Neural and Cognitive Bases of Human Punishment Behavior

By Matthew Raymond Ginther

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

May, 2017

Nashville, Tennessee

Approved:

Jeffrey Schall, Ph.D.

Rene Marois, Ph.D.

Baxter Rogers, Ph.D.

David Zald, Ph.D.

Note to Readers: The chapters that comprise this document contain, in some instances, previously published work or work that is currently under review. It is important to acknowledge the co-authors of this work for their contribution as well as the funding sources detailed therein.

The citations for the relevant work are provided below:

Chapter 1: Matthew Ginther, Richard J. Bonnie, Morris B. Hoffman, Francis X. Shen, Kenneth W. Simons, Owen D. Jones, and René Marois, *Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment,* 36 The Journal of Neuroscience 9420 (2016).

Chapter 3: Matthew Ginther, Lauren Hartsough, and René Marois, *Moral Outrage Drives the Interaction of Harm and Culpable Intent in Punishment Decisions* (under review).

Chapter 4: Matthew Ginther, Lauren Hartsough, and René Marois, *Distinct Emotional Profiles for Second- and Third-Party Norm Violations* (under review).

ACKNOWLEDGEMENTS

I am deeply indebted to many for their contribution towards this work. Amongst them I would like to particularly highlight the foundational work by Josh Buckholtz, Francis Shen, and Michael Treadway. Their work served as the stepping stone for the work contained here. I also would like to highlight the contributions by my co-authors: Richard Bonnie, Morris Hoffman, Owen D. Jones, Ken Simons, and Lauren Hartsough. Helpful comments were provided on chapter one by B.J. Casey, Terry Lohrenz, Marc Raichle, and Anthony Wagner. I would be remiss to not thank Benjamin Tamber-Rosenau for his intellectual and moral support. His contribution permeates almost everything I have done here at Vanderbilt. Finally, I am immensely grateful to my Ph.D. advisor, René Marois, for training, mentoring, and advising me these past seven years. It’s been quite the adventure.

TABLE OF CONTENTS

LIST OF TABLES

Page

Chapter One

Chapter Two

Chapter Three

LIST OF FIGURES

Page

Chapter One

Chapter Two

Chapter Three

Chapter Four

INTRODUCTION

There are many reasons to study human punishment decision-making. For one, it directly impacts the lives of millions whom are punished by the criminal justice system every day. A recent study found that one out of three Americans have been arrested for an offense (Brame et al., 2014). Beyond mere arrest, on a daily basis nearly 7 million individuals are either being actively imprisoned, jailed, or are under probation or parole (Glaze and Kaeble, 2014). Beyond the number of people, the costs are astounding. The most recent data indicate that the United States spent over a quarter of a trillion dollars on justice system expenditures in 2012 alone (Kyckelhahn, 2015). For comparison, the National Institutes of Health received approximately $30 billion in federal money that same year and all federal research expenditures totaled less than $65 billion. Of course, this is not to make a normative or empirical claim as to whether more or fewer people need to be punished or that more or less must be spent on justice, just that justice isn't cheap.

The amount spent on justice stands in stark contrast to what we know about the mechanisms of human punishment decision-making. Punishment decision-making is often irrational and easily swayed by spurious factors. The most salient, and popularly discussed, is the effect of offender race on punishment decisions (Steffensmeier et al., 1998; Tonry 1995). While the insidious effect of racial bias has been known–and even acknowledged–for some time, the discussion of its influence came to a head when, in 1986, Warren McCleskey appealed his death sentence to the United States Supreme Court on the ground that Georgia's capital sentencing process was administered in a manner that violated his rights to equal protection under the law, as guaranteed by the 14th amendment (McCleskey v. Kemp, 481 U.S. 279 (1987)). McCleskey cited a study authored by David C. Baldus and two colleagues that empirically examined the effect of race on the likelihood a defendant would receive a penalty of death (Baldus et al., 1983). Baldus et al. found that, after controlling for nearly 40 possible non-racial variables, a defendant charged with killing a white victim was 4.3 times as likely to be sentenced to death as a defendant charged with killing a black victim (McCleskey at 287). The decision by the Court, in a stunningly candid admission of how imperfect punishment intuitions are, noted that while the study may indicate that there is a discrepancy that "appears to correlate with race [,] . . . disparities in sentencing are an inevitable part of our criminal justice system" (McCleskey at 313) and thus did not, by itself, support a conclusion that the right to equal protection under the law had not been violated. To be sure, arbitrary influences beyond race exist as well. Studies have shown that punishment decisions can be easily biased by any number of factors. One particularly disturbing result demonstrated how a parole board's decisions appeared to correlate one to one with how long had lapsed since their last break (Danziger et al., 2011). Thus, if for no other reason, studying how we make punishment decisions is therefore important so that we can better understand how and when they go wrong.

Though studying punishment is also important for what it gets right. Simply put, punishment promotes cooperation. The type of large-scale cooperation amongst non-kin that defines the human species is impossible without the threat that defection will be met with punishment of some form (Fehr and Gachter, 2002a; Fehr and Rockenbach, 2004; Bowles and Gintis, 2011). When this punishment is carried out by an uninvolved third party ("third-party punishment"), it is particularly effective in ensuring stable cooperation on the massive scales seen in society today (Boyd and Richerson, 2005; Nowak and Sigmund, 2005; Mathew and Boyd, 2011). This is in contrast to second-party punishment, when the punishment is administered by the person who was the victim of the norm violation. The theoretical and observational research that has led to the conclusion that third-party punishment is central to what makes us human is only further buttressed by evidence that we are–seemingly–the only species to engage in third-party punishment behavior (Riedl et al., 2012). This stands in stark contrast to the observation that numerous species, besides humans, engage in acts of second-party punishment (Mulder and Langmore, 1993; Bshary and Grutter, 2005; Wenseleers and Ratnieks, 2006). This stark distinction in human and animal behavior only emphasizes the role that human neurobiology plays in punishment, and accordingly, in providing the framework that supports the fabric of modern society.

Given its importance, a number of studies have begun to address the question of why and how we punish. As for why, research has examined both the types of norm violations that lead to third-party punishment as well as the cognitive and behavioral states that may induce or motivate punishment behavior in response to the norm violation.

As for the characteristics of a norm-violation that are associated with punishment, behavioral studies have identified that there are two primary components that appear to dictate the outcome of third-party punishment. These are (1) the mental state of the offender; and (2) the severity of harm he caused (Carlsmith et al., 2002a; Cushman, 2008). Mental state of the offender refers to the intention of the offender in relation to the resulting harm, recognizing the same result can be attributable to both innocent or illicit intentions. The importance of mental state in punishment decisions aligns with real-world legal norms and practice dating back millennia (LaFave et al., 1986; Shen et al., 2011) and is embraced as a central tenet of our notion of due process: *actus reus non facit reum nisi mens sit rea*, or “the act is not culpable unless the mind is guilty." Though ancient in origin, the dynamics of how intentions and harm outcomes are integrated into a punishment decision have not been extensively studied. In the behavioral component of chapter one we examine how mental state and harm integrate into a punishment decision.

To understand why we punish, we must not only understand the types of norm-violations that trigger punishment behavior, but also uncover the internal motivation to respond to norm violations with punishment. Though there are highly rational reasons for why one should punish a wrongdoer–and our criminal justice system assumes that those that make punishment decisions are disinterested and impartial–there is much evidence to suggest that it is also subject to emotional influence (Carlsmith et al., 2002; Salerno and Peter-Hagene, 2013; Anon, 2014). Indeed, it has been suggested that emotional responses to a crime can strongly predict the administered punishment (Buckholtz et al., 2008), in line with the widely-held notion that emotions are powerful drivers of adaptive behavior (Plutchik, 1980). In chapters three and four we begin to explore the emotional motivators of punishment decisions.

Chapter three examines the correspondence between emotional states and punishment outcomes in the third-party context. Though it may be widely acknowledged that emotions serve as a driving force underlying TPP, there is considerable uncertainty regarding which emotional states are specifically linked to punishment. The past few decades have laid witness to research identifying moral outrage, contempt, anger, and disgust as all playing a role in third-party punishment (Rozin et al., 1999; Carlsmith et al., 2002a; Hutcherson and Gross, 2011; Russell and Piazza, 2013; Shweder et al., 2013). In chapter three we attempt to provide some clarity by independently and parametrically manipulating both mental state culpability and harm severity, and examining how contempt, anger, disgust, and moral outrage map onto these distinct components. We further examined how these emotions may differentially mediate the relationship between the norm violation and the punishment. We find that unlike anger, contempt, and disgust, moral outrage is evoked by the integration of culpable mental state and resulting harms, and it alone mediates the relationship between this integrative process and punishment decisions. We conclude that, distinct from other emotional responses, moral outrage captures the interaction of mental state and harm to fuel punishment.

As noted at the outset of this introduction, punishment can come from both an uninvolved third party, as well as from the victim of the norm-violation himself. Both second and third-party punishment have been observed in humans and some evidence has indicated that individuals punish similarly, regardless of whether they are the victim of the norm violation or not (Chavez and Bicchieri, 2013; Lergetporer et al., 2014). That we respond to second- and third-party norm violations similarly hints that both processes likely engage a similar cognitive process that responds to norm-violations generally. In Chapter four I examine how subjects respond to both second- and third-party norm violations in the course of interacting with other individuals in an interactive game. Our results indicate that while second- and third-parties respond to norm violations similarly in terms of how they punish wrongdoers, they report experiencing distinct emotional states as a result of the violation. Specifically, we observe that anger is almost entirely reported only by those who experience a second party violation while moral outrage is largely reported only by those who experience a third-party violation. These results indicate that while responses to second and third party violations may result in similar outcomes, they likely draw from distinct neurocognitive processes. These results can be related back to the results of chapter three as further evidence that moral outrage is a critical experience in the evaluation of third-party wrongs.

Beyond why we punish, it is also important to understand how we punish. That is, neurobiologically, what brain systems contribute to, and support, the integration of mental state and harm into a punishment decision. Prior research of punishment decision-making has hinted that these two different components rely on distinct brain systems to support their evaluation, with mentalization of others' thoughts (assumed to be critical to evaluation of the mental state of an actor) correlated with heightened activity in temporoparietal junction (TPJ), superior temporal sulcus (STS) and dorsomedial prefrontal cortex (DMPFC) (Corradi-Dell'Acqua et al., 2014). Likewise, studies have found that the evaluation of harmful events is associated with affective circuitry such as the amygdala and the insula (Jackson et al., 2005; Buckholtz et al., 2008; Shenhav and Greene, 2014). These results have been proposed to extend to the differential contribution of these two sets of brain regions to punishment decision-making (Buckholtz and Marois, 2012), but this has not been formally tested. It also remains unclear what brain systems support the integration of these two components. Prior studies have identified heightened activation in the dorsolateral prefrontal (DLPFC), medial prefrontal (MPFC) and posterior cingulate cortex (PCC) at the time of the punishment decision and this has been used to support the conclusion 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). In the neuroimaging component of chapter one we examine the brain systems engaged in the evaluation and integration of these separate components using fMRI, with a specific focus on testing the hypotheses detailed above. We overcome many limitations of previous research by implementing a novel experimental design that can parse the differential contribution of distinct components of punishment decision-making.

The results from this study confirm that evaluation of harms engaged brain areas associated with affective and somatosensory processing, whereas mental state evaluation primarily recruited circuitry involved in mentalization. Further, the results provided strong evidence that harm and mental state evaluations are integrated in medial prefrontal and posterior cingulate structures, with the amygdala appearing to mediate the interaction between harm and mental state. Dorsolateral prefrontal cortex, on the other hand, displayed heightened activation at the time of the decision and also was unique in that it encoded the punishment decision.

Among other findings, the results from the first chapter provided strong evidence that the amygdala was centrally involved in punishment decision-making. The one-to-one correspondence between activation in the amygdala and the punishment outcome hints at the possibility of a direct relationship between the two. However, functional imaging is not well-situated to test this hypothesis and contemporary techniques for addressing causality (e.g., TMS & TDCS) are not well-suited for studying the amygdala. In chapter two I explore whether inducing activity in the amygdala through exogenous stimuli may affect punishment decision-making. I test this by presenting subliminal cues that are known to induce activation in the amygdala. The results from this study fail to produce clearly interpretable data, but some possible interpretations are discussed.

1. PARSING THE BEHAVIORAL AND BRAIN MECHANISMS OF THIRD-PARTY PUNISHMENT

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 Gachter, 2002b). This, in turn, enabled our species, with uniquely extensive cooperation among non-kin, to flourish at massive scales reflecting 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., 2002b; Cushman, 2008b). Although this comports with real-world legal norms and practices (LaFave et al., 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 (DLPFC), medial prefrontal (MPFC) and posterior cingulate cortex (PCC) at the time of decision-making, suggesting 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), though 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. While some studies have associated DLPFC with implementation of cognitive control (Sanfey, 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; Treadway et al., 2014a; Buckholtz et al., 2015).

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.

METHODS

Subjects

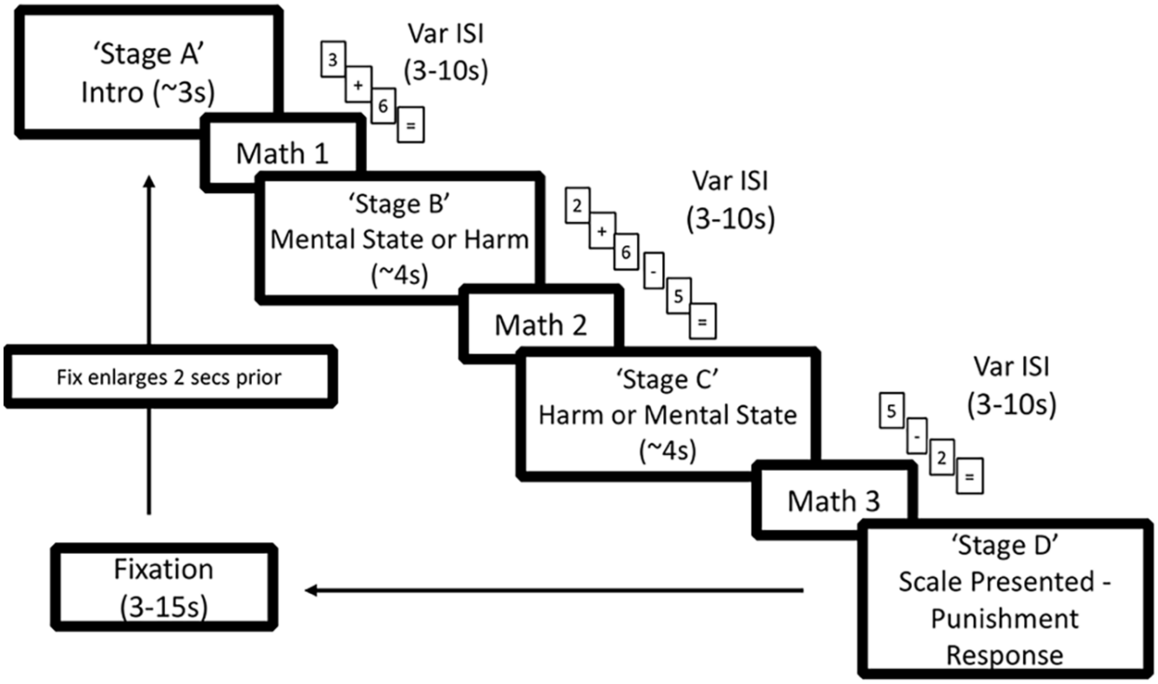
Twenty-eight right-handed individuals (13 females, ages 18-35 yrs) with normal or corrected-to-normal vision consented to participate for financial compensation. The Vanderbilt University Institutional Review Board approved the experimental protocol, and subjects provided their informed consent. Five 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 (> 3mm translation or 3 degrees of rotation) during the MRI scanning; and one failed to follow task instructions. That left twenty-three subjects (11 females, ages 18-35 yrs) 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; Treadway et al., 2014a; Buckholtz et al., 2015) 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 inter-stage-interval (ISI) drawn from an exponentially decaying distribution of approximately 3 to 10 s (Figure 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 (we capitalize these terms when referring to a specific component of our task). 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 seconds to make their response).

Fig. 1: Task Design



Note: Each round began with the presentation of 'Stage A' as a Rapid Serial Visual Presentation (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 inter-stimulus interval (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 'D' stage subjects were probed for their punishment response. The variable ITI lasted for a duration of 3-15 seconds, with the last two seconds accompanied by a larger fixation square. ISI, Interstimulus interval; ITI, Intertrial interval; Var, Variable.

Several details of the experiment were designed to optimize the likelihood that a given cognitive process occurred at a specific stage. First, in order to constrain the subjects' cognitive processing of each sentence to its presentation time and to preclude subjects from using the inter-stage intervals to ponder the appropriate response, the inter-stage intervals were filled with a secondary math task that lasted the duration of each ISI. Each math problem started 200ms 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 750ms and followed at the end by a '=', indicating that the subject should provide a response within 2 seconds. If no response was provided the task continued as if a response had been made. For example, a 3 second ISI would consist of two integers and one operation (e.g., 3, +, 5) for 2 s plus an average of a 1s-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, B and 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 to 9 response space accordingly.

As an example, a derivative scale may look like: '0' = 1 day in jail, '6' = 1 year in jail, and '9' = most severe punishment. For any given scenario, six of the 10 scales were available as possible options, with one of the six randomly selected for any given trial. The six scales per scenario were selected so as to ensure, based on pilot data, that the mean punishment response +/- 2 standard deviations 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 (M = 4.02, SD = 1.84) compared to 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 (M = 2.45, SD = 2.09). Second, at the time of the decision the distribution in reaction times was not uniform across levels of Harm or Mental State, as one would expect if subjects had made their decisions prior to Stage D. Instead there is a significant effect of both Mental State and Harm level on subject reaction time (see Figure 2B-C).

Following the subjects’ response, an inter-trial interval (ITI) drawn from a decaying exponential distribution from 3 to 15 seconds 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 two seconds of the ITI to signal to the participants the imminent start of the next trial (see Figure 1 for trial design).

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 four of the five 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 through Harm 4.

Some of the scenarios were based upon scenarios used in previous research (Shen et al., 2011) while 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 to 9 scale) to fine tune and verify our categorization of the scenarios along the four Harm levels (Harm 1: M = 1.49, SD = 0.29; Harm 2: M = 3.67, SD = 0.50; Harm 3: M = 6.13, SD = 0.37; Harm 4: M = 8.64, SD = 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 (Tables 1a and 1b).

The levels of the two 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, four levels of Mental State, and two possible orderings (Harm first or Mental State first), yielded a total of 512 different possible scenarios (64x4x2), 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 (i.e. subjects saw four scenarios in each Mental State (4 levels) x Harm (4 levels) cell in the factorial design). An example of a single theme and the eight derivative scenarios is presented in Tables 1a and 1b. Please note that 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., 2014a). 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 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 ($F(1,41) = 1.41$, $p = 0.241$).

<table>
<tr><th colspan="4"><b>Table 1a: Illustrative Theme (Planks & Bikes): Four “Mental-State First” Variations.</b></th></tr>
<tr><td colspan="4">Introductory Sentence</td></tr>
<tr><td colspan="4">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.</td></tr>
<tr><td>Purposeful Mental State</td><td>Reckless Mental State</td><td>Negligent Mental State</td><td>Blameless Mental State</td></tr>
<tr><td>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.</td><td>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.</td><td>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.</td><td>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.</td></tr>
<tr><td colspan="4">Harm Sentence</td></tr>
<tr><td colspan="4">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.</td></tr>
</table>

Table 1b: Illustrative Theme (Planks & 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.			
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.			
Purposeful Mental State	Reckless Mental State	Negligent Mental State	Blameless 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.	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.	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.	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.

Note: Subjects evaluated only one of the possible eight scenarios for each theme.

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 x 4 Harm-level design. The runs varied in duration given the variable response times, but never lasted more than 11.5 minutes. Each subject completed four of these fMRI runs. The experiment was programmed in Matlab (Mathworks, Natick MA) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997). 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 five-button keypads (Rowland Institute of Science, Cambridge, MA).

