2009

Brain Imaging for Legal Thinkers: A Guide for the Perplexed

Owen D. Jones
Vanderbilt University Law School

Joshua W. Buckholtz
Vanderbilt University

Jeffrey D. Schall
Vanderbilt University

Rene Marois
Vanderbilt University

Stanford Technology Law Review

Brain Imaging for Legal Thinkers: A Guide for the Perplexed

OWEN D. JONES, JOSHUA W. BUCKHOLTZ, JEFFREY D. SCHALL, RENE MAROIS[1]

CITE AS: 2009 STAN. TECH. L. REV. 5
http://stlr.stanford.edu/pdf/jones-brain-imaging.pdf

INTRODUCTION

¶1 It has become increasingly common for brain images to be proffered as evidence in civil and criminal litigation.[2] This Article offers some general guidelines to legal thinkers about how to understand brain imaging studies—or at least avoid misunderstanding them. And it annotates a published brain imaging study by several of the present authors (and others) in order to illustrate and explain, with step-by-step commentary.[3]

¶2 Brain images are offered in legal proceedings for a variety of purposes, as Professors Carter Snead and Gary Marchant have usefully surveyed.[4] On the civil side, neuroimaging has been offered in constitutional, personal injury, disability benefit, and contract cases, among others. For example, in *Entertainment Software Ass'n. v. Blagojevich,*[5] the court considered whether a brain imaging study could be used to show that exposure to violent video games increases aggressive thinking and behavior in

---

[1] Owen D. Jones is Professor of Law and Professor of Biological Sciences at Vanderbilt University. Joshua W. Buckholtz is a neuroscience graduate student at Vanderbilt University. Jeffrey D. Schall is E. Bronson Ingram Professor of Neuroscience at Vanderbilt University. Rene Marois is Associate Professor of Psychology at Vanderbilt University. Jones, Schall, and Marois are members of the MacArthur Foundation Law and Neuroscience Project, of which Jones also serves as Co-Director. The first two authors contributed equally to this Article. Correspondence to: owen.jones@vanderbilt.edu or rene.marois@vanderbilt.edu. This Article was prepared for the Stanford Technology Law Review 2009 Symposium on *Neuroscience and the Courts: The Implications of Advances in Neurotechnology.*

We received helpful comments from Gary Marchant and Teneille Brown, as well as from participants at conferences of the MacArthur Foundation Law and Neuroscience Project, the 2008 and 2009 Conferences on Empirical Legal Studies, and the Arizona State University College of Law conference "Law and Ethics of Brain Scanning: The Next Big Thing Coming Soon to a Courthouse Near You?" Bailey Spaulding and Francis Shen provided valuable research assistance.

Preparation of this Article was supported by the John D. and Catherine T. MacArthur Foundation, The Regents of the University of California, and Vanderbilt University.

[2] For an overview of issues, see Jeffrey Rosen, *The Brain on the Stand*, N.Y. TIMES MAG., Mar. 11, 2007, at 49; Stacey A. Tovino, *Functional Neuroimaging and the Law: Trends and Directions for Future Scholarship*, 7 AM. J. BIOETHICS 44 (2007). A sampling of the rapidly-growing scholarship at the law/neuroscience intersection appears *infra* note 32.

[3] The full complement of authors is: Joshua W. Buckholtz, Christopher L. Asplund, Paul E. Dux, David Zald, John C. Gore, Owen D. Jones, and Rene Marois. The article was published as *The Neural Correlates of Third-Party Punishment*, 60 NEURON 930 (2008).

[4] A very useful survey, on which we draw in part in the paragraphs that follow, has been prepared by Professor Carter Snead. *See* CARTER SNEAD, NEUROIMAGING AND THE COURTS: STANDARD AND ILLUSTRATIVE CASE INDEX, (2006), http://www.ncsconline.org/d_research/stl/June06/Snead.doc . Our research also benefitted from Gary Marchant, Brain Scanning and the Courts: Criminal Cases, Presentation to the Research Network on Legal Decision Making, MacArthur Foundation Law and Neuroscience Project (Oct. 11, 2008).

