological findings are remarkably consistent with recent work showing that the amygdalae's response to gruesome criminal scenarios is suppressed by means of a temporoparietal-medial-prefrontal circuit when the harmful outcome was purely accidental (Treadway et al., 2014). According to this account, the amygdalae are part of a corticolimbic circuit that, based on the offender's culpability, gates the effect of emotional arousal on punishment decisions (Treadway et al., 2014). Such a pivotal role of the amygdalae in

third-party punishment is in accord with the broader involvement of this brain region in mediating the influence of aversive states onto decision-making (Loewenstein and Lerner, 2002; Damasio, 2005; Phelps and LeDoux, 2005; Phelps, 2006; Miller and Cushman, 2013).

Finally, our results shed an important light on the role of the DLPFC in punishment decision-making. DLPFC activity in economic decision-making games has often been explained by a cognitive control account, according to which the DLPFC is promoting altruistic punishment behavior toward unfair players by inhibiting the prepotent response to act selfishly (Sanfey et al., 2003; Knoch et al., 2006). Such account of DLPFC function, however, is not easily reconcilable with third-party punishment studies showing greater DLPFC activity when subjects decided to punish (Buckholtz et al., 2008), or with other studies that have associated activity in this brain region across various cognitive tasks, such as working memory, analogical reasoning, rule-based decision-making, and amodal perceptual decision-making (Bunge et al., 2002; Heekeren et al., 2006; De Pisapia et al., 2007; Duncan, 2010; Hampshire et al., 2011). Furthermore, functional disruption of the DLPFC during third-party punishment decisions did not affect the severity of individuals' punishment decisions when the actor was blameless, but instead disrupted how they integrated the culpability of the actor and the severity of the harm in their punishment decisions (Buckholtz et al., 2015). Both of these observations favor an "integration and selection" hypothesis of DLPFC function in third-party punishment, in which the DLPFC integrates multiple neural representations from cognitive subtasks, such as the evaluation of the offender harm and mental state, to select an appropriate behavioral (punishment) response (Buckholtz and Marois, 2012; Buckholtz et al., 2015). Our results are highly consistent with this hypothesis. DLPFC activity was not only observed at the time of the decision response, it also selectively coded in a neurally distributed manner the amount of punishment assigned to the perpetrator. Thus, the DLPFC is not simply involved in the decision to punish, it is also implicated in assigning the appropriate punishment based on the relative weighing of the mental state of the transgressor and of the harm he caused.

In conclusion, the present study informs and extends proposed neural models of third-party punishment (Buckholtz and Marois, 2012). Evaluation of harm engages brain areas associated with affective and somatosensory processing, whereas mental state evaluation recruits primarily ToM/DMN circuitry. These representations are integrated in medial prefrontal cortical and subcortical (amygdala) structures, to be (presumably) routed to the DLPFC for the appropriate selection of a punishment response. Although many details remain to be worked out, this rigorous experiment paradigm reveals clear dissociations in the neural processing that underlies these complex, socially relevant, and legally important decisions.

References

Ames DL, Fiske ST (2013) Intentional harms are worse, even when they're not. Psychol Sci 24:1755–1762. CrossRef Medline

Ames DL, Fiske ST (2015) Perceived intent motivates people to magnify observed harms. Proc Natl Acad Sci U S A 112:3599–3605. CrossRef Medline

Baumgartner T, Schiller B, Rieskamp J, Gianotti LR, Knoch D (2014) Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. Soc Cogn Affect Neurosci 9:653–660. CrossRef Medline

Bowles S, Gintis H (2011) A cooperative species: human reciprocity and its evolution. Princeton, NJ: Princeton UP.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436. CrossRef Medline

Buckholtz JW, Marois R (2012) The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat Neurosci 15:655–661. CrossRef Medline

Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, Marois R (2008) The neural correlates of third-party punishment. Neuron 60:930–940. CrossRef Medline

Buckholtz JW, Martin JW, Treadway MT, Jan K, Zald DH, Jones O, Marois R (2015) From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron 87:1369–1380. CrossRef Medline