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 seven 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. While 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 Blood Oxygen Level Dependent (BOLD) images were acquired using a gradient-echo echoplanar imaging (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 (Brain Innovation, Maastricht, The Netherlands) in conjunction with custom Matlab software. 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 inter-stimulus 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 six distinct general linear models (GLMs) for each subject's data, with each GLM created to address a different question and avoid co-linearity issues between regressors. Specifically, to assess the evaluative process for Harm and Mental State separately, the first (GLM1) modeled each stage of the task as well as the inter-stimulus 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. In order 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 (GLM3) modeled only the Harm component, but with different regressors for each level of Harm in the sentence. The fourth (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 (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 four cells in order 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. Note that 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 beta 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 timecourses 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. While we display standard errors of the mean for these time courses, these are strictly for the purpose of visualizing the variance and shape of the hemodynamic responses. In order to avoid non-independent selective analysis of the data (the "double-dipping" problem) these timecourse 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 multi-voxel 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) 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 seconds after the decision prompt or 10 seconds after the start of the stage RSVP; which corresponds, on average, to 6 seconds 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 3mm 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 in order 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. Due to practical limitations we used the mean chance performance calculated on the ROI based MVPA analysis as chance for the searchlight analysis.

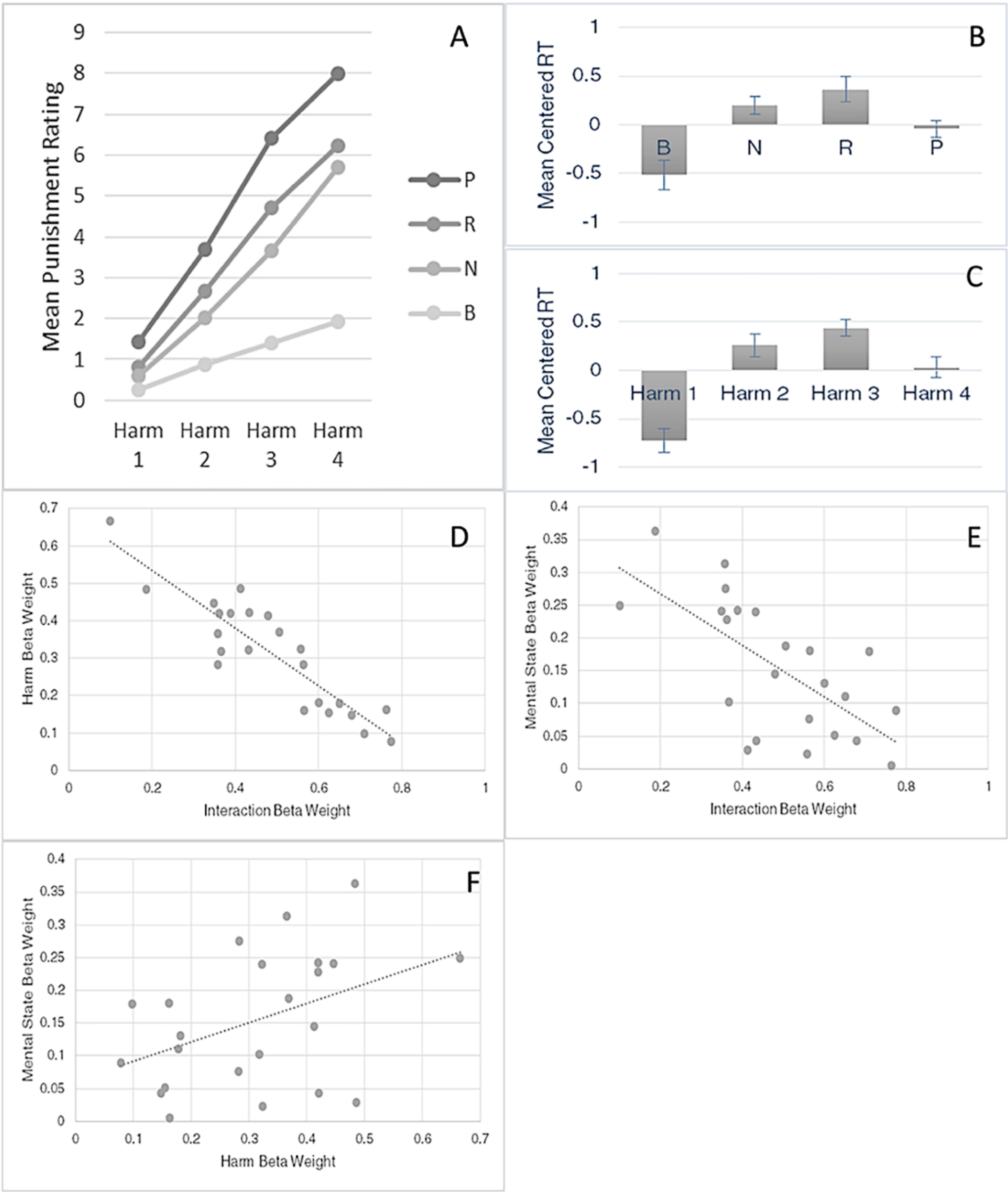
RESULTS

Behavioral Results

Figure 2A shows subjects' punishment ratings as a function of both Harm and Mental State levels. Using a repeated measures analysis of variance (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 ($F(6,144) = 3.84$, $p < 0.005$).

Figures 2B and 2C show subjects' mean reaction times (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 (Figures 2B-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 ($F(1,22) = 26.65$, $p < 0.001$).

Fig. 2: Results of Behavioral Analyses



Note: 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 display +/- 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.

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 2A, the nature of this interaction is a super-additive effect between Mental State and Harm. Mean r-square 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, Figure 2D) and Mental State (r = -0.67, p = 0.0005, Figure 2E), while Harm and Mental State showed a positive correlation (r = 0.43, p = 0.041, Figure 2F). 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.

Table 2: Behavioral Modelling 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	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

Note: Bolded model selected as the best model by means of Akaike Information Criterion (AIC). All beta coefficients standardized.

fMRI Data

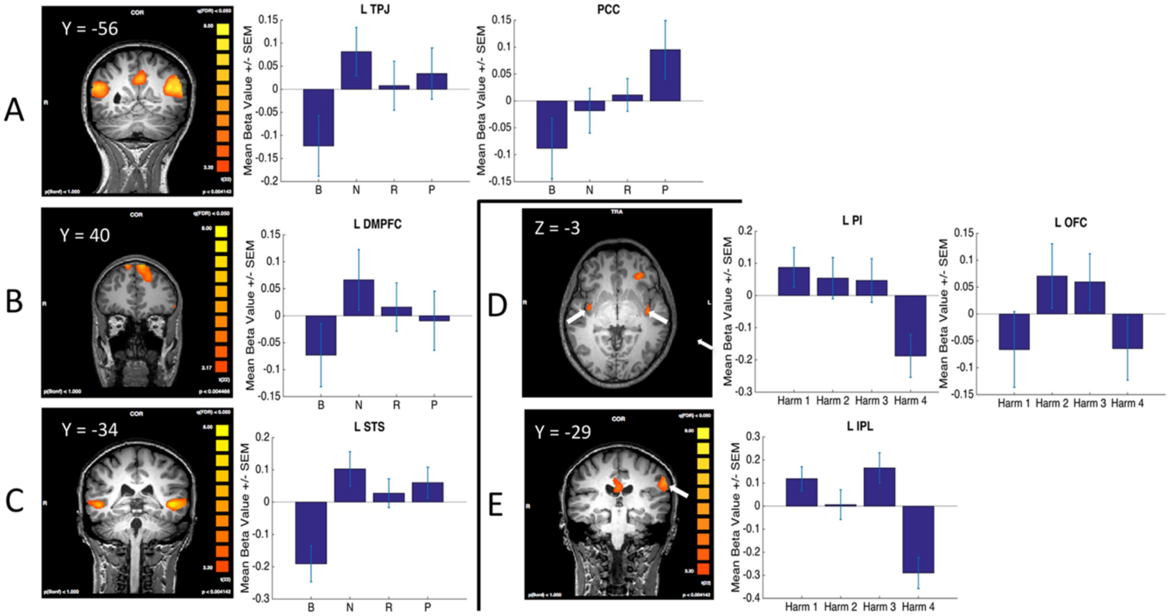
The analysis of the imaging data was directed at addressing three primary questions. First, to what extent do Mental State and Harm evaluation engage separable or common neural processes? Second, what regions support the integration of these two components? Third, is the punishment decision neurally 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 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, though we achieved qualitatively similar results when Mental State or Harm activity were 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' mind (Gallagher and Frith, 2003; Carrington and Bailey, 2009) including bilateral temporoparietal junction (TPJ), bilateral dorsomedial prefrontal cortex (dmPFC), and bilateral superior temporal sulcus (STS) (left panel Figure 3A-C, Table 3) as well as posterior cingulate cortex (PCC) (left panel Figure 3A-C, 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 (Mid. TG), left medial frontal gyrus (Med. FG), and left inferior frontal gyrus (IFG) (Table 3).

Fig. 3: fMRI Analysis of Evaluation of Mental State and Harm Information



Note: A-C. Left panel displays SPMs results of the contrast Mental State – Harm, highlighting A. TPJ and PCC, B. DMPFC, and C. STS. Right panel displays 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 panel displays SPMs results of the contrast Harm – Mental state illustrating D. PI and Left OFC, and E. L IPL. Right panel displays activity in the respective ROIs as a function of Harm level. SPM, Statistical Parametric Map; TPJ, Temporoparietal Junction; PCC, Posterior Cingulate Cortex; DMPFC, Dorsomedial Prefrontal Cortex; STS, Superior Temporal Sulcus; ROIs, Regions of Interest; PI, Posterior Insula; OFC, Orbitofrontal Cortex; IPL, Inferior Parietal Lobule; LPFC, Lateral Prefrontal Cortex.

Table 3: Mental State Evaluation at Time of Evaluation

	Talairach								Contrast with			
	Coordinates						Linear Contrast		MS difficulty		MS Decoding	
Region	**X**	**Y**	**Z**	***t***	***p***	**Size**	***F***	***p***	***F***	***p***	***t***	***p***
R Mid TG	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	0.08	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	4.8E-3	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	0.05	1.82	0.21
L DMPFC	-7	41	51	7.03	<1.0E-6	620	0.30	0.62	2.30	0.08	-3.06	0.08
L Med FG	-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	4.6E-3	8.34	7.6E-3	-1.66	0.21
L STS	-52	7	-22	11.47	<1.0E-6	266	8.20	2.7E-3	13.09	1.5E-3	-1.61	0.21
L TPJ	-43	-59	21	9.13	<1.0E-6	473	2.17	0.09	4.16	0.04	-0.08	0.94

Note: 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). Light shading indicates significance at p < 0.1. If both contrasts account for the data, dark shading indicates that the darkly shaded is significantly more consistent with the data than the other contrast (Rosnow and Rosenthal, 1996). 'MS Decoding column presents the results of a t-test compared to 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. TG, Temporal Gyrus; FG, Frontal Gyrus; IFG, Inferior Frontal Gyrus.

In each identified region of interest (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 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 beta 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: .22, R: .60, N: .52, B: .12. (Note that the quadratic fit of the classification accuracy data is similar to the RT data at response time for Mental States (see Fig. 2B). 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 if 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 (SVM) on brain activity during the period of evaluation. Note that for all MVPA analyses univariate differences were first subtracted out (see methods) so that the analysis was specific for multivariate patterns.

As displayed in Table 3 and visualized in Figure 3A-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 L 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, Figure 3A). Finally, we did not find above-chance levels of classification accuracy in any of the identified ROIs (Table 3). Taken 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 to 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 (left panel Figure 3D-E, 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 such that there was greater activation for the intermediate harms than the extreme harms (Figure 3D, Table 4), while 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 (Figure 3D-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.

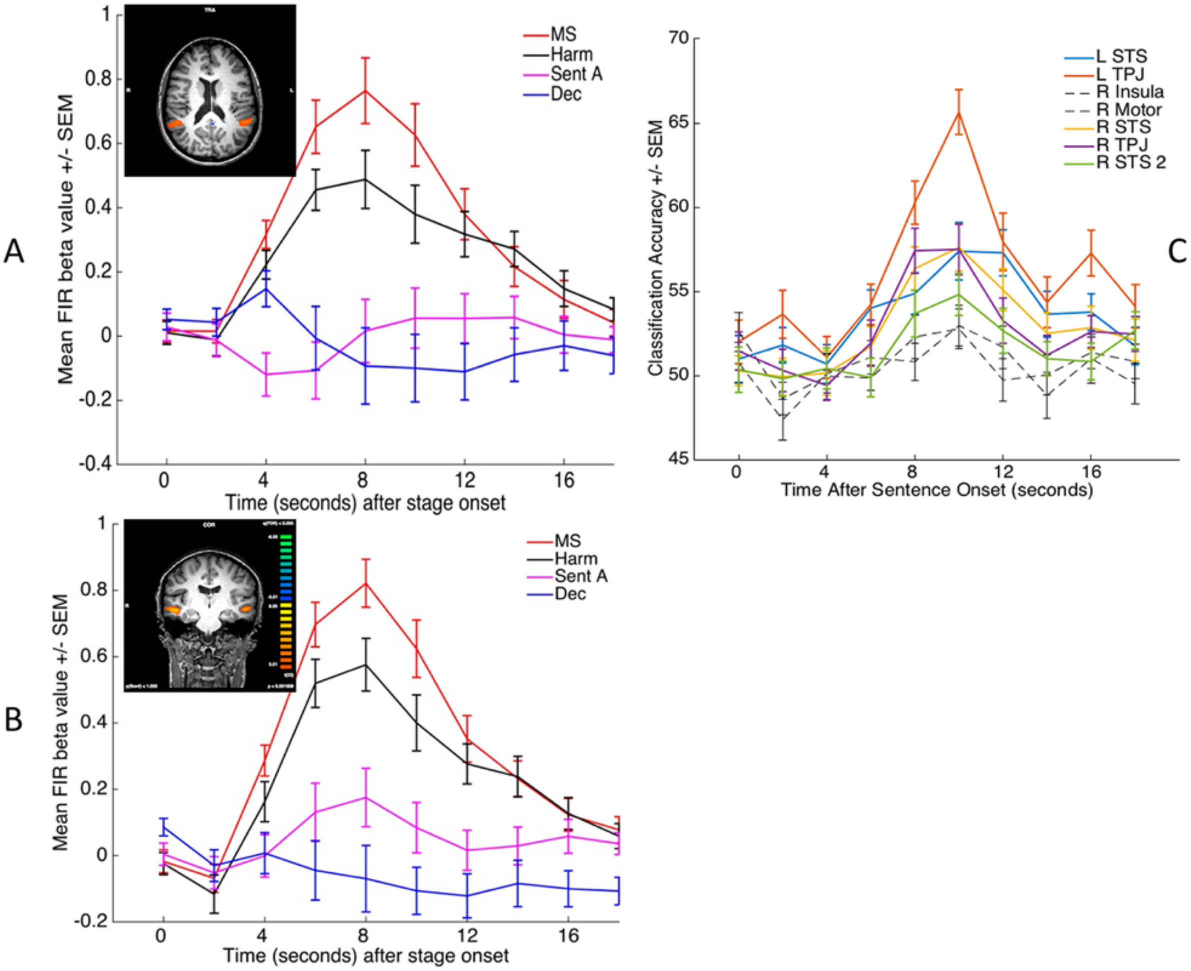
Table 4: Harm Evaluation at Time of Evaluation

	Talairach Coordinates						Linear Contrast		Difficulty Effect		Death Cond. Sig. 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	8.7E-5	0.95	0.25	18.74	4.9E-5	1.29	0.37
R PI	38	-8	-6	5.53	1.5E-5	15	7.55	5.4E-3	1.10	0.25	8.68	3.0E-3	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	0.04	1.51	0.18	-1.76	0.26
L PI	-40	-11	-3	5.17	3.5E-5	24	11.90	1.0E-3	3.46	0.07	16.14	1.1E-4	-0.90	0.53
L Fusiform Gyrus	-52	-53	-6	5.72	9.0E-6	30	10.79	1.3E-3	7.69	0.01	23.44	1.1E-5	-0.37	0.83
L IPL	-62	-29	33	5.61	1.2E-5	64	18.09	9.8E-5	9.41	0.01	35.74	<1.0E-6	1.67	0.26

Note: Table 4. Regions showing significant activation for Harm evaluation as contrasted with Mental State evaluation. Whole brain contrast corrected at q(FDR) = 0.05. 'Linear Contrast' column presents results of repeated measures ANOVA with a linear contrast. 'Quadratic Contrast' column presents the results of a repeated measures ANOVA with a quadratic contrast as a proxy of Harm evaluation difficulty. 'Death Cond. Sig. Lower' column presents the results of a repeated measures ANOVA with the contrast [-1 -1 -1 3]. Light shading indicates significance at p < 0.1. If more than one contrast accounts for the data, dark shading indicates the contrast accounts for significantly more of the variance in the data than the other two contrasts (Rosnow and Rosenthal, 1996). 'Harm Decoding' column presents the results of a t-test compared to chance level decoding of Harm level in each region. All ROI analyses corrected for multiple comparisons. TG, Temporal Gyrus; FG, Frontal Gyrus; IFG, Inferior Frontal Gyrus. LPFC, Lateral Prefrontal Cortex; OFC, Orbitofrontal Cortex.

Directly contrasting Harm and Mental State does not identify brain regions that may be commonly activated by the evaluation of the two components. In order to identify commonly recruited regions, we carried out 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); *i.e.* 1. Mental State > Stage A, 2. Harm > Stage A, 3. Mental State > Decision, and 4. Harm > Decision. This conjunction of contrasts revealed shared positive activations in bilateral STS and bilateral TPJ (Table 5, Figure 4A-B). Both STS and TPJ regions overlap substantially or entirely with the regions identified in the Mental State > Harm analysis (Cf. Tables 3, 5; Figures 3A, C; Figures 4A-B). As the timecourses in Figure 4A-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 see if 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 (Figure 4C), providing evidence for the conclusion that Harm and Mental State evaluation engage overlapping regions but employ largely distinct neural ensembles.

Fig. 4: fMRI Analysis of Regions Common to Mental State and Harm Evaluation



Note: A-B. Deconvolution timecourses of activity in A. temporoparietal junction (TPJ), and B. superior temporal sulcus (STS). Insets illustrate the locations of the relevant regions C. Event related MVPA timecourses illustrating mean classification accuracy as a function of time and ROI. Colored timecourses represent above chance classification. MS, Mental State; Sent A, Sentence A; Dec, Decision Stage.

Table 5: Mental State and Harm Evaluation vs. Other Task Components

	Talairach Coordinates						MS vs. Harm Decoding	
Region	**X**	**Y**	**Z**	***t***	***p***	**Size**	***t***	**p**
R STS	51	-19	-5	7.50	<1.0E-6	96	4.95	1.4E-4
R TPJ	48	-46	19	4.84	7.7E-5	35	5.54	5.1E-5
R STS2	45	5	-17	5.75	9.0E-6	29	2.63	0.02
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	1.2E-3
L TPJ	-48	-52	13	6.21	1.0E-6	110	8.03	7.0E-7

Note: Regions sensitive to a conjunction contrast of Mental State compared to 'Stage A' and 'Stage D' as well as Harm compared to 'Stage A' and 'Stage D'. 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. Shading indicates statistically significant declassification (corrected for multiple comparisons).

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 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 to Harm, or vice versa. For Mental State, in addition to the same PCC region identified in the Mental State > Harm analysis (Cf. Table 3 and Table 6) we identified positive linear relationships in left medial prefrontal cortex (MPFC), and left superior temporal gyrus (STG) (Table 6). The whole-brain approach did not reveal any areas with 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.