[5] 404 F. Supp. 2d 1051 (N.D. Ill. 2005).

adolescents. In *Fini v. General Motors Corp*,[6] brain scans were proffered to help determine the extent of head injuries from a car accident. In *Boyd v. Bert Bell/Pete Rozelle NFL Players Retirement Plan*,[7] a former professional football player proffered brain scans in an effort to prove entitlement to neuro-degenerative disability benefits. And in *Van Middlesworth v. Century Bank & Trust Co.*[8], involving a dispute over the sale of land, the defendant introduced brain images to prove mental incompetency, resulting in a voidable contract.

¶3 In criminal cases, brain images are sometimes invoked to support an argument that a defendant is incompetent to stand trial. In *United States v. Kasim*, for example, Kasim was found to be demented, and incompetent to stand trial for Medicaid fraud, on the basis of medical testimony that included brain images.[9] Brain images are also increasingly proffered by the defense at the guilt-determination phase, in an effort to negate the mens rea element of a crime, and to thereby avoid conviction. For example, in *People v. Weinstein*, [10] a defendant accused of strangling his wife and throwing her from a twelfth floor window sought to introduce images of a brain defect, in support of an argument that he was not responsible for his act. And in *People v. Goldstein,*[11] a defendant sought to introduce a brain image of an abnormality, in an effort to prove an insanity defense, after he pushed a woman in front of a subway train, killing her.

¶4 Brain images have also been proffered at the sentencing phase of criminal cases, in furtherance of mitigation. For example, in *Oregon v. Kinkel*,[12] a boy convicted of killing and injuring fellow students in a high school cafeteria sought to introduce brain images of abnormalities, in an effort to secure a more lenient sentence. Brain images have been offered—in *Coe v. State*,[13] for example—to argue that a convicted murderer is not competent to be executed. And accessibility to brain imaging technology has even been litigated—in *Ferrell v. State*[14] and *People v. Morgan*[15] for instance—in the context of a claim that a defense counsel's failure to procure a brain image for the defendant amounted to ineffective assistance of counsel.

¶5 For better or worse, the full complement of cases at the intersection of neuroscience and law is now too large for comprehensive overview—in part because many of the cases do not result in reported decisions.[16] While there is no denying that brain imaging is a powerful tool, whether used for medical or legal purposes, it is also clear that, like any tool, brain imaging can be used for good or for ill, skillfully or sloppily, and in ways useful or irrelevant.

¶6 We are concerned that brain imaging can be misused by lawyers (intentionally or unintentionally) and misunderstood by judges and jurors. Consequently, our aim in this Article is to provide information about the operation and interpretation of brain imaging techniques, in hopes that it will increase the extent to which imaging is properly interpreted, and conversely decrease the extent to which it is misunderstood or misused. We provide this information across two Parts and one Appendix.

¶7 Part I of the Article provides some very brief background on modern brain imaging, with particular emphasis on one wide-spread and powerful technique, known as functional magnetic

---

[6] No. 227592, 2003 Mich. App. LEXIS 884 (Mich. Ct. App. Apr. 8 2003).

[7] 410 F.3d 1173 (9th Cir. 2005).

[8] No. 215512, 2000 Mich. App. LEXIS 2369 (Mich. Ct. App. May 5, 2000).

[9] United States v. Kasim, No. 2:07 CR 56, 2008 U.S. Dist. LEXIS 89137 (N.D. Ind. Nov. 3, 2008). *See also* McMurtey v. Ryan, 539 F.3d 1112 (9th Cir. 2008); United States v. Gigante, 982 F. Supp. 140 (S.D.N.Y. 1997).

[10] 591 N.Y.S.2d 715 (N.Y. Sup. Ct. 1992).

[11] 786 N.Y.S.2d 428 (N.Y. Sup. Ct. 2004), *overruled on other grounds*, 6 N.Y.3d 119, 843 N.E.2d 727, 2005 N.Y. LEXIS 3389 (2005).