Buckner RL, Sepulcre J, Talukdar T, Krienen FM, Liu H, Hedden T, Andrews-Hanna JR, Sperling RA, Johnson KA (2009) Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. J Neurosci 29:1860–1873. CrossRef Medline

Bullmore E, Sporns O (2012) The economy of brain network organization. Nat Rev Neurosci 13:336–349. CrossRef Medline

Bunge SA, Dudukovic NM, Thomason ME, Vaidya CJ, Gabrieli JD (2002) Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. Neuron 33:301–311. CrossRef Medline

Carlsmith KM, Darley JM, Robinson PH (2002) Why do we punish? Deterrence and just desserts as motives for punishment. J Pers Soc Psychol 83:284–299. CrossRef Medline

Carrington SJ, Bailey AJ (2009) Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum Brain Mapp 30:2313–2335. CrossRef Medline

Castelhano MS, Muter P (2001) Optimizing the reading of electronic text using rapid serial visual presentation. Behav Information Tech 20:237–247.

Chang CC, Lin C-J (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Corradi-Dell'Acqua C, Hofstetter C, Vuilleumier P (2014) Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. Soc Cogn Affect Neurosci 9:1175–1184. CrossRef Medline

Cushman F (2008) Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. Cognition 108:353–380. CrossRef Medline

Damasio A (2005) Descartes' error: emotion, reason, and the human brain. London: Penguin.

Decety J, 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. CrossRef Medline

De Pisapia N, Slomski JA, Braver TS (2007) Functional specializations in lateral prefrontal cortex associated with the integration and segregation of information in working memory. Cereb Cortex 17:993–1006. CrossRef Medline

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172–179. CrossRef Medline

Fehr E, Gächter S (2002) Altruistic punishment in humans. Nature 415:137–140. CrossRef Medline

Fehr E, Rockenbach B (2004) Human altruism: economic, neural, and evolutionary perspectives. Curr Opin Neurobiol 14:784–790. CrossRef Medline

Gallagher HL, Frith CD (2003) Functional imaging of 'theory of mind.' Trends Cogn Sci 7:77–83. CrossRef Medline

Ginther MR, Shen FX, Bonnie RJ, Hoffman MB, Jones OD, Marois R, Simons KW (2014) The language of mens rea. Vanderbilt Law Rev 67:1327–1372.

Gray K, Wegner DM (2008) The sting of intentional pain. Psychol Sci 19:1260–1262. CrossRef Medline

Hacker CD, Laumann TO, Szrama NP, Baldassarre A, Snyder AZ, Leuthardt EC, Corbetta M (2013) Resting state network estimation in individual subjects. Neuroimage 82:616–633. CrossRef Medline

Hampshire A, Thompson R, Duncan J, Owen AM (2011) Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. Cereb Cortex 21:1–10. CrossRef Medline

Haushofer J, Fehr E (2008) You shouldn't have: your brain on others' crimes. Neuron 60:738–740. CrossRef Medline

Heekeren HR, Wartenburger I, Schmidt H, Prehn K, Schwintowski HP,

Villringer A (2005) Influence of bodily harm on neural correlates of semantic and moral decision-making. Neuroimage 24:887–897. CrossRef Medline

Heekeren HR, Marrett S, Ruff DA, Bandettini PA, Ungerleider LG (2006) Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. Proc Natl Acad Sci U S A 103:10023–10028. CrossRef Medline

Jackson PL, Meltzoff AN, Decety J (2005) How do we perceive the pain of others? A window into the neural processes involved in empathy. Neuroimage 24:771–779. CrossRef Medline

Janowski V, Camerer C, Rangel A (2013) Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Soc Cogn Affect Neurosci 8:201–208. CrossRef Medline

Keysers C, Kaas JH, Gazzola V (2010) Somatosensation in social perception. Nat Rev Neurosci 11:417–428. CrossRef Medline

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

LaFave WR (1986) Criminal law. St. Paul: West.