Table 6: Linear Whole-Brain Contrast of Mental State

	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

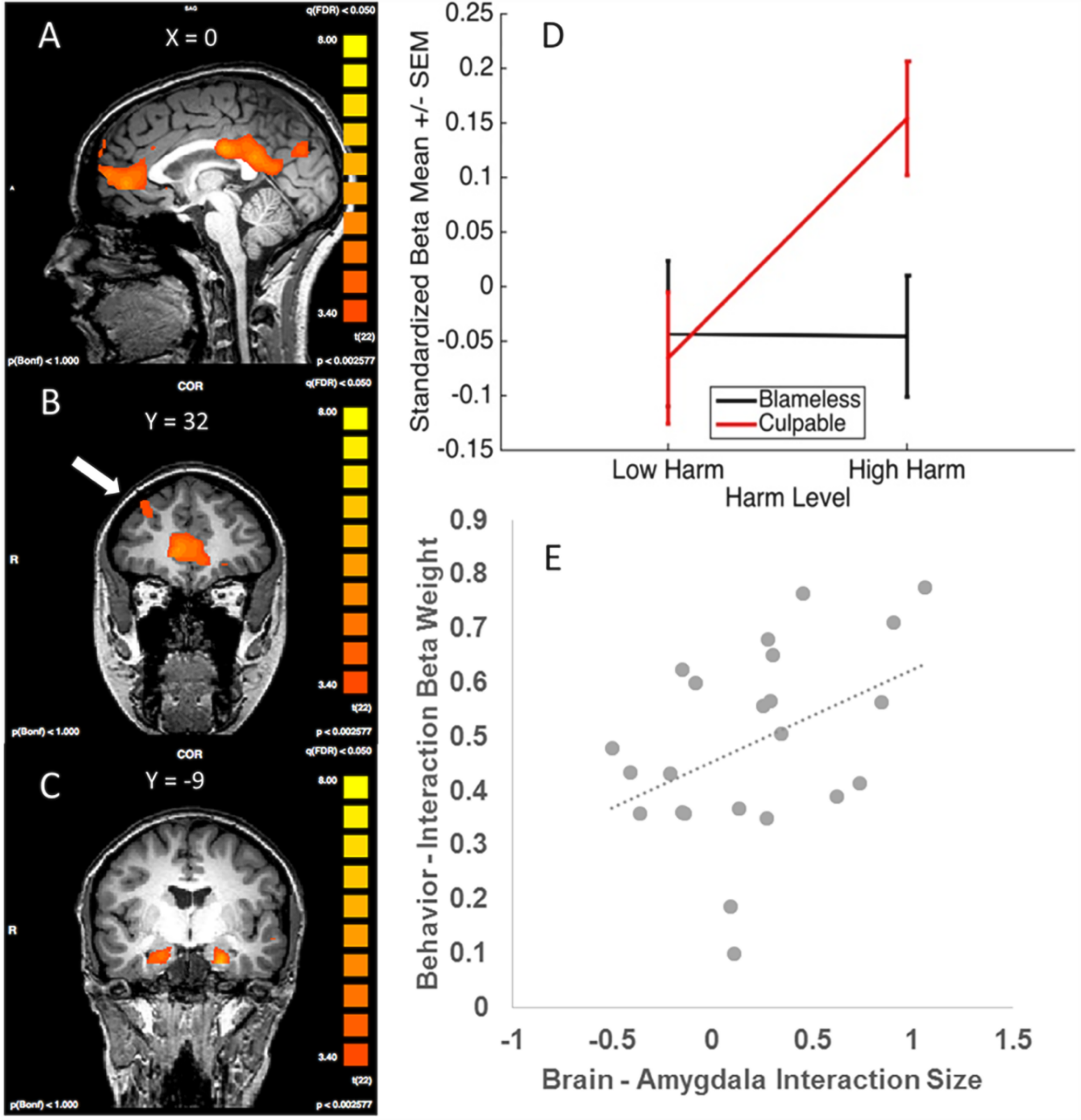
Note: Regions displaying a linear relationship between level of Mental State and brain activity in a whole-brain contrast. Whole brain contrast corrected at q(FDR) = 0.05. MPFC, Medial Prefrontal Cortex; STG, Superior Temporal Gyrus.

Taken 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 to Stage B (Stage C – Stage B) since 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 (WM) demand than Stage B, it is likely that at least some of the regions isolated may be related to WM 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 WM demands. The resulting SPM of this contrast revealed activation indicative of integration in bilateral amygdala, medial prefrontal cortex (MPFC), right dorsolateral prefrontal cortex (DLPFC), posterior cingulate cortex (PCC), and right middle occipital gyrus (Table 7, Figure 5A-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).

Fig. 5: fMRI Analysis of Integration Regions



Note: A. Medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), B. dorsolateral prefrontal cortex (DLPFC), and C. bilateral amygdala display activity consistent with integration using the following contrast: (Stage C-Stage B) – (Stage B – Stage A). D. The amygdala (left illustrated) 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* = .4195, *p* = .046).

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 super-additive effect of culpable Mental State and severe Harm). Specifically, using GLM5 (see Methods) we modeled conditions based on a 2 x 2 factorial design of Mental State (Blameless, Culpable) and Harm (Low, High) at Stage C. As displayed in Table 7 and Figure 5D, both left and right amygdala display a robust interaction mirroring the super-additive behavioral effect of Mental State and Harm integration (see Figure 2A). No other regions were observed when carrying out this interaction analysis on whole brains.

Table 7: Integration Regions

	Talairach Coordinates						Superadditive Harm x MS Intx.		Pun. Decoding (C)	
Region	**X**	**Y**	**Z**	***t***	***p***	**Size**	***F***	***p***	***F***	***p***
R Mid. Occ. Gyr.	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	-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	-0.41	0.69

Note: Regions showing evidence of supporting Mental State and Harm integration by means of the contrast (Stage C > Stage B) > (Stage B > Stage A). Whole brain contrast corrected at q(FDR) = 0.05. 'Superadditive Harm x MS Intx.' columns show 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 non-specific Mental State by Harm interaction, respectively. 'Pun. Decoding (C)' reports the significance of MVPA decoding of punishment amount during Stage C in each of these regions compared to chance. Shading indicates statistically significant interaction effect. 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 (Cf. Figure 3A, 5A; Tables 3, 5, and 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 (Cf. Tables 6 and 7). Mid. Occ. Gyr.. Middle Occipital Gyrus.

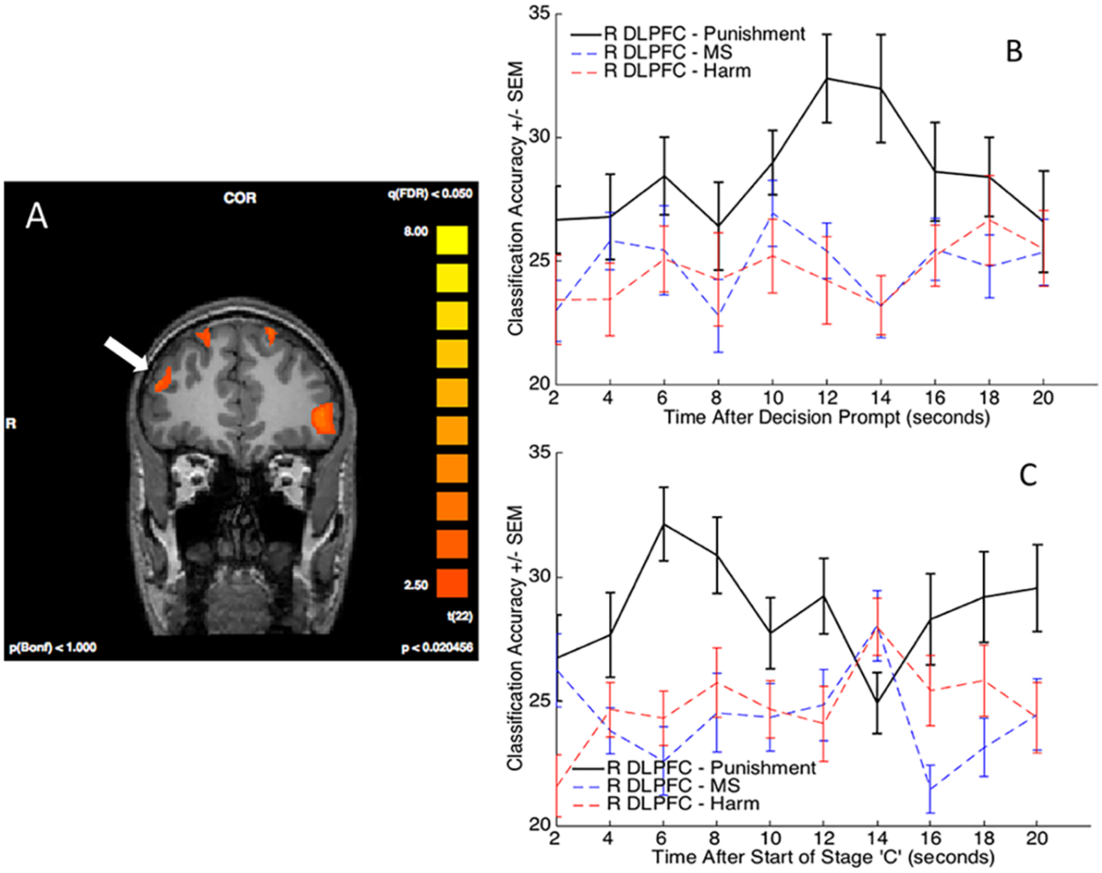
That the pattern of amygdalae activity mirrors subjects' punishment behavior is evidence for a relationship between the amygdalae and the ultimate punishment decision. To further explore this potential brain-behavior relationship, we examined how subjects' individual differences in amygdalae response 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 beta 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 = .42, p = .044, Figure 5E).

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.

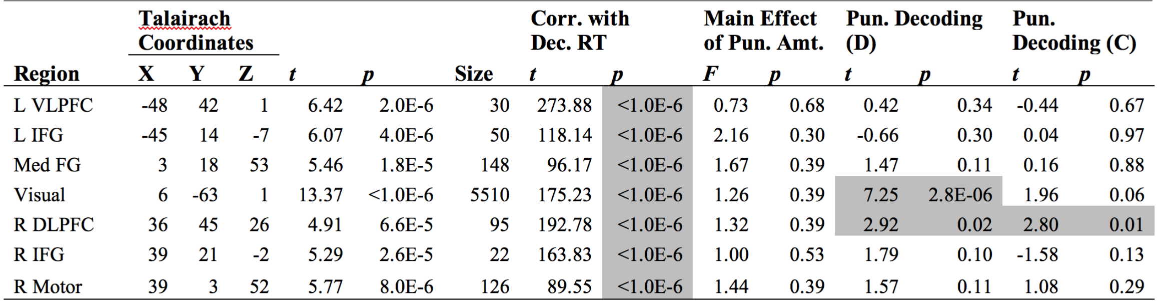
To search for such regions, we first identified those meeting the first criterion and then limited our analysis for the second criterion to the regions identified in the first step. To test the first criterion, we extracted subjects' beta values for each task stage and used GLM2 (which modeled each of the different task stages) to perform a conjunction analysis of the decision stage of the task compared to 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 as it is the only other task condition that involves response selection. Given the unique demands of Stage D compared to other task components this analysis expectedly revealed preferential activity in a number of regions, including right DLPFC, left ventrolateral prefrontal cortex (VLPFC), bilateral IFG, and visual and motor areas (Figure 6A, Table 8). Each of these regions displayed activity that was significantly correlated with RT at the decision screen (Table 8).

Fig. 6: fMRI Analysis of Decision Regions



Note: A. SPM showing regions (arrow points to right DLPFC) with preferential engagement at the time of decision by means of a 4-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 timecourses 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 B. the time of the decision and C. Stage C. DLPFC, Dorsolateral Prefrontal Cortex; MS, Mental State.

Table 8: Regions Showing Significant Activation for The Conjunction Contrast Between 'Stage D' And All Other Task Stages



	Talairach Coordinates						Corr. with Dec. RT		Main Effect of Pun. Amt.		Pun. Decoding (D)		Pun. 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	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	2.16	0.30	-0.66	0.30	0.04	0.97
Med FG	3	18	53	5.46	1.8E-5	148	96.17	<1.0E-6	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	1.26	0.39	7.25	2.8E-06	1.96	0.06
R DLPFC	36	45	26	4.91	6.6E-5	95	192.78	<1.0E-6	1.32	0.39	2.92	0.02	2.80	0.01
R IFG	39	21	-2	5.29	2.6E-5	22	163.83	<1.0E-6	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	1.44	0.39	1.57	0.11	1.08	0.29

Note: Whole brain contrast corrected at q(FDR) = 0.05. 'Corr. with Dec. RT' column tests whether there is a significant effect of 'Stage D' RT on activity in the identified regions. 'Main Effect of Pun. Amt.' column tests for a main effect on activity in each ROI as a function of subjects' punishment quantities. 'Pun. Decoding (D)' reports the significance of MVPA decoding of punishment amount during the decision stage in each of these regions compared to chance. 'Pun. Decoding (C)' reports the same for Stage C. Shading shows statistically significant correlation with decision RT, statistically significant main effect of punishment amount, or significant punishment amount classification accuracy. All ROI analyses corrected for multiple comparisons. FG, Frontal Gyrus; VLPFC, Ventrolateral Prefrontal Cortex; RT, Reaction Time.

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, Figure 6B). 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 reaction time 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 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, Figure 6B). 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, 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 presented with both Harm and MS 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 (Figure 6C, Tables 7 and 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 analysis 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 super-additive 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 which 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 Theory of Mind (ToM) network (Gallagher and Frith, 2003), though the regions also co-localize 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. While 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., 2002; 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 posterior insula (PI), left inferior parietal lobule (IPL) and left orbitofrontal cortex (OFC) show heightened activation. The functional profile of the PI and IPL are consistent with studies linking it with perceptions of others' bodily pain, perhaps coopting 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), though 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 and colleagues further observed effective connectivity between the amygdala and brain regions associated with the integration of intention and harm, though they did not observe an interaction effect in the amygdala. 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 amygdala is defined by a super-additive interaction wherein the amygdalae display robust activation only in the case of a culpable Mental State and substantial Harm. Most strikingly, the activation profile of the amygdala mimics 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., 2014a). According to this account, the amygdalae are part of a cortico-limbic circuit that, based on the offender's culpability, gates the effect of emotional arousal on punishment decisions (Treadway et al., 2014a). 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; Phelps et al., 2014).

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 towards unfair players by inhibiting the pre-potent response to act selfishly (Sanfey, 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), nor 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 sub-tasks (such as the evaluation of the offender harm and mental state) in order 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, while 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.

2. ASSESSING THE CAUSAL ROLE OF THE AMYGDALA IN PUNISHMENT DECISIONS THROUGH VISUAL MASKING

INTRODUCTION

A critical result of chapter one is the revelation of a one-to-one correspondence between activation in the amygdala and punishment behavior. Specifically, the super additive interaction between mental state and harm appears to correspond with both heightened punishment and heightened activation in the amygdalae. While this correspondence was purely correlational, the possibility of a causal brain-behavior relationship was buttressed by the relationship between individual subjects' amygdala profile and their individual punishment behavior.

That the amygdalae are critical to punishment decision-making is not, by itself, a novel result. Due to the fact that the amygdalae have been shown to be sensitive to aversive stimuli and are involved in threat detection, it is intuitive that they would be associated with punishment decision-making. This was initially supported by a study demonstrating the involvement of the amygdalae in moral judgment tasks (Heekeren et al., 2005) and more specifically so in Buckholtz et al. (2008), who found that activity in the amygdalae were correlated with the severity of a scenario's harm. Based on these results in combination with findings that negative affect influences legal decision-making (Feigenson and Park, 2006), Buckholtz and Marois (2012) proposed that the amygdalae influence punishment decision-making by being a heuristic for emotional affect, which, they postulate, is correlated with harm severity.

This proposed causal relationship between amygdala activity and punishment decision-making posits that increased activity in this brain region should, all else being equal, lead to increased punishment. Observational measurements of brain activity, such as what fMRI provides, are poorly situated to make inferences of this kind as they are not well suited to differentiate cause and effect (Smith, 2012).

Though fMRI is not the ideal technique for making claims of causality, using advanced analyses that examine temporal qualities of the BOLD timecourses can provide support for a claim of causality as opposed to mere correlation. Specifically, analyses such as Granger Causality Modelling (GCM) (Roebroeck et al., 2005), Structural Equation Modelling (SEM) (Kim et al., 2007), and Dynamic Causal Modelling (DCM) (Friston et al., 2017) can make claims about the casual relationship of two variables in time.

At least two studies have attempted to look at the causal role of the amygdala in punishment decisions using these types of functional imaging methods. Treadway et al. (2014) used GCM to identify heightened directional connectivity between the amygdala and dorsolateral prefrontal cortex when reading scenarios where the scenario protagonist was responsible for his conduct. Yu et al. (2015) found further support of integration of intentionality and harm in the amygdala by using DCM to demonstrate evidence for directionality from temporoparietal junction (claimed to be sensitive to accidental conduct) and anterior insula (claimed to be sensitive to attempted harm) with the amygdala.

Though analyses of this type are helpful, they have stark intrinsic limitations in their claims. For one, the analyses are still, ultimately, correlational. To borrow one analogy, GCM would indicate that sales of “Peeps” cause Easter. Further, these analyses are not able to rule out alternative causes for the observation, such as an independent brain region directing information to both the amygdala and the frontal regions analyzed in Treadway (2014) (Maziarz, 2015).

Lesion analyses provide a framework to make more robust causal claims. Broadly speaking, two types of lesion analyses are used in the field. In the classic sense of a lesion study, subjects with various brain damage are recruited and their brain damage is associated with a behavioral deficit. Studies such as these are the foundational work for modern neuropsychology (Cubelli and De Bastiani, 2011). However, because the lesions arise naturally or as the result of trauma, collecting subjects with consistently localized lesions in the region of interest is organizationally difficult. Ascribing varied lesions to a specific behavioral deficit is also challenging, but can be done in this field (Koenigs and Tranel, 2007), and modern neuroimaging and computational analyses have helped expand the usefulness of this approach (Bates et al., 2003), though it remains difficult and not frequently used because recruiting the number of subjects necessary to draw meaningful conclusions can take years, if not decades.

Another approach to lesion studies involves temporarily stimulating brain regions in order to create temporary "lesions." This is achieved by disrupting activity using two primary methods: transcranial direct current stimulation (TDCS) and transcranial magnetic stimulation (TMS). Such an approach was used to test the role of the dorsolateral prefrontal cortex in punishment decision-making in Buckholtz et al. (2015) and has also been used in similar tasks to parse the role of the temporoparietal junction (TPJ). However, the depth and size of the amygdala make both TDCS and TMS untenable solutions for targeted stimulation of the region.

Brain activity can, of course, also be modulated by exogenous cues or tasks. These stimuli and tasks, often used as functional localizers for imaging analyses, have been validated across numerous studies. While a number of functional localizers exist for the amygdala (e.g., (Hariri et al., 2002)), they are often unsuitable for use as an experimental manipulation due to the lack of a proper control (i.e., sham) condition and the possible interaction between the localizer task and the behavioral task. However, in the case of the amygdala, it is possible to induce localized activation using stimuli that do not reach conscious awareness. This is done by way of visual masking.

Visual masking involves the presentation of two stimuli, a target and a mask. The masking occurs when the mask interrupts the perception of the target (Kahneman, 1968). By presenting an angry or fearful face for a short duration (16 or 33ms), masked by a face expressing surprise for a longer duration, it is possible to induce activation in the amygdala without conscious perception of the target face (Whalen et al., 1998).