[12] 56 P.3d 463 (Or. Ct. App. 2002).

[13] 17 S.W.3d 193 (Tenn. 2000).

[14] 918 So.2d 163 (Fla. 2005).

[15] 719 N.E.2d 681 (Ill. 1999).

[16] One of the many efforts under way, within the MacArthur Foundation Law and Neuroscience Project, is a study by Hank Greely and Teneille Brown to find all actual and attempted uses of neuroimaging in criminal cases in California after January 1, 2006, regardless of whether such uses are mentioned in published opinions.

resonance imaging (fMRI). The physics of fMRI, and the statistics accompanying the analyses that generate brain images, are complicated. We will make no effort to provide a comprehensive or detailed exploration of the subject. There are many existing textbooks that cover this material to great depths, often far greater than legal thinkers will need to master, for the specific contexts in which brain images are (potentially) legally relevant.[17]

¶8 Instead, we will aim here to focus on what a lawyer needs to know, in order to have a basic understanding of what works how and why. Our goal is to present this in an accessible way, recognizing (as we trust our readers to allow us) that simplifying discussions are illustrative of general principles, but obviously ignore the richer detail that enables deeper appreciation of important caveats and subtleties.

¶9 Part II of this Article then turns to provide, in brief and accessible overview, a variety of key concepts to understand about the legal, biological, and brain imaging contexts at this particular law/neuroscience intersection, as well as a variety of guidelines we (and in some cases others) recommend to help avoid the various factual errors, logical traps, and analytic mis-steps that can all too quickly lead away from sound and sensible understandings of what brain images can mean—and equally what they cannot. Make no mistake: we are not the only researchers concerned about potential misunderstandings of brain images.[18] A great many cautions have been swirling about in the literature, often offering multiple versions of key and basic points about the limitations of the technologies, and we hope here to distill some of those, add others, and explain the set in a way that we hope provides a concise and useful introduction to legal thinkers approaching this interdisciplinary nexus for the first time.

¶10 The Appendix to this Article then provides a concrete illustration of how to read an fMRI study. We will not over-claim. Some of the details of fMRI defy short descriptions, involve technical details unlikely to be relevant in legal contexts, or both. On the other hand, much of the technical jargon, and many of the basic concepts one will encounter in an fMRI study, are clear with just a little explanation, oriented toward the audience we anticipate. We attempt to provide this in an accessible, informative way—assuming no particular scientific sophistication of the reader.

¶11 Specifically, the core of the Appendix is a 2008 fMRI study (co-authored by three of us and others) that used fMRI techniques to investigate how brains are activated during punishment decisions. Though we do not anticipate that the substantive findings will necessarily find immediate utility in litigation, we believe that legal thinkers reading an fMRI study will learn most from a study

---

[17] *See, e.g.*, SCOTT A. HUETTEL ET AL., FUNCTIONAL MAGNETIC RESONANCE IMAGING (2d ed. 2009); ALFRED L. HOROWITZ, MRI PHYSICS FOR RADIOLOGISTS: A VISUAL APPROACH (3d ed. 1995); FUNCTIONAL MRI: AN INTRODUCTION TO METHODS (Peter Jezzard et al., 2001). Useful introductions to broader cognitive neuroscience, of which brain-imaging is but a part, appear in: MICHAEL S. GAZZANIGA ET AL., COGNITIVE NEUROSCIENCE: THE BIOLOGY OF THE MIND (3d ed. 2008); JAMIE WARD, THE STUDENT'S GUIDE TO COGNITIVE NEUROSCIENCE (2006); ESSENTIALS OF NEURAL SCIENCE AND BEHAVIOR (Eric R. Kandel et al. eds., 1995); MARIE T. BANICH, COGNITIVE NEUROSCIENCE AND NEUROPSYCHOLOGY (2d ed. 2004); MARK F. BEAR ET AL., NEUROSCIENCE: EXPLORING THE BRAIN (3d ed. 2006); NEUROSCIENCE (Dale Purves et al. eds., 4th ed. 2007).