Lamm C, Decety J, Singer T (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage 54:2492–2502. CrossRef Medline

Leech R, Sharp DJ (2014) The role of the posterior cingulate cortex in cognition and disease. Brain 137:12–32. CrossRef Medline

Liang X, Zou Q, He Y, Yang Y (2013) Coupling of functional connectivity and regional cerebral blood flow reveals a physiological basis for network hubs of the human brain. Proc Natl Acad Sci U S A 110:1929–1934. CrossRef Medline

Loewenstein G, Lerner JS (2002) The role of affect in decision making. In: Handbook of affective sciences. Oxford: Oxford UP.

Maddock RJ, Garrett AS, Buonocore MH (2002) Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. Hum Brain Mapp 18:30–41. CrossRef Medline

Mathew S, Boyd R (2011) Punishment sustains large-scale cooperation in prestate warfare. Proc Natl Acad Sci U S A 108:11375–11380. CrossRef Medline

Miller R, Cushman F (2013) Aversive for me, wrong for you: first-person behavioral aversions underlie the moral condemnation of harm. Soc Pers Psychol Compass 7:707–718. CrossRef

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660. CrossRef Medline

Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442. CrossRef Medline

Phelps EA (2006) Emotion and cognition: insights from studies of the human amygdala. Annu Rev Psychol 57:27–53. CrossRef Medline

Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175–187. CrossRef Medline

Rand DG (2012) The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. J Theor Biol 299:172–179. CrossRef Medline

Rosnow RL, Rosenthal R (1996) Contrasts and interactions redux: five easy pieces. Psychol Sci 7:253–257. CrossRef

Rozzi S, Ferrari PF, Bonini L, Rizzolatti G, Fogassi L (2008) Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. Eur J Neurosci 28:1569–1588. CrossRef Medline

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. Science 300:1755–1758. CrossRef Medline

Shen FX, Hoffman MB, Jones OD, Greene JD, Marois R (2011) Sorting guilty minds. N Y Univ Law Rev 86:1306–1360. Medline

Shenhav A, Greene JD (2014) Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. J Neurosci 34:4741–4749. CrossRef Medline

Simons KW (2003) Should the Model Penal Code's Mens Rea provisions be amended. Ohio State J Criminal Law 1:179.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD (2004) Empathy for pain involves the affective but not sensory components of pain. Science 303:1157–1162. CrossRef Medline

Singer T, Critchley HD, Preuschoff K (2009) A common role of insula in feelings, empathy and uncertainty. Trends Cogn Sci 13:334–340. CrossRef Medline

Sporns O, Honey CJ, Kötter R (2007) Identification and classification of hubs in brain networks. PLoS One 2:e1049. CrossRef Medline

Tamber-Rosenau BJ, Dux PE, Tombu MN, Asplund CL, Marois R (2013) Amodal processing in human prefrontal cortex. J Neurosci 33:11573–11587. CrossRef Medline

Tassy S, Oullier O, Duclos Y, Coulon O, Mancini J, Deruelle C, Attarian S, Felician O, Wicker B (2012) Disrupting the right prefrontal cortex alters moral judgement. Soc Cogn Affect Neurosci 7:282–288. CrossRef Medline

Todd MT, Nystrom LE, Cohen JD (2013) Confounds in multivariate pattern analysis: theory and rule representation case study. Neuroimage 77:157–165. CrossRef Medline

Treadway MT, Buckholtz JW, Martin JW, Jan K, Asplund CL, Ginther MR, Jones OD, Marois R (2014) Corticolimbic gating of emotion-driven punishment. Nat Neurosci 17:1270–1275. CrossRef Medline

Wallis JD (2007) Orbitofrontal cortex and its contribution to decision-making. Annu Rev Neurosci 30:31–56. CrossRef Medline

Yu H, Li J, Zhou X (2015) Neural substrates of intention: consequence integration and its impact on reactive punishment in interpersonal transgression. J Neurosci 35:4917–4925. CrossRef Medline