Based on our knowledge that a visually masked fearful face can augment activity in the amygdala without conscious awareness of the stimulus, we sought to test the hypothesis that the amygdala was causally linked to increased punishment amounts. Subjects made punishment decisions while, on half of trials, a masked fearful face was presented just prior to the punishment decision. In order to be able to better parse the nature of the amygdala's involvement, subjects evaluated scenarios that were parametrically and independently manipulated across level of harm and level of mental state. This allows for weights (i.e., the relative contribution of the different components) to be calculated, as in chapter one, for the contribution of harm, mental state, and their interaction on the punishment decision. By parsing the relative contribution of these three components in this way, our analyses may be able to distinguish the contribution of amygdala activation on punishment decisions

METHODS

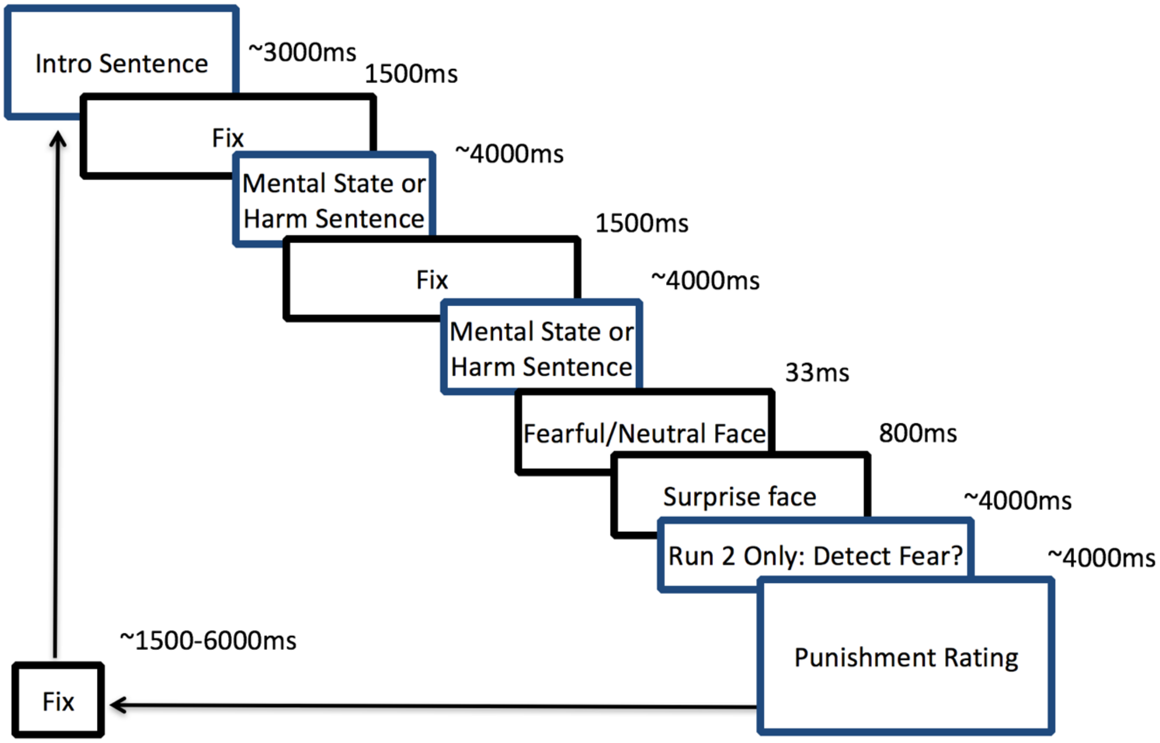
Twenty-five subjects (10 males) were recruited to participate in the study. Subjects completed the task for course credit. The task design mirrored that used in chapter one ('Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment') aside from the following changes.

Subjects completed only two runs, with each run comprising 18 trials. Only two runs were used due to pilot data indicating strong habituation effects after the first run (see below). Because this study was not designed with neuroimaging in mind, the math problems during the ISI period were replaced with fixation on a point (1500ms). Scenarios presented on individual trials were manipulated across both mental state (3 levels: blameless, negligent, and purposeful) and harm (3 levels: *de minimis*, substantial, and death). Thus, in each run there were two trials for each cell of the 3x3 design. These two trials consisted of one treatment and one control.

A number of steps were taken to ensure that the task provided subliminal affective priming: (1) adapting the masking procedures to follow those previously demonstrated to induce affective priming; (2) ensuring that the punishment task did not require the processing of prime faces (subjects were thus instructed that the image of the person was unrelated to the scenario, but, to explain the presence of the face, were told that the research question was how the presence of the face may impact their punishment decisions); (3) subjects were not told that there was a target face presented in addition to the mask; (4) the use of a funnel questionnaire at the end of the task to determine awareness of the target face; (5) trial-by-trial two-alternative forced choice probe to test for possible conscious perception; (6) a pilot study confirming the design induced changes in affect perception (Li et al., 2008).

Adapting previously used procedures, following the last sentence, but prior to making the punishment decision, subjects were presented with a target face (33ms) masked with a face expressing surprise (800ms). A surprise face was used as the mask because its ambiguity may facilitate subliminal processing of the target face (Kim et al., 2003; 2004). For the treatment condition the target face expressed fear, while for the control the face was neutral. This aspect to the design, and the accompanying timing procedures were based on (Whalen et al., 1998; Li et al., 2008; Kim et al., 2010), which demonstrated targeted activation in the amygdala without conscious awareness of the target using similar parameters. The face images were drawn from the KDEF database (KDEF, D. Lundqvist, A. Flykt and A. Ohman, Karolinska Hospital, Stockholm, Sweden). Images were colormatched, centered, and grayscaled using custom Matlab software.

Fig. 1: Task Design

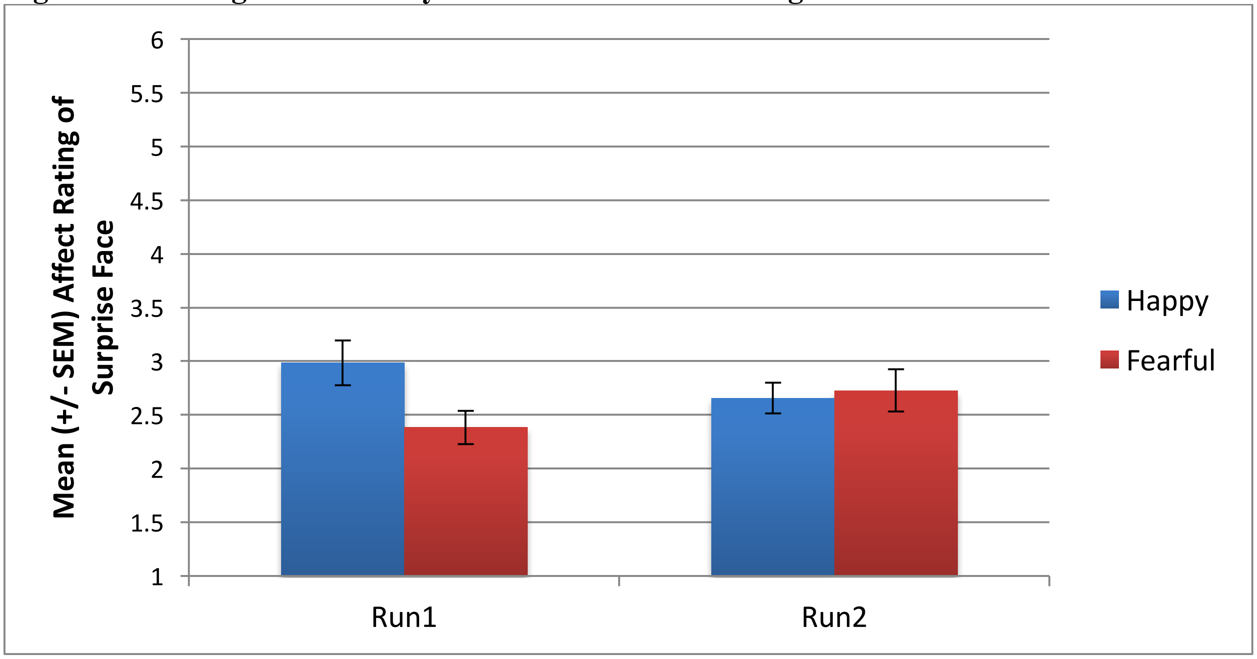


After the first run, subjects were debriefed using a funnel questionnaire adapted from Li et al. (2008) to determine if they had any perception of the target. Specifically, subjects were asked: (1) Did you see anything besides the surprise faces? (2) Did you see anything right before the surprise faces? (3) There was actually a flicker before the surprise faces. What did you see? (4) Did you see a face? (5) What expression did you see in the face?

After the debriefing, but before the second run, subjects were informed that there was a target face and that during run two an additional prompt was added prior to the punishment rating. This prompt directed subjects to indicate whether or not they detected fear in the target face. This was a forced-choice probe and were thus asked to make their best guess, even if they could not detect the presence of a target face.

In a pilot study, subjects completed the same task but before making a punishment decision were asked to rate the affect of a surprise face on a 6 point Likert scale (1 = Extremely Negative, 6 = Extremely Positive). This pilot study revealed that the paradigm successfully induced affective priming as seen by the affect ratings provided when a happy target was used compared to a fearful target. However, the priming habituated by the time of the second run (Figure 2).

Fig. 2. Task Design Successfully Induces Affective Priming

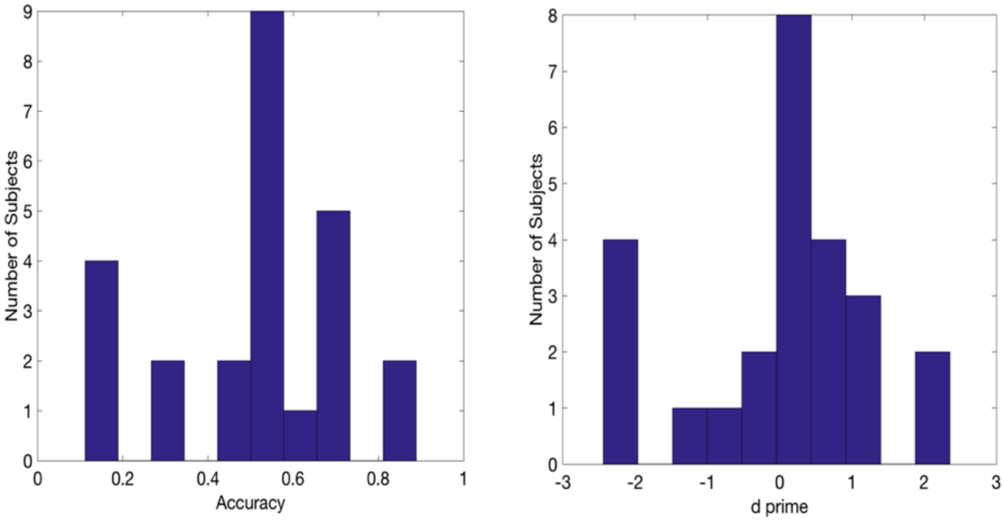


Note: Error bars indicate +/- 1 SEM.

Results

Four of 25 subjects reported detection of a target face following run 1. Subjects' response accuracy for the detection probe in the second run as well as the corresponding d' were calculated. Sample accuracy and d' were not significantly greater than .50 ($t(24) = 0.0526$, $p = 0.9585$) and 0 ($t(24) = -0.0388$, $p = 0.9694$), respectively (Figure 3). Two subjects' accuracy and d' values indicated a likelihood of awareness of the target face's emotional expression.

Fig. 3: No Conscious Perception of the Target Face

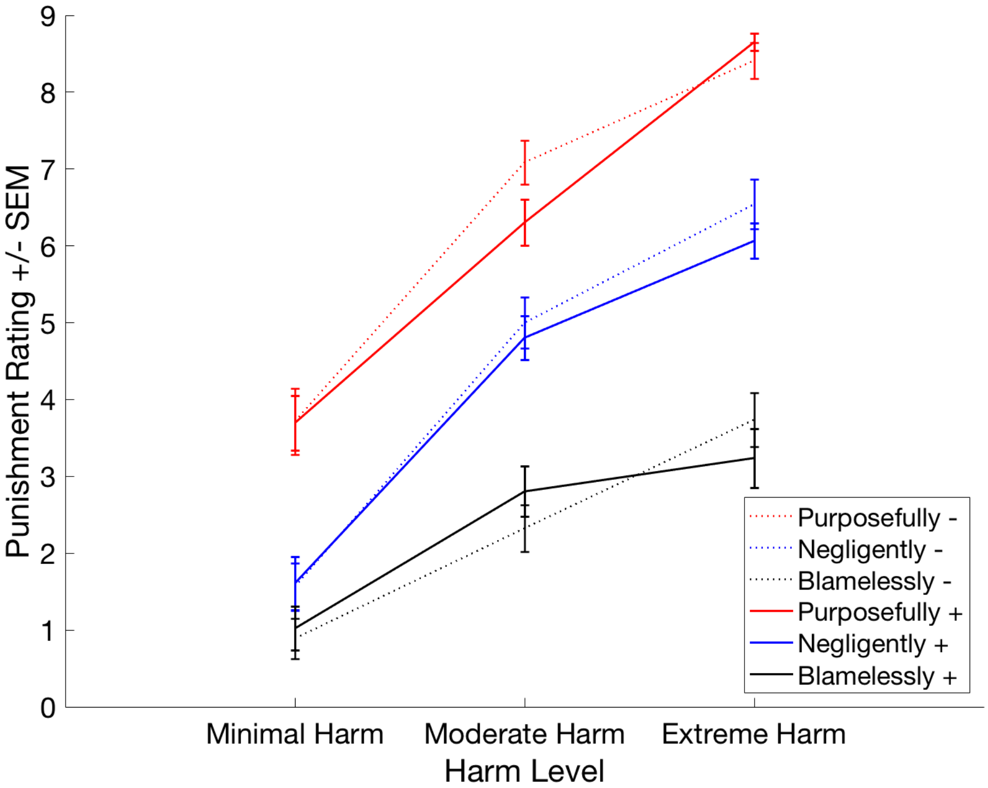


To examine the effect of treatment on punishment decisions, a 3-way analysis of variance (ANOVA) was performed, with subject as a random effect. In addition to examining a simple effect of treatment, an interaction effect of treatment with mental state, harm, and their interaction was also examined. Results are presented in Table 1. No effects for treatment were observed, either as a simple effect or as an interaction. Results were not materially affected by excluding those subjects who self-reported awareness of the target image or otherwise evinced awareness in run two. Results were also not materially affected by excluding run two, on suspicion of habituation, as demonstrated in the pilot data (Figure 2). For purposes of visualization, figure 4 displays punishment ratings as a function of mental state, harm, and treatment.

Table 1: Results of ANOVA

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
MS	1090.37	2	545.19	245.92	0.00
Harm	1153.98	2	576.99	260.27	0.00
Treatment	1.51	1	1.51	0.68	0.41
MS*Harm	82.91	4	20.73	9.35	0.00
MS*Treatment	1.28	2	0.64	0.29	0.75
Harm*Treatment	1.63	2	0.82	0.37	0.69
MS*Harm*Treatment	12.07	4	3.02	1.36	0.25
Error	877.89	396	2.22	[]	[]
Total	3221.65	413	[]	[]	[]

Fig. 4: Effect of Visual Masking of Fearful Faces on Punishment Decision



Note: Error bars display +/- 1 SEM. '+' indicates the presence of the treatment (fearful target), while '-' indicates the presence of the control (neutral target).

Discussion

We find no evidence that visually masked fearful faces affect punishment decision-making. Null effects such as these are challenging since they prove difficult to interpret. This is especially true here.

The result could, potentially, support the conclusion that activation in the amygdala does not feed forward to affect punishment decision-making. This would be contrary to the hypothesis in Buckholtz et al. (2008), Buckholtz & Marois (2012) and Ginther et al. (2016). It would also be directly opposing the main interpretation of the results in Treadway et al. (2014). Of course, as noted before, the GCA results that served as the backbone for the results in Treadway are not dispositive as to the Amygdala causally driving punishment decisions since they can also be explained by a third region driving activation in both the Amygdala and the target region (in that case, DLPFC). Instead, the result would be consistent with the amygdala's activation reflecting the result of other feed forward processes. One potential interpretation is that, perhaps, the heightened amygdala activation relates to the mnemonic encoding of the highly aversive stimuli (Hamann et al., 1999).

Unfortunately, other interpretations of the result are just as convincing, if not more so. For one, it is possible that the manipulation did not sufficiently augment activation in the amygdala to have any material effect. While the protocol was adapted from a study where amygdala activation was demonstrated using fMRI, the model study involved 168 blocked trials per condition, compared to 18 in the present study. It is thus possible that the amygdala activation was not sufficient to detect any difference on the punishment decision with the amount of power provided by the present study. Practical constraints limited our ability to have more trials per subject, per condition.

At the same time, the amygdala is known to show rapid habituation effects (Zald, 2003). This is demonstrated not only by reduced BOLD activity in the amygdala but also by suppressed behavioral effects (Dijksterhuis and Smith, 2002). We further validated the presence of habituation effects in the present task by demonstrating a robust effect for the masked fearful target in run one, but no effect in run two. Nonetheless, we found similar behavioral results for both runs, inconsistent with the theory that habituation may, alone, account for the null result. Future research may seek to sacrifice the benefit of a 3x3 design where both mental state and harm are parametrically manipulated and instead focus on a fully between-subjects design that did not allow for habituation.

Further confounding the interpretation of the results is evidence that the amygdala may not be sensitive to visually masked fearful faces. Pessoa et al. (2006) found that amygdala activation could be detected using fMRI when using a 67ms target and when the face could be detected by participants but not when using a 33ms target, as used in the present study. Pessoa argues that the Whalen et al. (1998), which served as a model for our purposes here, did not accurately conclude a lack of awareness since they relied entirely on self-reports and accuracy and did not apply a signal detection framework to support their conclusions. Pessoa et al.'s conclusion is buttressed by a 2011 study observing that masked faces, as opposed to words or other highly valent items, did not induce affective priming (Andrews et al., 2010).

Nonetheless, applying a signal detection framework to the present results, as called for in Pessoa et al. (2006), we reliably concluded a lack of awareness using the 33ms target, contrary to Pessoa's account of Whalen et al. (1998). Further, our evidence demonstrating affective priming in the pilot study weighs against the argument presented by Andrews et al. (2011) that masked emotional faces will not induce affective priming and other investigators have also identified that masked emotional faces can induce activation in the amygdala absent conscious awareness (Kim et al., 2010). Ultimately, it is difficult to isolate the different explanations for the various findings across the literature as well as in the present null effect.

3. MORAL OUTRAGE DRIVES THE INTERACTION OF HARM AND CULPABLE INTENT IN PUNISHMENT DECISIONS

INTRODUCTION

In chapters one and two we examined the neurobiological mechanisms that give rise to punishment decisions. In chapters three and four we shift to examining the cognitive and affective states that motivate punishment behavior. As noted in the introduction, punishment of defectors is essential to human cooperation and is thought to be a major factor underlying our unparalleled social, technological, and economic achievement (Fehr and Fischbacher, 2004; Buckholtz and Marois, 2012). However, even if the adaptive benefits of TPP are by now well understood, much less can be said about the proximate factors that drive TPP behavior, especially when one considers that it is often carried out in the absence of concrete and immediate benefits to the punisher (Fehr and Gächter, 2002). The answer to this riddle requires not only understanding the external factors that trigger punishment behavior, but also elucidating the source of internal motivation to respond with punishment. By now, much is understood about the elements of norm violation to which we respond: namely the harm that was caused, and the extent to which that harm was produced with intent (Carlsmith et al., 2002b; Cushman, 2008b), consistent with real-world legal practices (LaFave et al., 1986; Shen et al., 2011). It is particularly the superadditive interaction between culpable mental state and substantial harm (Cushman, 2008a; Treadway et al., 2014b) that leads to punishment. After all, we do not punish bad deeds if they occur purely accidentally nor the desire to harm another if no action is taken to do so.

But confronted with a reprehensible act committed willfully, what drives a third-party observer to respond with punishment? While folk theory suggests that TPP results from cold-headed reasoning – indeed, it is often considered a key rationale for the establishment of an uninvolved and impartial adjucator –, there is much evidence to suggest that it is also subject to emotional influence (Darley, Carlsmith, & Robinson, 2000; Gummerum et al., 2016; Salerno & Peter-Hagene, 2014). Indeed, it has been suggested that emotional responses to a crime can strongly predict the administered punishment (Buckholtz et al., 2008), in line with the widely held notion that emotions are powerful drivers of adaptive behavior (Plutchik, 1980).

If there is now converging evidence that emotions serve as the proximate driving force(s) underlying TPP, there is much debate about which specific emotions may be the drivers. Early work proposed the 'CAD' (contempt, anger, and disgust) triad hypothesis to describe the emotional response to social norm violations (Shweder et al., 1997). This hypothesis suggests that contempt is elicited in response to violations of community standards, anger by autonomy (individual rights) violations, and disgust by violations of divinity. While some research has supported this model (Rozin et al., 1999), the hypothesized 'CAD' associations with violations of community, autonomy, and divinity have been inconsistent (Hutcherson and Gross, 2011; Russell and Piazza, 2013). Other research suggests that it is expressions of both anger and disgust that strongly predict punishment severity, but their relative contributions to punishment behavior are often difficult to distinguish (Gutierrez et al., 2012; Piazza et al., 2013), if they are different at all (Nabi, 2002; Royzman et al., 2014).

The emotion of moral outrage has emerged as a potential key player in punishment behavior. Early punishment research found that moral outrage mediates the effects of offense severity and mitigating circumstances (i.e., facts that lessened the actor's culpability) on punishment (Carlsmith et al., 2002). However, because that early study did not assess other emotions, the importance of moral outrage in punishment decision-making relative to contempt, anger, and disgust remains unknown. More recent studies have examined how moral outrage may relate to other emotional states. In particular, (Salerno and Peter-Hagene, 2013) indicated that moral outrage may be related to the combined experience of anger and disgust. However, because they did not independently manipulate the nature or severity of the norm violation and did not examine the expression of moral outrage as compared to other emotions, it remains unclear how norm violations, emotional states, and punishment decisions are inter-related.