[18] The limits of brain imaging techniques are widely known to brain imaging researchers, and many brain imaging researchers are broadly concerned about misunderstandings among laypeople. A non-exhaustive list of important cautionary and explanatory articles, which have influenced some of our approaches below, include: John T. Cacioppo et al., *Just Because You're Imaging the Brain Doesn't Mean You Can Stop Using Your Head: A Primer and Set of First Principles*, 85 J. PERSONALITY AND SOC. PSYCHOL. 650 (2003); Dean Mobbs et al., *Law, Responsibility, and the Brain*, 5 PLOS BIOLOGY 693 (2007); Eric Racine et al., *fMRI in the Public Eye*, 6 NATURE REVS. NEUROSCIENCE 159 (2005); J.D. Trout, *Seduction Without Cause: Uncovering Explanatory Neurophilia*, 12 TRENDS COGNITIVE SCI. 281 (2008); Society of Nuclear Medicine Brain Imaging Council, *Ethical Clinical Practice of Functional Brain Imaging*, 37 J. NUCLEAR MED. 1256 (1996); Michael S. Gazzaniga, *The Law and Neuroscience*, 60 NEURON 412 (2008); Joseph H. Baskin et al., *Is A Picture Worth A Thousand Words? Neuroimaging in the Courtroom*, 33 AM. J.L. & MED. 239 (2007); Russell A. Poldrack et al., *Guidelines for Reporting an fMRI Study*, 40 NEUROIMAGE 409 (2008); Nikos K. Logothetis, *What We Can Do and What We Cannot Do With fMRI*, 453 NATURE 869, (2008); WILLIAM R. UTTAL, NEUROSCIENCE IN THE COURTROOM: WHAT EVERY LAWYER SHOULD KNOW ABOUT THE MIND AND THE BRAIN (2008). One of the works most critical of how brain imaging results can be interpreted is WILLIAM R. UTTAL, THE NEW PHRENOLOGY: THE LIMITS OF LOCALIZING COGNITIVE PROCESSES IN THE BRAIN (2003).

that inherently addressed matters relevant to law—in this case, the decision whether or not to punish someone for criminal behavior and, if so, how much.

¶12 To facilitate that learning in this concrete application, the *Stanford Technology Law Review* has generously afforded us the unique opportunity to annotate the Article in the margin with explanations of various terms and contexts, as they appear throughout the study.

I. BRAIN-IMAGING: A VERY BRIEF OVERVIEW

¶13 There are many kinds of brain images. All readers are likely familiar with the way x-rays, and the closely aligned technique known as computed tomography (CT) scanning, can show various structural anomalies in the body, including in the brain. In these techniques, radiation aimed at and passing through the body forms images on photographic film. The varying density of different tissues in the body results in varying levels of radiation reaching the film—creating, in turn, an image of internal structures. (For example, bone tissue appears as white, while soft tissue appears gray.) CT scanning varies from conventional x-rays by virtue of collecting images from multiple angles rotating around the body, which images are then combined by computers into cross-sectional representations. These techniques (like magnetic resonance imaging, which will be discussed in a moment) are used for information about how various parts of the body are structured. They can show whether structures are intact, and can reveal damage, atrophy, intrusions, and developmental anomalies. They do not, however, collect or provide information about how those body parts are actually functioning.

¶14 *PET scanning*, which refers to positron emission tomography, is one of the techniques that enable researchers to learn about how the brain *functions*, as it is actually doing so. With PET, a researcher injects a subject with radioactive tracers that move through the bloodstream and accumulate in different locations and concentrations in the brain, over time, as different parts of the brain increase and decrease activity (such as glucose metabolism) that is associated with brain function. (A similar technique, known as *SPECT*, uses single photon emission computed tomography.)