The present study is designed to understand this relationship. We achieve this by independently and parametrically manipulating both mental state culpability and harm severity, and examining how contempt, anger, disgust, and moral outrage map onto these distinct components. We further examined how these emotions may differentially mediate the relationship between the norm violation and the punishment. Because sadness is often expressed in response to scenarios involving harm to others (Winterich et al., 2010), we also examined its potential importance.

Methods

We recruited a total of 455 participants via Amazon Mechanical Turk, with each receiving $6/hour as compensation. Most participants completed the survey in 5-8 minutes. Responses were excluded from analysis if participants failed to complete the full survey or incorrectly answered an attention check question, resulting in 387 participants (Ages 19 to 76 years, Mean=37, SD=12; 52% male) ultimately included in the analysis. Sample size was chosen based on a power analysis indicating that roughly 400 participants were needed to obtain a power of 0.95 with a moderate effect size (.25) for main effects and the interaction (a moderate effect size is consistent with prior studies examining mediation effects of emotion on punishment).

After providing informed consent, participants read four scenarios (in addition to the 'attention check' scenario described further below) depicting the actions of a protagonist named 'John' that resulted in harm to another person. These four scenarios consisted of three anchoring scenarios (which were not analyzed and were indicated as practice scenarios to subjects) followed by the test (i.e., analyzed) scenario. The anchoring scenarios served to introduce subjects to the task design and the full spectrum of possible harm and mental state levels, as well as the punishment response scale.

The test scenario was derived from one of 64 different scenario stems previously used by Ginther et al. (2016), with each stem describing a specific set of events. The 64 stems varied in the level of harm that John caused – from negligible to moderate to severe and to death – thus producing 16 stems in each of the 4 harm categories. Each of the individual 64 stems could vary amongst four different levels of John's mental state in causing the harm. These mental states corresponded to those of purposeful, reckless, negligent, and blameless conduct as described in the Model Penal Code's mental state hierarchy. With 64 different stems, each of which with four different possible mental states, there was a total of 256 scenarios from which the test scenario was randomly sampled for each subject.

Each scenario was presented in three phases. The first phase was an introduction sentence that provided relevant background information about the scenario. The second and third phases presented either the mental state of the actor or the resulting harm, with the order of presentations of the mental state and harm counterbalanced across subjects (i.e harm in second phase was followed by mental state in the third phase, or vice versa) (see Table 1 for an example fact pattern).

Table 1a: Illustrative Theme (Planks & Bikes): Four "Mental-State First" Variations.

Introductory Sentence (Phase 1)			
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.			
Mental State Sentence (Phase 2)			
Purposeful Mental State	Reckless Mental State	Negligent Mental State	Blameless 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.	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.	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.	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 (Phase 3)			
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.			

Table 1b: Illustrative Theme (Planks & Bikes): Four "Harm First" Variations.

Introductory Sentence (Phase 1)			
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.			
Harm Sentence (Phase 2)			
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.			
Mental State Sentence (Phase 3)			
Purposeful Mental State	Reckless Mental State	Negligent Mental State	Blameless 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.	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.	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.	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.

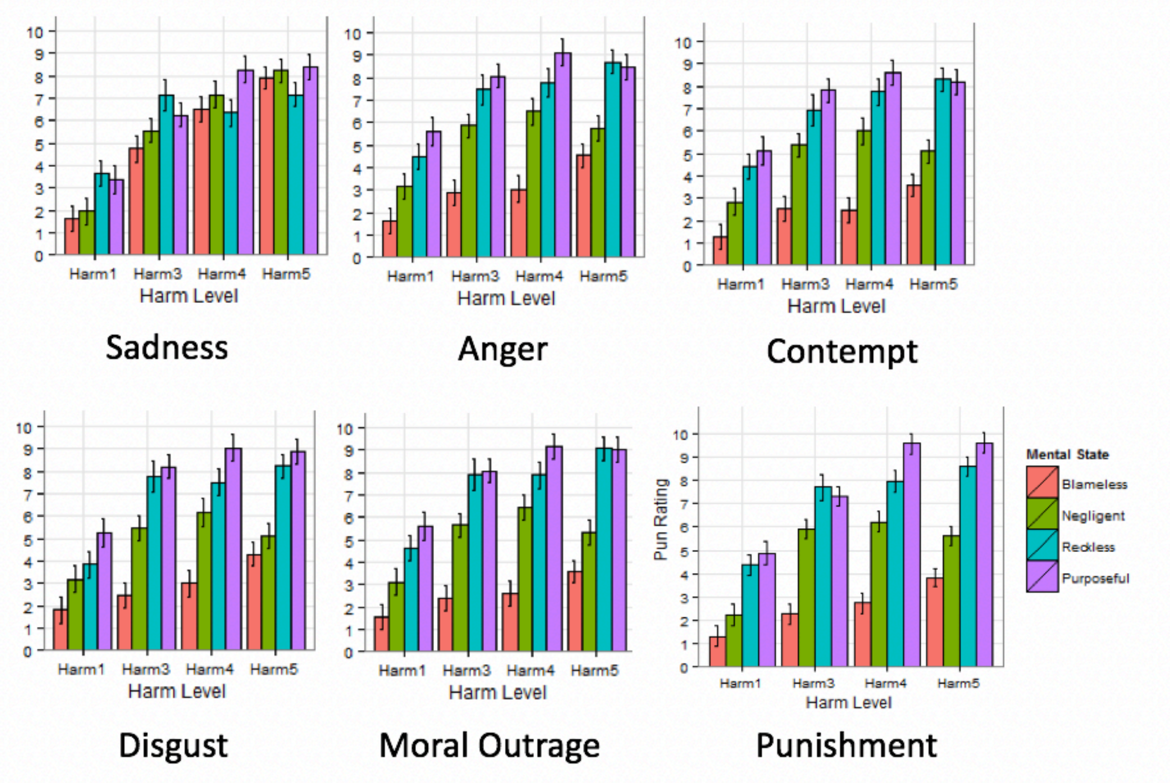
Note: This example is drawn from one of 64 possible themes. Further, subjects evaluated only one of the possible eight scenarios from the theme they evaluated.

Subjects' emotions were assessed as described below at two points during each trial. First, after the presentation of the second phase (i.e. after presentation of either harm or mental state) and second, after the presentation of the third phase (i.e. after presentations of both harm or mental state). Thus, subjects provided two distinct emotional responses: one after only being aware of either the mental state or the harm, and another after becoming aware of the full contents of the scenario. The purpose of the first emotional assessment was to isolate the emotional response to harm and mental state independent of one another, whereas the second assessment was used determine the emotional response to the integration of harm and mental state. After the second, and final, emotion assessment subjects provided a punishment rating, also as described below. Subjects' progression through these steps was self-paced.

A pilot study (N=400) in which participants were asked to rate the extent to which they experienced anger, disgust, contempt, moral outrage, and sadness (each on a 10-point Likert scale with "0" as "Not at all" and "9" as "Extremely") after reading the scenarios established that collecting emotion responses in this fashion (i.e. with multiple Likert ratings) would result in data that could not be used to disambiguate the emotions of anger, contempt, disgust, and moral outrage as they relate to human punishment behavior.

In the pilot study for this experiment (N=400), when providing the emotion ratings participants were asked to rate the extent to which they experienced anger, disgust, contempt, moral outrage, and sadness, each on a 10-point Likert scale with "0" as "Not at all" and "9" as "Extremely". All emotions were presented simultaneously, with the ordering on the screen randomly determined, on the screen. For the punishment rating, participants were asked to indicate how much they felt John should be punished for his behavior on a 10-point Likert scale with "0" as "No punishment" and "9" as "Most severe punishment". As part of our pre-processing procedures we analyzed the data for collinearity in the predictors (the ratings for the individual emotions would serve as predictors for the punishment ratings). We were specifically concerned with potential collinearity issues because a major component of this study is focused on disambiguating the relationship of contempt, anger, disgust, sadness, and moral outrage with regards to norm violations and punishment decisions, and serial Likert scale ratings have a tendency to be correlated (Nabi, 2002; Gutierrez et al., 2012; Royzman et al., 2014). If the ratings provided for these emotions are highly collinear, regression analyses will not be suitable and will result in both spurious and inconsistent results (Baayen, 2008). The collinearity analysis on the emotion and punishment ratings did reveal high collinearity (variance inflation factors for anger, disgust, and moral outrage all >8). Plots of the emotion and punishment scores by task condition from this pilot experiment are displayed in Figure 1).

Fig. 1: Mean Emotion and Punishment Ratings for The Pilot Study Wherein All Ratings Were Acquired Using Serial Likert Scales



Sadness

Anger

Contempt

Disgust

Moral Outrage

Punishment

Note: Error bars show +/- 1 standard error of the mean.

Table 2: Principal Component Analysis of Factors When Using Multiple Likert Scales

Factor	**Component 1**	**Component 2**
Anger	0.9	0.34
Contempt	0.9	0.25
Disgust	0.89	0.22
Moral Outrage	0.93	0.28
Sadness	0.29	0.96
Punishment	0.86	0.32

In order to better understand the nature of this collinearity and whether it could be avoided by means of data reduction, we performed a principal component analysis (PCA) with varimax rotation. A scree plot analysis on the results supported the conclusion that two latent variables underlie the data: First, a component consisting of anger, contempt, disgust, moral outrage, and the punishment decision; and a second component consisting of sadness. The first component included nearly identical and high loading values from each of the emotions and the punishment ratings (Table 2). Consequently, the pilot study indicated that collecting emotion responses in this fashion would result in data that could not be used to disambiguate the emotions of anger, contempt, disgust, and moral outrage as they relate to human punishment behavior.

In order to overcome the limitations illustrated by the results of our pilot study, we implemented a new method to assess emotional responses (Rozin et al., 1999). Rather than rating each emotion, subjects were instructed to select a primary emotion (either anger, disgust, contempt, moral outrage, or sadness) that best identified their emotional state. Again, this occurred after phases 2 and 3 of scenario presentation. The different emotion response options were presented as a list, though the order of this list was randomized across trials. Subjects were instructed to click on one of the emotions to select it. After selecting a primary emotional response, subjects were asked to rate how strongly they experienced that emotion on a 10-point Likert scale with “0” as “Not at all” and “9” as “Extreme”. Participants then selected, on a subsequent screen, the second strongest emotion they felt in response to the scenario, and subsequently rated this emotional experience on a Likert scale of the same format as the primary emotion. Thus each participant provided a primary and secondary emotional response after the second phase of the scenario (having seen only information about either harm or mental state), and again after complete presentation of the scenario (with information about both harm and mental state). Measurement of the punishment response was the same as in the pilot experiment (0 to 9 Likert scale).

After completing the test scenario, participants were presented with an 'attention check' scenario. This scenario appeared identical to the test and anchoring scenarios in structure except that embedded in the scenario was a sentence with an instruction to provide a specific response to the emotion and punishment questions. This allowed us to screen out individuals who did not carefully read the scenarios. Following the attention check, basic demographic information was collected and individuals were debriefed and provided with instructions for compensation.

Our statistical analyses focused on answering two primary questions: First, what are the emotional responses to varying levels of harm and mental state, both independently and when these factors are integrated? Second, can the experience of specific emotions be linked to punishment behavior?

To address the first question, we relied on regression analyses to examine the relationships between the norm violation, emotion, and the punishment decision. For all regression analyses, predictors were standardized through z-transformation so as to enable meaningful comparison between the results of the regression.

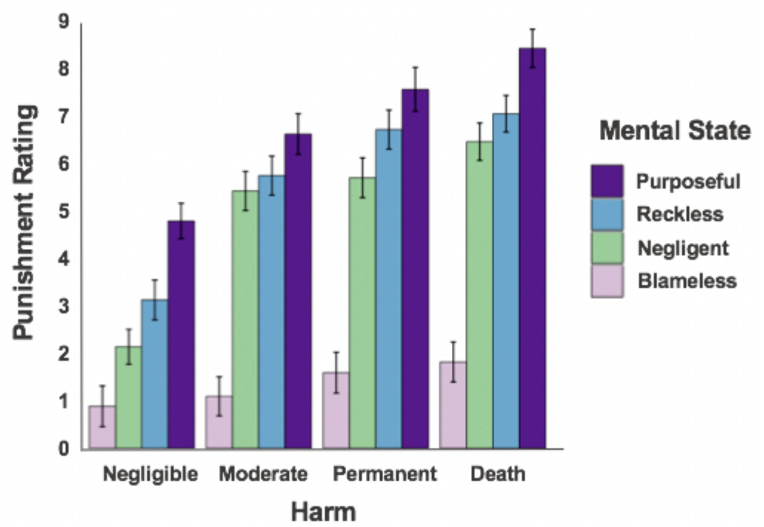
When analyzing emotion selected as the outcome variable and not as predictor (e.g. when examining the effect of the type of norm violation on the emotion the respondents selected) we used linear probability models with each emotion expressed as a binary variable (selected or not) regressed, in independent analyses, on harm level, mental state level, and their interaction, with heteroskedastic robust standard errors to allow for non-normality of variances around the probability estimates (Aldrich and Nelson, 1984). Lines of best fit for the predicted probability of selecting each emotion as a function of harm and mental state were generated to further visualize these relationships.

To address the second question, separate mediation analyses were run to test whether the expression of each emotion mediated the effect of harm, mental state, and the interaction of mental state and harm on subjects' punishment ratings. Emotions were again treated as binary variables (selected or not). We used a counterfactually defined causal mediation method, as the product-of-coefficients approach to calculating indirect effects cannot be applied to binary mediators (Steen et al.; Imai et al., 2010; Pearl, 2012; Valeri and Vanderweele, 2013). We obtained estimates for the natural indirect effect for each emotion as a mediator of the effect of the harm, mental state, and their interaction on punishment using the R package Medflex (Steen et al., 1AD). Standard errors were calculated using the bootstrap method with 1000 draws (Preacher and Hayes, 2008). Additionally, 95% confidence intervals were generated for each indirect effect.

RESULTS

Punishment behavior was characterized by not only an effect of harm (b=0.38, se=0.04, p<.001, 95% CI [0.31, 0.45]) and mental state (b=0.60, se=0.04, p<.001, 95% CI [0.53, 0.67]) but also a superadditive interaction between the two (b=0.10, se=0.04, p=.01, 95% CI [0.03, 0.17]) (Figure 2), consistent with prior findings (Ginther et al., 2016).

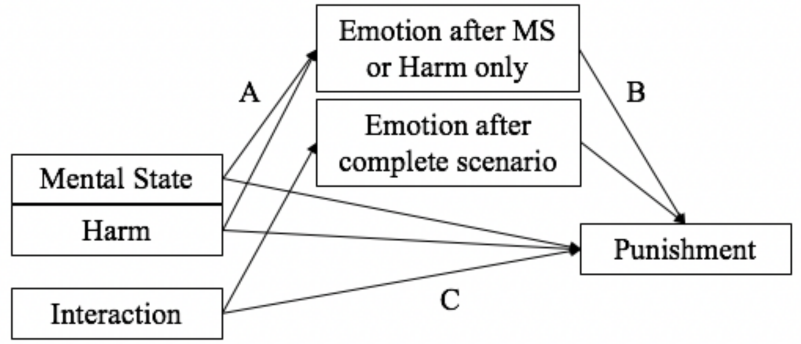
Fig. 2: Mean Punishment Ratings as a Function Of Mental State And Harm Level



Note: Error bars display +/- 1 standard error of the mean.

How do presentations of harm and mental state – independent of one another – affect the emotional response of the participants and, ultimately, their punishment decision? To assess the first part of this question, we examined subjects' first emotion response, after they had only been presented with information as to the level of harm or mental state, but not both. Specifically, we tested for linear trends in subjects' emotional responses as a function of level of mental state and level of harm, independently (path "A" in Figure 3).

Fig. 3: Graphical Depiction of Regression and Mediation Analyses Performed in the Present Study

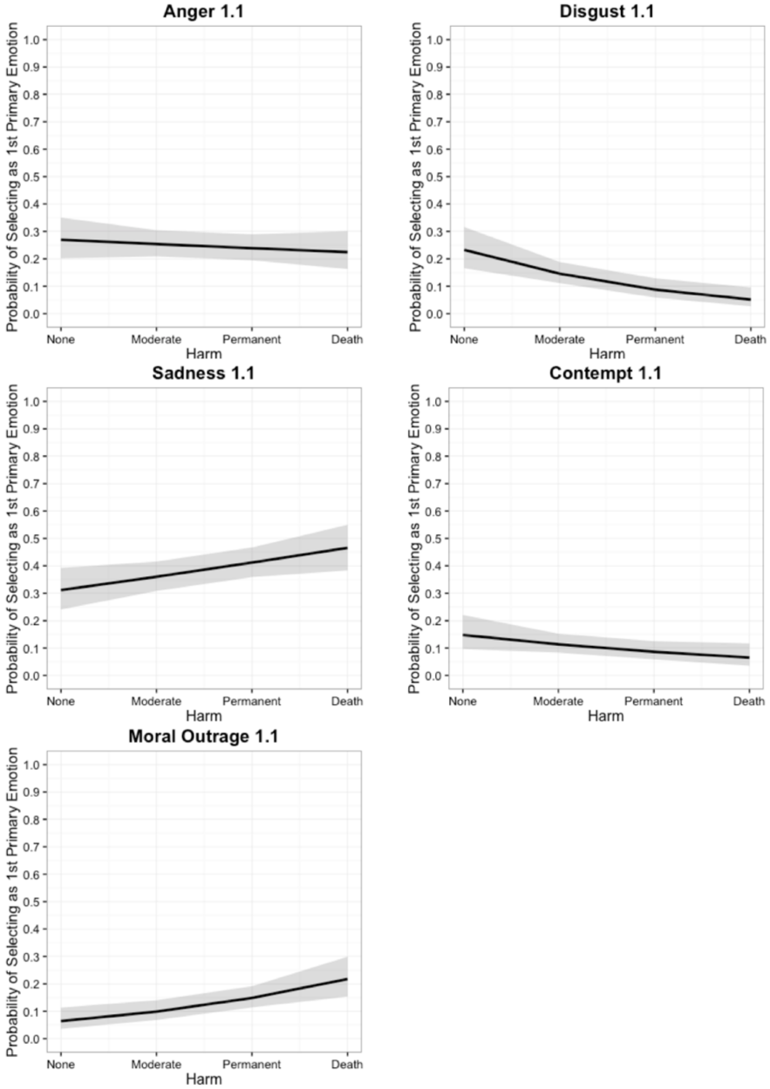

Emotion after MS or Harm only
Emotion after complete scenario
A
B
Mental State
Harm
Interaction
Punishment
C

Note: Independent regression analyses were performed with level of mental state, harm, and the interaction as predictors. Linear regression was performed between 3 different paths: (1) path 'A' between the stimulus (either Mental State, Harm, or the interaction) and the emotion response, (2) path 'B' between the emotion response and the punishment amount, and (3) path 'C' between the stimulus and the punishment amount. Note that when examining the effect of Mental or Harm alone, the emotion selected following the presentation of that respective predictor (Mental State or Harm) was used. For the interaction we used the emotion selected after subjects evaluated the complete scenario. In a separate analysis, to determine the presence of a mediation effect, for each set of predictors (Mental State, Harm, and their interaction) we examined the size of the indirect effect (a*b[1]) in the presence of the direct effect (c).

This revealed that expressions of sadness and moral outrage increased with increasing levels of harm severity while disgust and contempt decreased (note beta values in "Path A" column in Table 3 and see Figure 4 for a graphical depiction). For mental state, expressions of anger, disgust, and moral outrage increased as harm severity increased while sadness decreased (note beta values in "Path A" column in Table 4 and see Figure 5 for a graphical depiction).

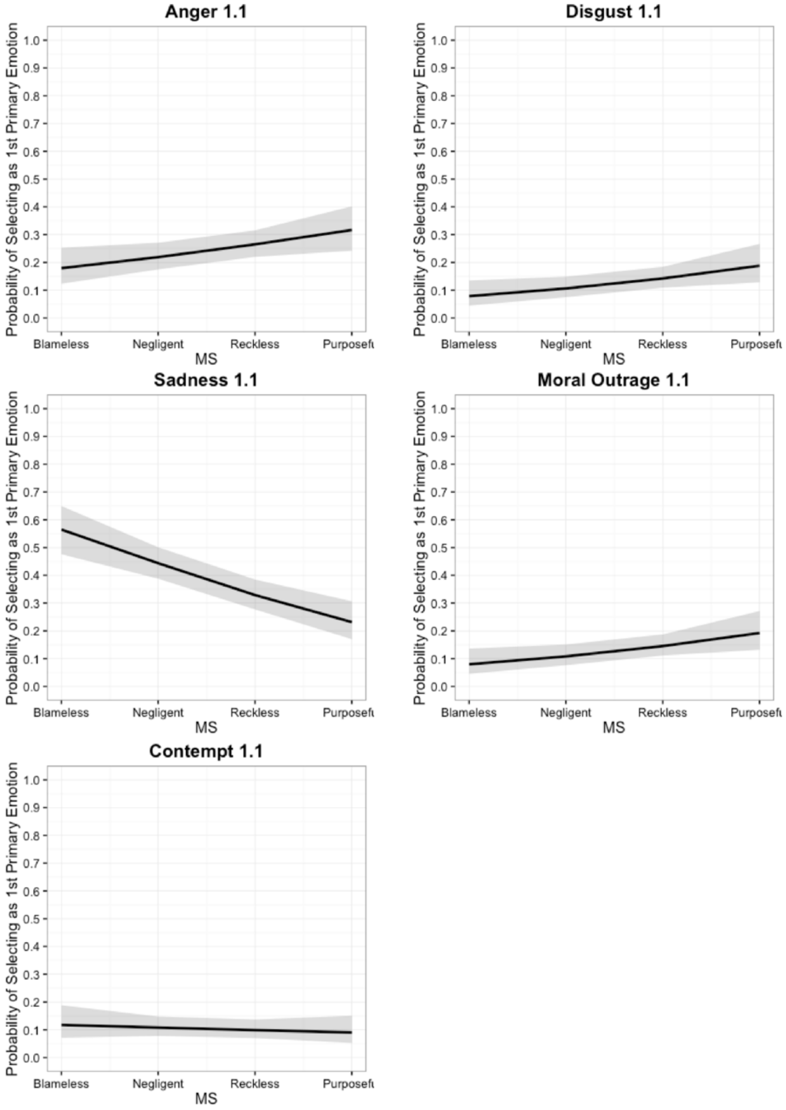
[1] To be sure – as noted in the methods – we used a counterfactually defined causal mediation method in order to calculate the indirect effect due to concerns about the effect of a binary regressor (the emotion selection) on the product of coefficients approach (a*b). Nonetheless, the product of coefficients approach provides nearly identical results in this case and is helpful for conceptualizing the nature of the mediation effect.

Fig. 4: Predicted Probability Fits for Each Emotion for The First Emotion Rating When Harm Was Presented First



Note: Shaded areas depict 95% confidence intervals.

Fig. 5: Predicted Probability Fits for Each Emotion for the First Emotion Rating When Mental State was Presented First



Note: Shaded areas depict 95% confidence intervals.

Using a mediation analysis, we assessed which of these emotional responses to harm or mental state might demonstrate an indirect effect on punishment decisions. The mediation analyses examine the size and directionality of the indirect effects (Path A and Path B in Figure 3) in the presence of the direct effect (Path C in Figure 3). Moral outrage manifested an indirect effect for both mental state and harm, while sadness manifested an indirect effect for harm alone.

No other mediation effects were observed (see “Mediation Effect” columns in Tables 3 & 4).

Table 3: Relationship Between Harm, Emotion, and Punishment

	Path A: Harm to Emotion				Mediation Effect on Punishment			
Emotion	*b*	se	*p*	95% CI	*b*	se	*p*	95% CI
Anger	-0.02	0.02	0.485	[-0.06, 0.03]	-0.004	0.01	0.525	[-0.02, 0.01]
Disgust	-0.07	0.02	<.001**	[-0.10, -0.04]	-0.01	0.01	0.466	[-0.03, 0.02]
Contempt	-0.03	0.02	0.035	[-0.07, -0.002]	0.01	0.01	0.389	[-0.01, 0.02]
Sadness	0.06	0.03	<.001**	[0.01, 0.11]	-0.03	0.01	0.024*	[-0.05, -0.004]
Moral Outrage	0.06	0.02	0.001**	[0.02, 0.09]	0.02	0.01	0.032*	[0.003, 0.05]

Note: “Path A: Harm to Emotion” column presents standardized regression coefficients of path A (see Figure 2) between the harm stimulus and the emotional response for each emotion at the first emotion rating (when only the harm had been presented). The “Mediation Effect on Punishment” column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
**p* < .05. ***p* < .005.

Table 4: Relationship Between Mental State, Emotion, and Punishment

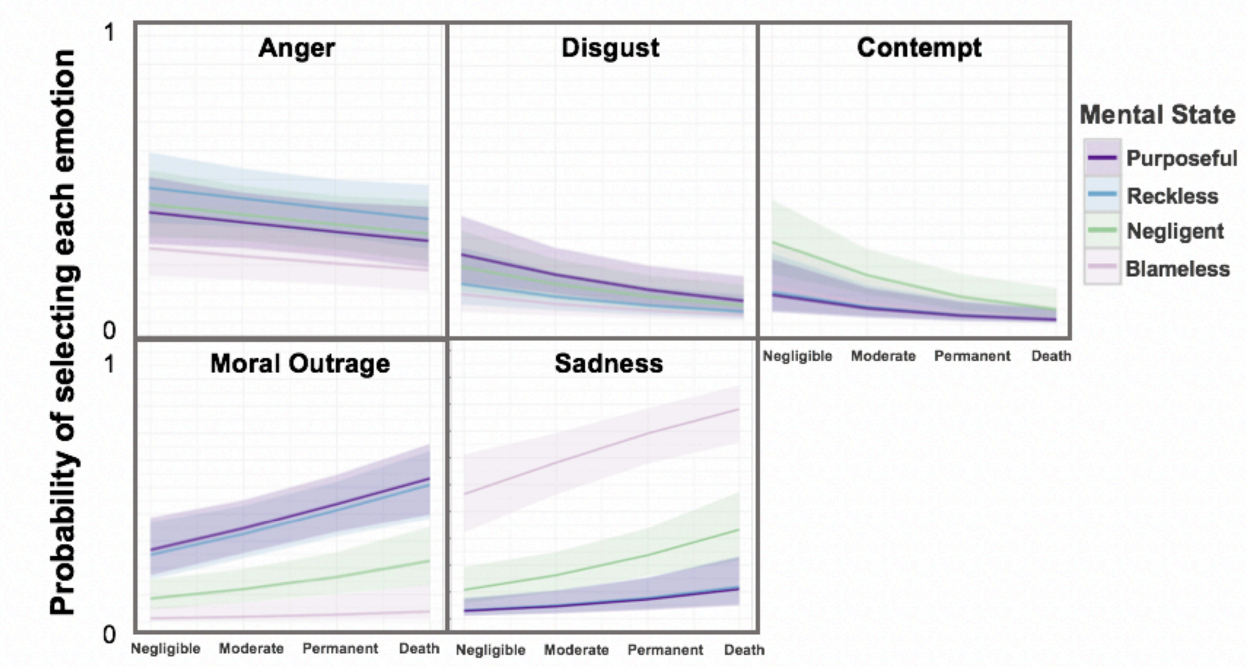
	Path A: Mental State to Emotion				Mediation Effect on Punishment			
Emotion	*b*	se	*p*	95% CI	*b*	se	*p*	95% CI
Anger	0.05	0.02	0.034*	[0.004, 0.09]	0.01	0.01	0.259	[-0.01, 0.02]
Disgust	0.05	0.02	0.006*	[0.01, 0.08]	-0.01	0.01	0.126	[-0.03, 0.004]
Contempt	-0.01	0.02	0.579	[-0.04, 0.02]	0	0	0.633	[-0.01, 0.01]
Sadness	-0.13	0.02	<.001**	[-0.18, -0.08]	0	0.01	0.745	[-0.02, 0.02]
Moral Outrage	0.04	0.02	0.014*	[0.01, 0.08]	0.02	0.01	0.037*	[0.001, 0.04]

Note: “Path A: Mental State to Emotion” column presents standardized regression coefficients of path A (see Figure 3) between the mental state stimulus and the emotional response for each emotion at the first emotion rating (when only the mental state had been presented). The “Mediation Effect on Punishment” column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
*$p < .05$. **$p < .005$.

Thus, while various emotions tracked increasing levels of either mental state or harm, only moral outrage increased commensurately with both increasing levels of harm and mental state. Further, only moral outrage and sadness displayed a mediating effect between the triggering factors (harm, mental state) and the punishment response.

In addition to examining the characteristics of mental state and harm independently, we examined the integration of mental state and harm and the commensurate effect on punishment by repeating the above analyses on the second emotional response (*i.e.,* the response provided after the evaluation of both mental state and harm). While before we tested for linear trends as a function of mental state and harm independently, here we examine a linear trend with the interaction of mental state and harm.

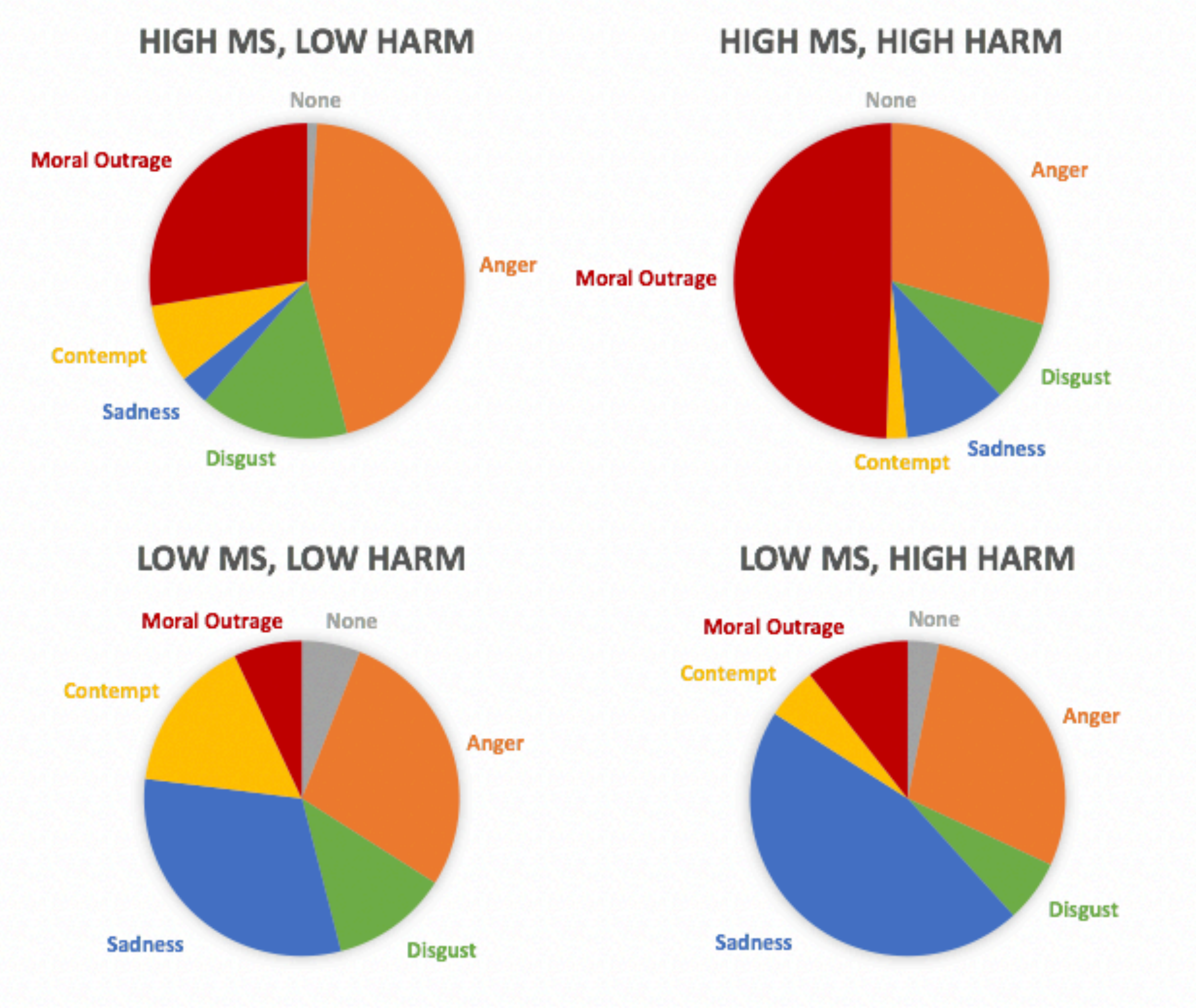
Fig. 6: Lines of Best Fit for The Predicted Probability of Selecting Each Emotion in the Second Emotion Rating as a Function of Harm and Mental State



Note: Shaded areas depict 95% confidence intervals.

Moral outrage, contempt, and sadness all demonstrated an interaction between the mental state and harm factors (see "Path A" column in Table 5 and Figure 6 for a graphical depiction). Importantly, we only observed a superadditive interaction–paralleling the punishment behavior–in the case of moral outrage: there was no effect on anger and disgust, and the interaction of harm and mental state *negatively* predicted contempt and sadness. Specifically, subjects experienced sadness primarily in response to unintentional harms, and the likelihood of experiencing sadness scaled with the severity of the harm (Figure 6 and Figure 7). The interaction of harm and mental state also negatively predicted contempt, with the interaction effect driven by a large proportion of subjects experiencing contempt for negligent accidents that resulted in low severity harms (Figure 6).

Fig. 7: Emotion Response Proportions for The Second Emotion Rating as a Function of Different Interactions of Mental State and Harm



Note: High MS refers to Purposeful and Reckless; Low MS refers to Blameless and Negligent. High Harm refers to life altering and death; Low Harm refers to negligible and substantial.

Table 5: Relationship Between the Interaction Of Mental State And Harm, Emotion, And Punishment

	Path A: Interaction to Emotion				**Mediation Effect on Punishment**			
Emotion	*b*	se	*p*	95% CI	*b*	se	*p*	95% CI
Anger	-0.002	0.02	0.943	[-0.05, 0.05]	0	0	0.945	[-0.01, 0.01]
Disgust	-0.01	0.02	0.385	[-0.05, 0.02]	-0.003	0.01	0.526	[-0.02, 0.01]
Contempt	-0.05	0.01	0.001**	[-0.07, -0.02]	-0.004	0.01	0.529	[-0.02, 0.01]
Sadness	-0.09	0.02	<.001**	[-0.14, -0.05]	0.07	0.02	<.001**	[0.03, 0.11]
Moral Outrage	0.16	0.02	<.001**	[0.11, 0.20]	0.06	0.01	<.001**	[0.03, 0.08]

Note: "Path A: Interaction to Emotion" column presents standardized regression coefficients of path A (see Figure 2) between the interaction term and the emotional response for each emotion at the second emotion rating (when both mental state and harm had been presented). The "Mediation Effect on Punishment" column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
**p* < .05. ***p* < .005.

The above analyses indicate that moral outrage was the only emotion sensitive to increasing harms and culpable intent, and that moral outrage displayed a superadditive interaction effect in relation to mental state and harm that mirrors the superadditive effect present in subjects' punishment decisions (as in Figure 2). To determine whether the expression of moral outrage in response to the interaction of harm and mental state might drive the superadditive punishment behavior, we assessed whether that emotion – or any other – mediated the effect of the interaction of harm and mental state on subjects' punishment decisions. Mediation effects were present only for sadness and moral outrage, though they mediated the opposite effects on punishment (see "Mediation Effect" column Table 5). Specifically, subjects were more likely to experience moral outrage when there were high levels of harm and culpable mental states, but less likely to experience sadness. Furthermore, expression of moral outrage was associated with greater punishment while the expression of sadness was associated with lesser punishment. We compared the strength of this mediation effect with the strength of the mediation effect that moral outrage displayed for harm and mental state independently. The mediation effect for the interaction is substantially greater than for either harm or mental state ($Z = 2.86$, $p = 0.0043$) (Paternoster et al., 1998).

Discussion

Our results indicate that the expression of moral outrage is uniquely critical to third-party punishment (TPP) behavior in humans in two ways. First, we demonstrate that while most of the emotions tested respond to increases in harm or increases in mental state, moral outrage is selectively expressed by the interaction of culpable intent and harmful outcomes. Second, we observe that the expression of moral outrage selectively mediates the effects of both harm and mental state on punishment decisions. We discuss these two links between moral outrage and punishment below.

As noted in the introduction, recent findings have observed that punishment behavior is characterized by a superadditive effect of culpable intent and harmful outcome (Ginther et al., 2016; Treadway et al., 2014). A primary goal of the present study was to determine what emotion(s) may be associated with the interaction of these two components. While a number of studies have investigated the relationship between emotion and punishment decisions (Fehr and Gachter, 2002a; Kogut, 2011; Salerno and Peter-Hagene, 2013; Laurent et al., 2014), few have empirically investigated the types of norm violations that induce the expression of specific emotions (Rozin et al., 1999; Hutcherson and Gross, 2011; Russell and Piazza, 2013). Moreover, to our knowledge, none have examined the emotional response specific to the interaction of a culpable mental state and severe harms, the lynchpin of punishable behavior. Our observation that the expression of moral outrage is selective to, and predominates at, the intersection of culpable intent and harmful outcomes is consistent with Carlsmith (2002) who put this emotion at the junction of serious offenses and the absence of mitigating circumstances. However, Carlsmith (2002) did not demonstrate that moral outrage – as compared to contempt, anger, or disgust – uniquely reflected the intersection of these twin components that define human punishment behavior. By revealing that subjects overwhelmingly expressed moral outrage at this junction, our results indicate that moral outrage is uniquely reflective of the human emotional response to severe harms that are the result of culpable conduct.

This aspect of our results does conflict, in part, with a recent study by Landmann and Hess (2016) that found that anger ratings were predicted by moral violation regardless of the outcome caused. They concluded that this anger at the intent to commit a violation independent of the harm can be defined as moral outrage. While our findings do support the notion that anger is the predominant response in the case of culpable intent without a harmful outcome (see Figure S4), our results are inconsistent with their conclusion that moral outrage is selective to intent. This discrepancy in findings is likely due to the fact that Landmann and Hess did not explicitly gauge subjects' moral outrage but instead relied on inference to reinterpret expressed anger as moral outrage.

The second major finding of the present study, derived from the mediation analysis, is that expression of moral outrage is selectively linked to augmented punishment decisions in the case of severe intentional harms. Insofar as moral outrage appears to be associated with punishment decisions, this result is consistent with two prior studies (Carlsmith, 2002; Salerno & Peter-Hagene, 2013). However, in contrast to these two prior studies, we demonstrate that expression of moral outrage is, in this way, unique among emotions. That we did not observe an indirect effect of contempt, anger or disgust on punishment decisions is evidence that these emotional states are not, even when experienced by subjects, influencing the punishment outcome in the same manner as moral outrage. Thus, it would appear that moral outrage may be a unique emotional motivator linking severe culpable norm violations and punishment of third-parties.

Studies of punishment decision-making often speak only of the emotional drivers of punishment, not of emotions that may act to suppress action (Carlsmith et al., 2002a; Salerno and Peter-Hagene, 2013). Two of our results support a conclusion that sadness may operate in such a fashion. First, we observed that sadness is the predominant response, at an eight-to-one ratio, when a severe harm is unintentionally caused, or put another way, when an accident occurs. Second, and critically, expression of sadness induced the opposite effect to moral outrage of the norm violation on punishment; that is, sadness mediated reduced punishment ratings. Previous studies have found evidence that sadness can blunt subsequent anger and reduce the influence of anger on cognitive judgments (Winterich et al., 2010), perhaps as a result of empathy towards the offender (Skorinko et al., 2014), thus reducing punishment. While further research is necessary to establish a causal link between expression of sadness and punishment suppression, this observation can potentially have substantial real-world application in both judicial and policy domains.