¶15 *EEG* and *MEG,* short for electroencephalography and magnetoencephalography respectively, records electromagnetic fluctuations in various parts of the brain, as the brain is functioning, using non-invasive sensors applied to the scalp.[19] In research laboratories, the EEG signals can be analyzed in relation to stimuli or responses to obtain *event-related potentials* (ERP) which were used before brain imaging was developed to make inferences about the brain processes underlying perceptual, cognitive and motor processes.[20]

¶16 *fMRI* (functional magnetic resonance imaging[21]) uses the technology of regular magnetic resonance imaging adapted to detect changes in hemodynamic (literally "blood movement") properties of the brain occurring when the subject is engaged in very specific mental tasks. In a nutshell (and with a reminder that we are over-simplifying for heuristic purposes) here's how it works.

¶17 At its most basic, fMRI can be understood as a tool for learning which regions of the brain are working, how much, and for how long, during particular tasks. In much the same way that the body delivers more oxygen to muscles that are working harder, the body delivers more oxygen to brain regions that work harder. The fMRI technique measures blood oxygenation levels—within small cubic volumes of brain tissue known as "voxels"—as those levels change across time with the

[19] This signal is used in conjunction with measures like heart-rate and skin electrical conductance to constitute the polygraph procedure that is used commonly in a context of detecting deception. Although used commonly by the U.S. government and police departments, the fundamental limitations of these procedures have been thoroughly described. *See, e.g.*, COMM. TO REVIEW THE SCIENTIFIC EVIDENCE ON THE POLYGRAPH, NAT'L RESEARCH COUNCIL, THE POLYGRAPH AND LIE DETECTION (2003)

[20] STEVEN J. LUCK, AN INTRODUCTION TO THE EVENT-RELATED POTENTIAL TECHNIQUE (2005). Some have attempted to use ERP signals in legal settings, but the limitations of this approach are well-known and can serve as lessons for the interpretation of brain imaging information.

[21] The leading "f" remains lower-case, by convention.

varying metabolic demands of active neurons.[22] Changes in demand for oxygen are widely considered to be reliable proxies for inferring the fluctuating activity of the underlying neural tissue.[23]

¶18 The physical principles underlying fMRI are quite complex. But in general terms the technology works as follows: An fMRI machine creates and manipulates a primary magnetic field,[24] as well as several smaller magnetic fields (one in each three-dimensional plane) that can be quickly varied in orientation and uniformity. Recall (from basic physics) that protons within the nuclei of atoms spin on an axis and carry a positive charge. As they spin, these electric charges form what can be thought of as tiny magnets. When a person is inserted (typically horizontally) into the open bore of an fMRI machine, the previously random axes of spin, for many protons, align, like iron filings along a magnet. That is, the axes begin to point in the same direction. Researchers then administer to the subject's head brief radio frequency pulses (which usually originate from a device looking rather like a small bird-cage that surrounds the subject's head). Those pulses deflect the protons' axes of spin temporarily. When the pulses stop, the axes gradually return to their original orientation, releasing energy during that "relaxation" process. The machine can detect characteristics of the released energy because it depends on a proton's "local" magnetic environment, and this environment is affected by the relative concentrations of oxygenated and deoxygenated blood in local brain tissue. Crucially, as these concentrations are affected by regional changes in brain activity, they provide indirect markers of neural activity that form the basis of the fMRI signal. The machine enables localization of these signals in space—i.e. "spatial resolution"—by collecting them from many different "slices" of the brain. And the technique enables localization of these signals in time—i.e., "temporal resolution"—by recording the signals many times over a period of several seconds for each mental event. A "stack" of slices comprising the whole brain is acquired every couple of seconds or so, enabling the rapid collection of many of these three-dimensional "volumes" of brain activity over the period of an experimental paradigm.

II. KEY CONCEPTS AND GUIDELINES

¶19 This Part is divided into four sections. These address the legal context, the biological context, the intersection of law and biology, and finally, with that preparatory background, the brain imaging context. We proceed in this way because one cannot gain a clear understanding of brain imaging, and its intersection with the legal system, without first considering the underlying legal and biological contexts, and their background interactions.

A. The Legal Context

¶20 With terrific, new, whiz-bang technology—which can reveal inner structures and workings of the brain—it is all too tempting to jump past the more mundane legal issues, and to race to apply new techniques to solve new problems in new ways.

¶21 But hold the horses. Although our principal purpose here is to discuss how to read (and not read) brain imaging evidence, we would be remiss not to first anchor the discussion in the legal contexts in which those images might, arguably, be admissible. The territory here is broad, and could occupy us for some time. But to be brief, there are a variety of questions to keep in mind, at the outset, in order to understand the specific legal context in which brain imaging might be considered in the courtroom.

¶22 The threshold consideration, of course, is: *Are the proffered brain images relevant?* Because behavior comes from the brain, and the legal system often cares not only about how someone acted but also

---

[22] *See generally* HUETTEL ET AL., *supra* note 17.

[23] There are varying opinions in the neuroscience community about how conclusive an understanding there is of the fMRI signal's relationship to the activity of neurons, and about how much fMRI can reveal—beyond where brain activation occurred—about behavior and mental states. *See, e.g.,* Logothetis, *supra* note 18; Poldrack, *supra* note 18.

[24] Magnetic fields are described in Tesla units. A 3-Tesla machine (which uses super-cooled electrical coils) generates a magnetic field roughly 60,000 times the magnetic field of the Earth.

why, it is tempting to assume that brain images of people important to the litigation will provide legally relevant information, of one sort or another. But this is, in fact, not a decision to reach lightly.

¶23 *What specific legal questions do the images purportedly address?* Contexts vary considerably, even within the civil and criminal halves of the docket (each of which bears differing underlying standards of proof). Within civil cases, for example, there are a wide variety of different legal purposes into which brain images might conceivably plug. Are brain images proffered to help establish liability, such as in the case of a medical malpractice action? To demonstrate a pre-existing condition, such as in the case of a dispute over insurance coverage? To help estimate damages, such as in the case of a car accident? And within criminal cases, are brain images proffered during the liability phase, in an effort to defeat the prosecution's claim that the defendant had (and was therefore capable of having) the mental state requisite for conviction? Are they instead proffered during the sentencing phase, in an effort to mitigate penalty? Are they proffered as evidence of lying or truthfulness?

¶24 It is important to remember that the admissibility of brain images is not simply a matter of whether they are scientifically sound. The potential relevance and hence admissibility of brain images will vary, according to *the specific legal issue at hand* within civil and criminal contexts. Put another way, the admissibility of brain images depends largely on their perceived potential relevance (if any) to the issue to be determined, *independent of* (and often before) considering the quality and interpretation of the specific images themselves.

¶25 *What, specifically, do the images allegedly demonstrate, and how well does that connect to the legal issues at hand?* Some of the many variables that may come into play here include: Are these structural or functional images? When were they taken? (For example, before or after events in question?) How recently? Under what circumstances were they procured? (For example, what specific mental tasks was the subject executing during functional imaging?) What is being compared to what? (For example: Are these before and after images of the same brain?; Are these comparisons between a party's brain and a group-averaged composite, for contrast?)

¶26 *What are the applicable standards for the admissibility of scientific evidence?* As is well known, the federal and state systems can have (and often do have) different standards for the admission of scientific evidence. And the state standards vary among the states. It is therefore necessary to note that the backdrop of all that follows below is the specific legal regime under which images are to be evaluated for potential relevance, within the specific context of the specific matters in dispute. Although it is not our purpose here to explore the applicability of scientific evidence law to brain images, we would be remiss not to flag the centrality of evidentiary rules and contexts to all that follows. Interested readers will find comprehensive discussion of scientific evidence generally in the treatise MODERN SCIENTIFIC EVIDENCE.[25]

B. The Biological Context

¶27 Understanding the potential relevance of brain images to law also requires a few words of general background about the relationship between biology and behavior generally. Key things to keep in mind (generally speaking) include[26]:

- All behavior results from the interaction of genes, environments (including social contexts), developmental history, and the evolutionary processes that built the brain to function in the ways it does.
- Behavior originates in the physical and chemical activities of the brain.[27]

---

[25] MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (David L. Faigman et al., eds., 2006). Chapter One provides an excellent overview of the "general acceptance" and validity tests. It examines the cases that established those tests and discusses subsequent cases that applied and further developed those tests.