Together, our results present further support that moral outrage may be an emotional experience distinct from the experience of anger, contempt, and disgust, which have all been studied much more extensively in the literature, especially in the realm of punishment decision-making (Gutierrez et al., 2012; Russell and Piazza, 2013). Whether or not moral outrage is a specific emotion or instead a combination of distinct cognitive, affective, and behavioral states is debatable and perhaps, ultimately, an issue of semantics (Haidt, 2001; Batson et al., 2007; Hutcherson and Gross, 2011; Cameron et al., 2015; Landmann and Hess, 2016). What is clear, however, is that when subjects are asked to identify their emotions after evaluating a severe culpable harm they select moral outrage over anger at a two-to-one ratio and disgust at a three-to-one ratio. Combined with the results from the regression and mediation analyses, the evidence indicates that moral outrage – more so than contempt, anger, or disgust – reflects the emotional backbone of human TPP behavior.

4. DISTINCT EMOTIONAL PROFILES FOR SECOND- AND THIRD-PARTY NORM VIOLATIONS

INTRODUCTION

In chapter three we reveal the central role of moral outrage in third-party punishment (3PP) behavior. This result raises the question of whether moral outrage similarly drives punishment by second-parties (2PP), or if it is unique to third-party norm violations.

Prior research has failed to provide any evidence of 3PP outside of the human species (Riedl et al., 2012). Furthermore, studies have demonstrated that 3PP has a distinct developmental timecourse from 2PP (McAuliffe et al., 2015). And thus, while 2PP and 3PP are–in many ways–similar, they may be fundamentally distinct in how they evolved and, commensurately, in the cognitive mechanisms that support them. One way of exploring this possibility is by contrasting the cognitive and affective states that arise in response to the conditions that induce each behavior (i.e., second-party (2P) and third-party (3P) norm violations).

As detailed in chapter three, a number of studies have done this by exploring the emotional response to norm violations and accompanying punishment decisions. Unfortunately, few have focused on 2PP and rarely have studies compared 2PP and 3PP specifically. Early work exploring the emotional correlates of 3P norm violations proposed the 'CAD' (contempt, anger, disgust) triad hypothesis as a means of connecting emotional responses to norm violations (Rozin et al., 1999; Shweder et al., 2013). This hypothesis associates contempt with violations of community/hierarchy, anger with violations of autonomy/individual rights, and disgust with violations of divinity/purity. Although, as noted in chapter three, these proposed associations have been found to be relatively inconsistent, researchers have nonetheless built on the ‘CAD’ triad hypothesis to inform our understanding of 3PP (Hutcherson and Gross, 2011; Russell and Piazza, 2013).

In regards to 2P norm violations, few studies have examined the corresponding emotional response. Those that have, have identified anger as the primary motivating emotion (Small and Loewenstein, 2005). Anger is typically evoked in response to unequal treatment or the violation of a social norm and has been associated with increased 2PP (Pillutla and Murnighan, 1996).

Directly comparing 2P and 3P is challenging because it is difficult to create a task that can provide for a 2P and 3P violation that controls for other variables while also being believable to subjects. There have been some attempts, however. Batson et al. (2007) had participants rate their anger experiencing an unfair allocation or, alternatively, witnessing a third-party receive an unfair allocation. They found that individuals expressed greater anger in response to a personal harm (second-party) than for third-party violations (Batson et al., 2007). In contrast, Hardecker et al., 2016 found no difference in anger expression in response to norm violations between 2PP and 3PP conditions for children aged 2-3 years.

While contempt, anger and disgust have received the lion’s share of the research, recent work has called attention to the role of moral outrage in 3PP (Carlsmith et al., 2002; Salerno and Peter-Hagene, 2013). The authors have identified that moral outrage is particularly evoked by 3P norm violations, and it–unlike anger, contempt, and disgust–mediates the relationship between the norm violation and the punishment decision (see chapter three). Left unanswered from the latter study is whether moral outrage plays a similar role in responding to 2P norm violations or if it is specific to 3PP. A few scholars have hypothesized that moral outrage is 3P specific (Batson et al., 2007; Landmann and Hess, 2016), though this has not been empirically tested. The present experiment seeks to answer this very question by comparing the emotional profiles of those who experience both second and third-party norm violations under similar conditions. We hypothesize that moral outrage may be expressed specifically to 3P norm violations, lending support to the possibility that humans' unique proclivity towards 3PP may be rooted in the neurocognitive experience of moral outrage.

METHODS

We recruited 360 subjects to play one round of a widely used economic game (the "Investment Game") where individuals participated as either second or third parties to a norm violation and also made punishment decisions.

In the Investment Game a player, labeled the Investor, is bestowed points. Investors have the option of investing none, some, or all of these points with a Trustee. Invested points are multiplied threefold. The rub of the game, however, is that the Trustee can keep as much of the investment, and the returns on the investment, without any obligation, beyond social norms, to return some of the investment to the Investor (Berg et al., 1995).

We modified the game slightly by adding two features, both of which have been successfully applied in previous studies. First, we provided Investors with the ability to spend some of their money on punishing (or rewarding) the Trustee at the end of the game (Pedersen et al., 2013) by paying to adjust their ultimate payout. The punishment was costly to the Investor (4:1 ratio of punishment:cost). Second, we added the role of an Observer on one half of trials (Charness et al., 2008). The Observer witnessed, without the knowledge of the Investor or the Trustee, the behavior of both and could punish (or reward) the Trustee at the end of the game just as the Investor could. As a result we could evaluate both 2PP (via Investors) and 3PP (via Observers) in the same general task.

Participants were recruited to participate in this online game through Amazon Mechanical Turk (AMT). AMT has become a widely used and accepted means of recruiting participants for social science and psychology research (Paolacci et al., 2010; Sprouse, 2010; Mason and Suri, 2011; Crump et al., 2013). In addition to convenience and cost, AMT provides a population sample far more representative than samples of convenience used in most psychological studies (Buhrmester et al., 2011). The Vanderbilt University Institutional Review Board approved the experimental protocol and all participants provided their informed consent.

In order to increase the likelihood that subjects believed they were actually interacting with other people, the study consisted of three separate stages. Subjects were told that the three stages allowed us to collect the responses of the people they were participating with before proceeding. We took this approach based on pilot data indicating that when the entire study took place in one sitting many subjects did not believe they were actually participating with other individuals.

The first stage contained the instructions for the investment game and a test for comprehension of the instructions. The test was comprised of three questions and was used to ensure careful reading of the instructions. The instructions stated that there was only one round to the game and that subjects would be randomly assigned to either the role of the Investor or the Trustee (in reality, subjects were never assigned the role of the Trustee but were, as described below, assigned the role of the Observer in one half of trials). After reading the instructions and correctly answering the comprehension questions, subjects were asked to provide their first name and provide a few sentences on why, if they were selected as the Trustee (again, this was never the case) the Investor should trust them. They were then told they would be contacted once matched with other participants.

Approximately 5-10 minutes after completing the first stage, subjects received an email to continue the study. Upon clicking the link, they were presented with a page that told them that they had either been assigned as the Investor or the Observer.

If selected as the Investor (2PP), subjects were provided the name of the person selected as the Trustee as well as their short statement, which indicated a willingness to cooperate. The name and short statement were drawn from one of two individuals' information provided during piloting of the study and were counterbalanced across subjects. Subjects assigned to be Investors were then provided with a prompt asking how much they wanted to invest from a sum of 20 "points" they were provided. Subjects were told that these points would be converted into a "not insubstantial?" bonus payment at the end of the study and that this bonus was in addition to the payment (approximately $6/hr) for their participation.

If selected as the Observer (3PP), the role was briefly explained and subjects were told that they were not previously informed of this position out of concern that knowledge of its existence would affect how Investors and Trustees acted in the 2PP condition (e.g., if subjects in the 2PP condition were aware of the Observer they may be less likely to punish the Trustee). Observers were also told that they would see the actions of the other players as they occurred. Finally, Observers were presented with the names and statements provided by the "subjects" selected as both the Investor and the Trustee. The names and statements were the same as those used in the 2PP condition, again counterbalanced across subjects. In the 3PP condition, as in the 2PP condition, it was made clear that only one round would be played.

After another 5-10 minute break, subjects received one last email that directed them to the third and final stage of the study.

Here, Investors were told what percentage of the points they invested (and that were tripled) were returned to them by the Trustee. The amount was set to be a random value between 10 and 30% of the point total held by the Trustee. Subjects were also told the monetary value of their final point total. The monetary value was fixed at $0.40, regardless of how many points they ended up with. The dollar amount that the trustee finished with was proportionally scaled to the amount of points the Trustee ended with compared to the Investor. For example, if the Trustee ended with 40 points and the Investor with 10, then the Trustee would end with $1.60 compared to the Investor's $0.40.

Observers were similarly informed of the number of points returned to the Investor by the Trustee. Again, the amount was fixed to between 10 and 30% of the total. The Observer was also provided the monetary equivalent of the ending point total, which was calculated in the same way, with the Investor's total fixed at $0.40.

Alongside this information, Investors and Observers were asked to provide their emotional state in relation to the Trustee's decision ("Which of the following emotions best describes how you feel in response to the Trustee's decision?"). Subjects could choose between Anger, Disgust, Sadness, Contempt, and Moral Outrage. These emotions were selected on the basis of chapter 3. Accompanying this question, subjects were asked to provide a rating of the strength of the selected emotion on a six-point scale from "Not at all" to "Extreme".

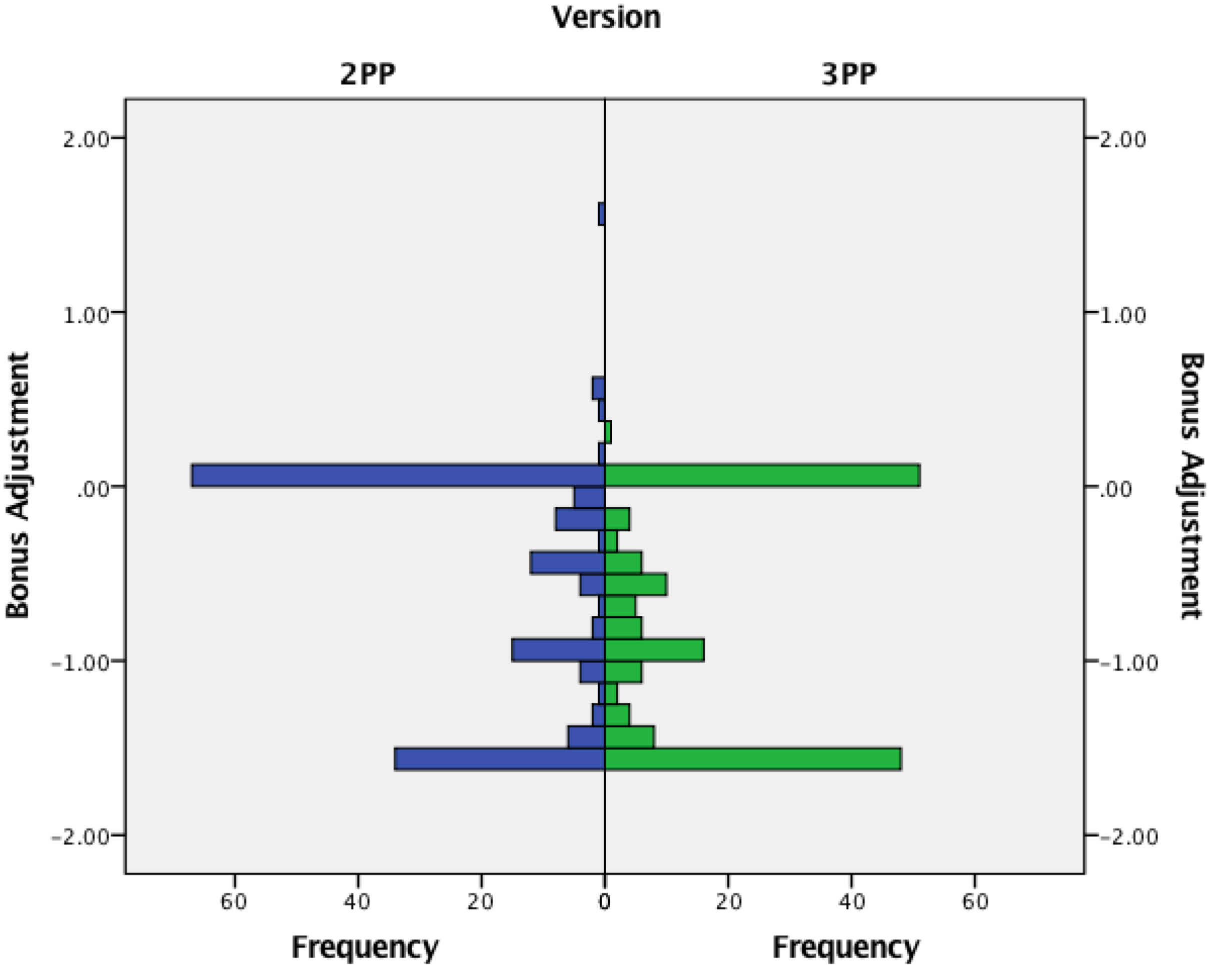
On the next screen, both Investors and Observers were informed that the Trustee had not yet been informed of the monetary equivalent of his bonus and that they could adjust the monetary payout of the Trustee at the 4:1 cost. In the case of the Investor, the cost of adjusting the score was paid out of their \$0.40 bonus. In the case of the observer, they were instructed that they received a cash bonus of \$0.40 for their participation and that any amount used to adjust the bonus of the trustee would come from that total. Subjects were told they could increase or decrease the bonus. To avoid possible demand effects, the word punish was never used.

After completing the study, participants filled out a funnel questionnaire to determine whether the participants were aware of the possibility that they were not actually interacting with other participants(Pedersen et al., 2013). Subjects that demonstrated suspicion on the basis of the questionnaire as determined by a blind researcher were excluded from the analysis. After the questionnaire, subjects were debriefed, paid, and awarded their bonuses.

RESULTS

Both second and third parties punished Trustees who acted unfairly (Figure 1). Overall, 57% of second parties punished unfair Trustees while 69% of third parties punished unfair Trustees. Average punishment amounts in the 2PP and 3PP condition were significantly greater than zero for both (one sample Wilcoxon signed rank test (2P: z = -8.17, p < 0.001 n = 167; 3P: z = -8.00, p < 0.001, n = 169)). For those participants that punished, we did not observe a difference in the amount they punished Trustees (z = 1.80, p = 0.181). Thus, it appears from the results of the behavior that 2PP and 3PP are quite similar.

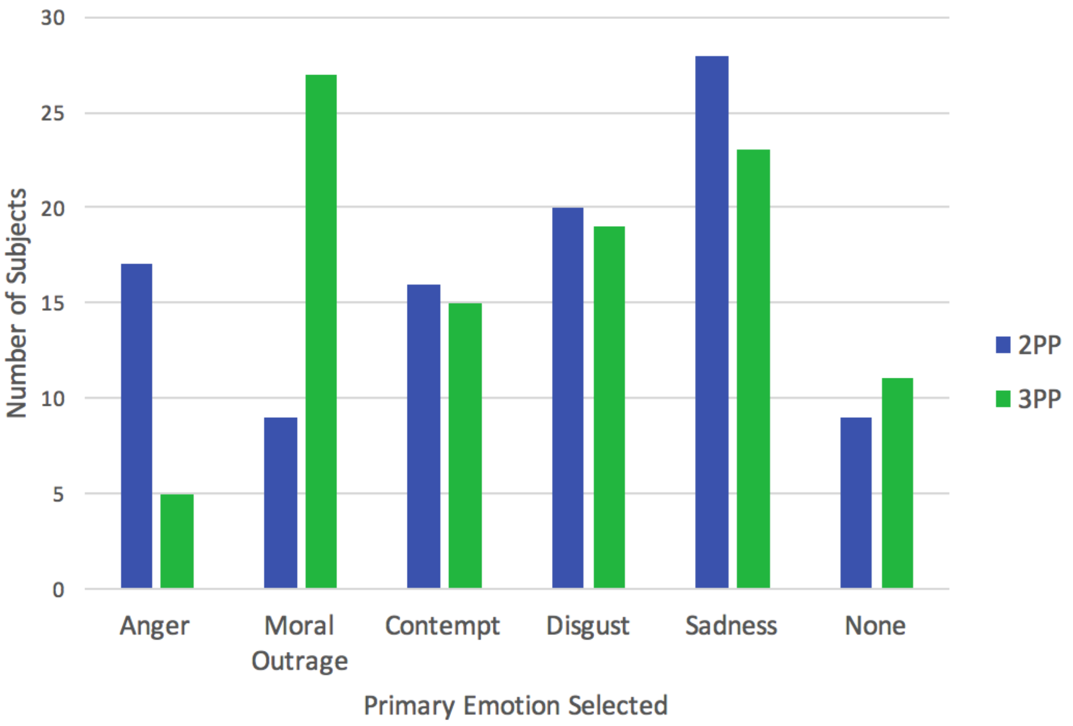
Fig. 1: Dual Histograms Displaying Amount That Subjects Changed the Trustee's Bonus In Both the 2PP and 3PP Versions of the Game



Note: The bonus adjustment cost subjects 25% of the amount of the adjustment.

Contrasting the emotion selected in response to the norm violation, we observed no difference in the frequency in which subjects selected contempt, sadness, or disgust for second-party violations as compared to third-party violations (p's > 0.3) (Figure 2). However, moral outrage was significantly more highly represented for third-party violations ($\chi(1) = 3.89$, $p = 0.048$) and anger was significantly more highly represented for second-party violations ($\chi(1) = 6.00$, $p = 0.014$).

Fig. 2: Number of Subjects Selecting Each of The Emotions as Their Primary Emotional Response to The Second- Or Third-Party Norm Violation



Discussion

Both 2P and 3P norm violations spur punishment behavior in most individuals. As an initial matter, the parity between the frequency of 2PP and 3PP indicates that subjects found a norm violation to have occurred in both conditions. However, the parity also stands in contrast to a number of studies that have indicated that 3PP is less robust when compared to 2PP and even that 3PP is, perhaps, illusory, a product of methodological limitations in the laboratory studies that produce the behavior (Pedersen et al., 2013). Since the present study accounts for the methodological issues raised in Pedersen et al. (demand effects, audience effects, inadvertent consequences of using the strategy method and affective forecasting), our results support the conclusion 3PP behavior is reliably expressed in humans. We believe our finding that 3PP is expressed at similar levels as 2PP can be attributed to the efforts made in the study design to create a specific norm (the promise to cooperate), which was expressly violated. To our knowledge most studies rely on the expectation that the other will cooperate, which may or may not be a reasonable expectation in a staged interaction. Additionally, our efforts to make subjects believe that they were actually interacting with other participants should also contribute to the validity of our construct.

Though 2P and 3P norm violations appear to be associated with a similar frequency and amount of punishment, our results indicate that they are not associated with a similar emotional response. Specifically, the data indicate that expression of anger is largely limited to 2P violations while expression of moral outrage is similarly limited to 3P violations. Responses of contempt, disgust, and sadness, however, appear to be non-distinct between 2P and 3P violations.

The contrast in the expression of anger and moral outrage for 2P and 3P norm violations has two likely explanations. First, moral outrage may simply be a synonym for anger in a third-party context(Batson et al., 2007). That is, the cognitive, affective, and behavioral properties of moral outrage may be identical to that of anger but subjects in a 3P context label the anger as being moral outrage. Alternatively, moral outrage may be a distinct emotional experience from anger. Our study is limited in its ability to distinguish between the two interpretations.

Nonetheless, we believe the evidence favors the interpretation that moral outrage reflects a distinct neurocognitive state. Three observations are particularly compelling. First, studies in young children reveal that the onset of 3PP lags in relation to 2PP (McAuliffe et al., 2015). If moral outrage were simply anger by another name, it would be more likely that 3PP would have a similar developmental timecourse. Second, that researchers have been unable to observe 3PP in other species indicates that 3PP is supported by distinct neurocognitive mechanisms (Riedl et al., 2012). Of course, there are other explanations for both of these observations beyond the possibility that moral outrage and anger are distinct. For instance, the lack of 3PP in young children and non-human primates may be due to the failure of both to detect 3P violations and not the fact that 3PP is motivated by a distinct emotional experience. Our third observation–that expressed moral outrage, but not anger, mediates the relationship between norm violations and punishment decisions (see chapter 3)– does indicate that anger and moral outrage do play distinct roles in motivating 3PP.

Distinguishing whether or not moral outrage and anger are the same neurocognitive state may ultimately be difficult to determine solely on the basis of behavioral observations. Neuroimaging may thus prove helpful. One study has indicated distinguishable brain activity in a 2PP and 3PP task (Corradi-Dell'Acqua et al., 2013), though the design did not allow for attributing this distinction with the subject's emotional experience. Some studies have purported to be able to distinguish emotional state using multivariate analyses (Saarimäki et al., 2016), though the validity of these conclusions has proven a matter of debate (Clark-Polner et al., 2016).

CONCLUSION

Across four separate studies we examined the properties of human punishment decision-making. Specifically, we sought to explore not only the brain systems that support punishment decisions, but the conditions that motivate punishment more generally.

In the behavioral component of chapter one we revealed how punishment appears to be driven by the superadditive interaction of a culpable mental state and a harm. Using fMRI, we built upon this behavioral foundation to identify neural systems that supported the evaluation of the separate components of mental state and harm and their integration. Our findings further 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.

While chapter one indicated a key relationship between the profile of amygdala activation and subjects' punishment decisions, the data were unable to indicate whether or not a causal relationship existed. Chapter two attempted to provide evidence of a causal relationship by inducing targeted activation in the amygdala during a punishment decision-making task. In other words, chapter two assessed whether targeted activation of the amygdala could shift punishment decisions, and if so, how? The null results, however, left open many possible interpretations and thus proved largely uninformative.

While the first two chapters focused on the brain mechanisms that support punishment decision-making, chapters three and four examine what motivates people to punish, with a specific attention to the role of emotion.

Chapter three focused on identifying the emotional correlates of the superadditive interaction of mental state and harm identified in the examination of third-party punishment in chapter one. We found that–unlike anger, contempt, and disgust–moral outrage is evoked by the interaction of culpable mental state and severe harms, and it alone mediates the relationship between this interaction and punishment decisions. We observed that sadness has the opposite effect of mediating a dampening of punishment in response to accidental harms.

The results from chapter three raised the question of whether the privileged role of moral outrage in punishment decision-making is specific to third-party punishment. Given the similarities between second- and third-party punishment, it is reasonable that they may draw from similar cognitive and affective states. On the other hand, a number of observations hint at the possibility that third-party punishment has a unique neurocognitive framework. By exposing participants to second- and third-party norm violations, while controlling for other factors, we identified that moral outrage does appear to be unique to third-party violations. Anger, on the other hand, appears unique to second-party violations.

Together, the results map out both the network of brain regions that support punishment decision-making, as well as the neurocognitive states that motivate the behavior. A ripe avenue for future research is an attempt to link the two. In other words, beyond the regions that support punishment decision-making, can we identify the mechanisms that drive punishment decision-making. The activation profile of the amygdala is tantalizing evidence of a possible connection, but chapter two was unable to provide further evidence of a connection between the two. The difficulty of linking cognitive and affective states with brain systems extends beyond this very research and is an area of active study, and debate, in the neurocognitive sciences more generally.

Beyond the results themselves, these studies hopefully provide a framework for future research at this exciting junction between neuroscience, psychology, and law. As just one example, the results from Chapter one provide a detailed examination, through behavioral modelling and neuroimaging analyses, of how punishment decisions are made under ideal conditions. As noted in the introduction, punishment decision-making is often subject to any number of biases. By starting with an existing framework we can better understand, for instance, how punishment decisions might go wrong. One intriguing example of how this work can be so leveraged is examining the effect of character information on punishment decision-making.

Every jurisdiction in the United States has adopted some version of Federal Rule of Evidence 404. This rule provides that evidence of a person's character or trait may not be admitted to prove behavior in accordance with the character or trait. Rule 404 is based on the presumption that character evidence is of little probative value while also being highly prejudicial. The rules committee that wrote rule 404 was concerned that such evidence "subtly permits the trier of fact to reward the good man or to punish the bad man because of their respective characters despite what the evidence in the case shows actually happened." However, while the rationale for rule 404 extends beyond the bounds of a trial, its protections do not. Every day, police officers, district attorneys, and even the public make decisions and intuitions regarding guilt and punishment. These decisions are made without any of the protections afforded by Rule 404. Three fascinating questions present themselves. First,
does character evidence actually influence third party punishment decisions? Second, to what extent can individuals restrain these influences upon instruction to do so? Third, what are the neural mechanisms that mediate the influence of such character evidence, as well as the mechanisms that support the potential disruption of this influence? By using the results from

these studies as a starting point, answering these questions becomes a much simpler proposition and one our lab has already started pursuing.

The most salient conclusion from the research presented above may be that neuroscience and psychology provide a rich toolbox to examine longstanding legal phenomena and assumptions. From the psychophysics of conscious perception, to massive online studies recruiting thousands of participants, and to machine learning on multivariate data acquired using fMRI, the tools used here are a demonstration of what's possible when leveraging established knowledge and techniques from multiple fields. I hope it provides a useful foundation for what's to come.

REFERENCES

Aldrich JH, Nelson FD (1984) Linear probability, logit, and probit models.

Ames DL, Fiske ST (2013) Intentional Harms Are Worse, Even When They're Not. Psychological Science 24:1755–1762.

Ames DL, Fiske ST (2015) Perceived intent motivates people to magnify observed harms. Proceedings of the National Academy of Sciences 112:3599–3605.

Andrews V, Lipp OV, Mallan KM, König S (2010) No evidence for subliminal affective priming with emotional facial expression primes. Motiv Emot 35:33–43.

Anon (2014) Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. 133:97–103.

Baayen RH (2008) Analyzing linguistic data: A practical introduction to statistics using R.

Baldus DC, Pulaski C, Woodworth G (1983) Comparative review of death sentences: An empirical study of the Georgia experience. The Journal of Criminal Law and Criminology 74:661.

Bates E, Wilson SM, Saygin AP, Dick F, Sereno MI, Knight RT, Dronkers NF (2003) Voxel-based lesion-symptom mapping. Nat Neurosci 6:448–450.

Batson CD, Kennedy CL, Nord LA (2007) Anger at unfairness: Is it moral outrage? European Journal of Social Psychology, 37(6), 1272–1285

Baumgartner T, Schiller B, Rieskamp J, Gianotti LRR, Knoch D (2014) Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. Social Cognitive and Affective Neuroscience 9:653–660.

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. Games and Economic Behavior 10:122–142.

Bowles S, Gintis H (2011) A cooperative species: Human reciprocity and its evolution.

Boyd R, Richerson PJ (2005) The Origin and Evolution of Cultures. Oxford University Press, USA.

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

Brame R, Bushway SD, Paternoster R, Turner MG (2014) Demographic Patterns of Cumulative Arrest Prevalence by Ages 18 and 23. Crime & Delinquency 60:471–486.

Bshary R, Grutter AS (2005) Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. Biology Letters 1:396–399.

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.

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.

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.

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. Journal of Neuroscience 29:1860–1873.

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science 6:3–5.

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

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

Cameron CD, Lindquist KA, Gray K (2015) A constructionist review of morality and emotions: no evidence for specific links between moral content and discrete emotions. Pers Soc Psychol Rev 19:371–394.

Carlsmith KM, Darley JM, Robinson PH (2002) Why do we punish?: Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology 83:284–299.

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

Castelhano MS, Muter P (2001) Optimizing the reading of electronic text using rapid serial visual presentation. Behaviour & Information Technology.

Chang C-C, Lin C-J (2001) LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Charness G, Cobo-Reyes R, Jiménez N (2008) An investment game with third-party intervention. PNAS 68:18–28.

Chavez AK, Bicchieri C (2013) Third-party sanctioning and compensation behavior: Findings from the ultimatum game. Journal of Economic Psychology 39:268–277.

Clark-Polner E, Johnson TD, Barrett LF (2016) Multivoxel Pattern Analysis Does Not Provide Evidence to Support the Existence of Basic Emotions. Cereb Cortex.

Corradi-Dell'Acqua C, Civai C, Rumiati RI, Fink GR (2013) Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Social Cognitive and Affective Neuroscience 8:424–431.

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. Social Cognitive and Affective Neuroscience 9:1175–1184.

Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research Gilbert S, ed. PLoS ONE 8:e57410–e57418.

Cubelli R, De Bastiani P (2011) 150 years after Leborgne: why is Paul Broca so important in the history of neuropsychology?

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

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

Damasio A (2005) Descartes' Error. Penguin.

Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences 108:6889–6892.

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. Cerebral Cortex 17:993–1006.

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

Dijksterhuis A, Smith PK (2002) Affective habituation: Subliminal exposure to extreme stimuli decreases their extremity. Emotion 2:203–214.

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in Cognitive Sciences 14:172–179.

Fehr E, Fischbacher U (2004) Social norms and human cooperation. Trends in Cognitive Sciences 8:185–190.

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

Fehr E, Rockenbach B (2004) Human altruism: economic, neural, and evolutionary perspectives. Current Opinion in Neurobiology 14:784–790.

Feigenson N, Park J (2006) Emotions and attributions of legal responsibility and blame: A research review. Law and Human Behavior 30:143.

Friston KJ, Preller KH, Mathys C, Cagnan H, Heinzle J, Razi A, Zeidman P (2017) Dynamic causal modelling revisited. NeuroImage.

Gallagher HL, Frith CD (2003) Functional imaging of "theory of mind." Trends in Cognitive Sciences 7:77–83.

Ginther MR, Shen FX, Bonnie RJ, Hoffman MB, Jones OD, Marois R, Simons KW (2014) The Language of Mens Rea. Vanderbilt Law Review.

Ginther MR, Bonnie RJ, Hoffman MB, Shen FX, Simons KW, Jones OD, Marois R (2016) Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. Journal of Neuroscience 36:9420–9434.

Glaze LE, Kaeble D (2014) Correctional Populations in the United States, 2013. Bureau of Justice Statistics NCJ 248479.

Gray K, Wegner DM (2008) The sting of intentional pain. Psychological Science 19:1260–1262.

Gutierrez R, Giner-Sorolla R, Vasiljevic M (2012) Just an anger synonym? Moral context influences predictors of disgust word use. Cognition & Emotion 26:53–64.

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.

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review.

Hamann SB, Ely TD, Grafton ST, Kilts CD (1999) Amygdala activity related to enhanced memory for pleasant and aversive stimuli. Nat Neurosci 2:289–293.

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

Hardecker S, Schmidt MF, Roden M, Tomasello M (2016) Young children's behavioral and emotional responses to different social norm violations. Journal of Experimental Child Psychology 150:364–379.

Hariri AR, Tessitore A, Mattay VS, Fera F (2002) The amygdala response to emotional stimuli: a comparison of faces and scenes. NeuroImage.

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

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. PNAS 103:10023–10028.

Heekeren HR, Wartenburger I, Schmidt H, Prehn K, Schwintowski H-P, Villringer A (2005)

Influence of bodily harm on neural correlates of semantic and moral decision-making. NeuroImage 24:887–897.

Hutcherson CA, Gross JJ (2011) The moral emotions: a social-functionalist account of anger, disgust, and contempt. Journal of Personality and Social Psychology 100:719–737.

Imai K, Keele L, Tingley D (2010) A general approach to causal mediation analysis. Psychological Methods 15:309–334.

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.

Janowski V, Camerer C, Rangel A (2013) Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Social Cognitive and Affective Neuroscience 8:201–208.

Kahneman D (1968) Method, findings, and theory in studies of visual masking. Psychological Bulletin 70:404.

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

Kim H, Somerville LH, Johnstone T, Alexander AL, Whalen PJ (2003) Inverse amygdala and medial prefrontal cortex responses to surprised faces. Neuroreport 14:2317–2322.

Kim H, Somerville LH, Johnstone T, Polis S, Alexander AL, Shin LM, Whalen PJ (2004) Contextual modulation of amygdala responsivity to surprised faces. Journal of Cognitive Neuroscience 16:1730–1745.

Kim J, Zhu W, Chang L, Bentler PM, Ernst T (2007) Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. Human Brain Mapping 28:85–93.

Kim MJ, Loucks RA, Neta M, Davis FC, Oler JA, Mazzulla EC, Whalen PJ (2010) Behind the mask: the influence of mask-type on amygdala response to fearful faces. Social Cognitive and Affective Neuroscience 5:363–368.

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.

Koenigs M, Tranel D (2007) Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game. Journal of Neuroscience 27:951–956.

Kogut T (2011) The role of perspective taking and emotions in punishing identified and unidentified wrongdoers. Cognition & Emotion 25:1491–1499.

Kyckelhahn, T. (2015) Justice Expenditure and Employment Extracts, 2012-Preliminary. U.S.

Bureau of Justice Statistics. NCJ 248628. www.bjs.gov

LaFave WR (1986) Criminal law. St. Paul, MN: West Pub. Co.

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.

Landmann H, Hess U (2016) What elicits third-party anger? The effects of moral violation and others' outcome on anger and compassion. Cognition & Emotion:1–15.

Laurent SM, Clark BAM, Walker S, Wiseman KD (2014) Punishing hypocrisy: the roles of hypocrisy and moral emotions in deciding culpability and punishment of criminal and civil moral transgressors. Cognition & Emotion 28:59–83.

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

Lergetporer P, Angerer S, Glätzle-Rützler D, Sutter M (2014) Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. PNAS 111:6916–6921.

Li W, Zinbarg RE, Boehm SG, Paller KA (2008) Neural and behavioral evidence for affective priming from unconsciously perceived emotional facial expressions and the influence of trait anxiety. Journal of Cognitive Neuroscience 20:95–107.

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. PNAS 110:1929–1934.

Loewenstein G, Lerner JS (2002) The role of affect in decision making. In: Handbook of Affective Sciences. Oxford University Press, USA.

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

Mason W, Suri S (2011) Conducting behavioral research on Amazon’s Mechanical Turk. Behav Res 44:1–23.

Mathew S, Boyd R (2011) Punishment sustains large-scale cooperation in prestate warfare. PNAS 108:11375–11380.

Maziarz M (2015) A review of the Granger-causality fallacy. The Journal of Philosophical Economics: Reflections on Economic and social issues VIII.:86–105.

McAuliffe K, Jordan JJ, Warneken F (2015) Costly third-party punishment in young children.

Cognition 134:1–10.

Miller R, Cushman F (2013) Aversive for Me, Wrong for You: First-person Behavioral Aversions Underlie the Moral Condemnation of Harm. Social and Personality Psychology Compass 7:707–718.

Mulder RA, Langmore NE (1993) Dominant males punish helpers for temporary defection in superb fairy-wrens. Animal Behaviour 45:830–833.

Nabi RL (2002) The theoretical versus the lay meaning of disgust: Implications for emotion research. Cognition & Emotion.

Nichols T, Brett M, Andersson J, Wager T, Poline J-B (2005) Valid conjunction inference with the minimum statistic. NeuroImage 25:653–660.

Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437:1291–1298.

Paolacci G, Chandler J, Ipeirotis P (2010) Running experiments on Amazon Mechanical Turk. Judgment and Decision Making 5:411–419.

Paternoster R, Brame R, Mazerolle P, Piquero A (1998) Using the correct statistical test for the equality of regression coefficients. Criminology.

Pearl J (2012) The causal mediation formula--a guide to the assessment of pathways and mechanisms. Prev Sci 13:426–436.

Pedersen EJ, Kurzban R, McCullough ME (2013) Do humans really punish altruistically? A closer look. Proceedings of the Royal Society B: Biological Sciences 280:20122723–20122723.

Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat Vis.

Pessoa L, Japee S, Sturman D, Ungerleider LG (2006) Target Visibility and Visual Awareness Modulate Amygdala Responses to Fearful Faces. Cerebral Cortex 16:366–375.

Phelps EA (2006) Emotion and Cognition: Insights from Studies of the Human Amygdala. Annu Rev Psychol 57:27–53.

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

Phelps EA, Lempert KM, Sokol-Hessner P (2014) Emotion and Decision Making: Multiple Modulatory Neural Circuits. Annu Rev Neurosci 37:263–287.

Piazza J, Russell PS, Sousa P (2013) Moral emotions and the envisaging of mitigating circumstances for wrongdoing. Cognition & Emotion 27:707–722.

Pillutla MM, Murnighan JK (1996) Unfairness, anger, and spite: Emotional rejections of ultimatum offers. Organizational Behavior and Human Decision Processes 68:208–224.

Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res 40:879–891.

Rand DG (2012) The promise of Mechanical Turk How online labor markets can help theorists run behavioral experiments. Journal of Theoretical Biology 299:172–179.

Riedl K, Jensen K, Call J (2012) No third-party punishment in chimpanzees. PNAS 37: 14824–14829.

Roebroeck A, Formisano E, Goebel R (2005) Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage 25:230–242.

Rosnow RL, Rosenthal R (1996) Contrasts and Interactions Redux: Five Easy Pieces. Psychological Science 7:253–257.

Royzman E, Atanasov P, Landy JF, Parks A, Gepty A (2014) CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. Emotion 5:892-907.

Rozin P, Lowery L, Imada S, Haidt J (1999) The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). Journal of Personality and Social Psychology 76:574–586.

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. European Journal of Neuroscience 28:1569–1588.

Russell PS, Piazza J (2013) CAD revisited effects of the word moral on the moral relevance of disgust (and other emotions). Social Psychological & Personality Science 4:62-68.

Saarimäki H, Gotsopoulos A, Jääskeläinen IP, Lampinen J, Vuilleumier P, Hari R, Sams M, Nummenmaa L (2016) Discrete Neural Signatures of Basic Emotions. Cereb Cortex 26:2563–2573.

Salerno JM, Peter-Hagene LC (2013) The interactive effect of anger and disgust on moral outrage and judgments. Psychological Science 24:2069–2078.

Sanfey AG (2003) The Neural Basis of Economic Decision-Making in the Ultimatum Game. Science 300:1755–1758.

Shen FX, Hoffman MB, Jones OD, Greene JD, Marois R (2011) Sorting guilty minds. New York University Law Review 86:1306–1360.

Shenhav A, Greene JD (2014) Integrative Moral Judgment: Dissociating the Roles of the

Amygdala and Ventromedial Prefrontal Cortex. J Neurosci 34:4741–4749.

Shweder RA, Much NC, Mahapatra M and Park L. (2003) The "Big Three" of Morality (Autonomy, Community, Divinity) and the Big Three Explanations of Suffering. In: Shweder RA (ed.) *Why Do Men Barbecue? Recipes for Cultural Psychology*. Cambridge, MA: Harvard University Press.

Simons KW (2003) Should the Model Penal Code's Mens Rea Provisions Be Amended. Ohio State Journal of Criminal Law 1:179– 205.

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

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.

Skorinko JL, Laurent S, Bountress K (2014) Effects of perspective taking on courtroom decisions. Journal of Applied Social Psychology 44:303-318.

Small DA, Loewenstein G (2005) The devil you know: the effects of identifiability on punishment. J Behav Decis Making 18:311–318.

Smith SM (2012) The future of FMRI connectivity. NeuroImage 62:1257–1266.

Sporns O, Honey CJ, Kötter R (2007) Identification and classification of hubs in brain networks. PLoS ONE.

Sprouse J (2010) A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. Behav Res 43:155–167.

Steen J, Loeys T, Moerkerke B, Vansteelandt S (1AD) Medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models.

Steffensmeier D, Ulmer J, Kramer J (1998) The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. Criminology 36:763–798.

Tamber-Rosenau BJ, Dux PE, Tombu MN, Asplund CL, Marois R (2013) Amodal Processing in Human Prefrontal Cortex. J Neurosci 33:11573–11587.

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. Social Cognitive and Affective Neuroscience 7:282–288.

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

Tonry M (1995) Malign neglect: Race, crime, and punishment in America. Oxford University Press.

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

Valeri L, Vanderweele TJ (2013) Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods 18:137–150.

Wallis JD (2007) Orbitofrontal Cortex and Its Contribution to Decision-Making. Annu Rev Neurosci 30:31–56.

Wenseleers T, Ratnieks FLW (2006) Enforced altruism in insect societies. Nature 444:50–50.

Whalen PJ, Rauch SL, Etcoff NL, McInerney SC, Lee MB, Jenike MA (1998) Masked Presentations of Emotional Facial Expressions Modulate Amygdala Activity without Explicit Knowledge. J Neurosci 18:411–418.

Winterich KP, Han S, Lerner JS (2010) Now that I'm sad, it's hard to be mad: the role of cognitive appraisals in emotional blunting. Personality and Social Psychology Bulletin 36:1467–1483.

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.

Zald DH (2003) The human amygdala and the emotional evaluation of sensory stimuli. Brain Research Reviews 41:88–123.