[26] Interested readers can find further information about these background principles in a variety of sources (as well as in the citations that they, in turn, provide). *See, e.g.*, Jeffrey D. Schall, *On Building A Bridge Between Brain And Behavior*, 55 ANN. REV. PSYCH. 23 (2004).

[27] Yes, the alert reader will point out that some behavior, such as reflexes, leaps right out of the spinal cord. In the text, we are speaking in generalities.

- All behavior is thus "biological."
- Understanding behavior as biological in nature does not mean that behavior is "biologically determined" in a reductionist or reliably predictive way.
- The brain is an evolved information-processing organ that, generally speaking, and through differing processes, associates various environmental inputs with various behavioral outputs.
- Those environmental influences are (generally speaking) unique for each individual.
- Each person's brain, though highly flexible, is both anatomically and functionally specialized. (That is, brains do not consist of undifferentiated all-purpose tissue.)
- Humans share, across the species, a common brain plan of anatomical and functional specialization,
- Each brain is slightly different in size, shape, and other anatomical features.
- One area of the brain can affect multiple behaviors.
- A given behavior arises from multiple areas of the brain.
- Different individuals can use different parts of the brain, in different ways, on the same cognitive tasks.
- Behavior is a complex phenomenon, neither attributable to single causes, nor easily parsed among multiple causes.
- Cognitive phenomena rarely originate from a single region in the brain.

C. The Intersections of Biology and Law

¶28 The potential relevance of brain imaging to law must be evaluated against the broader background of the intersections of law and human biology (both structural and behavioral) generally.[28]

- Like the rest of behavior, both criminal and law-abiding behavior originates in the brain.
- There is no brain structure, or set of brain structures, that is specifically "for" criminal or law-abiding behavior (since those categorizations of behavior are socially determined).
- To say that brain features influence behavior relevant to crime does not mean that brain features can necessarily explain why certain individuals behaved criminally.
- No explanation of any kind, brain-based or otherwise, has an automatic bearing on justification or exculpation or mitigation in law.
- Legal responsibility for behavior is a legal conclusion, not a scientific finding.
- Establishing a "biological basis" for behavior carries no automatic, normative relevance to anything (legal or otherwise).
- Norms, though influenced by biology, can never be justified by biology alone.

D. The Brain Imaging Context (using fMRI)

¶29 With that brief but foundational background, drawing attention to the legal and biological contexts, and the interaction of them, we can now turn to discuss key concepts about brain imaging that legal thinkers should know[29]:

---

[28] *See, e.g.,* Owen D. Jones & Timothy H. Goldsmith, *Law and Behavioral Biology*, 105 COLUM. L. REV. 405 (2005). *See also* LAW & THE BRAIN (Semir Zeki & Oliver Goodenough, eds., 2006); LAW, MIND, AND BRAIN (Michael Freeman & Oliver Goodenough, eds., 2009); THE IMPACT OF BEHAVIORAL SCIENCES ON CRIMINAL LAW (Nita Farahany ed., 2009); Owen D. Jones, *Behavioral Genetics and Crime, In Context*, 69 L. & CONTEMP. PROBLEMS 81 (2006); bibliographic sources compiled on the website of *The Society for Evolutionary Analysis in Law* (www.sealsite.org).

[29] For more details, see sources cited *supra* note 17.

1. Anatomical imaging and functional imaging are importantly different.

¶30 Two anatomical images, taken one minute apart, will ordinarily look identical. Yet two functional images, from data collected one minute apart, could look completely different. One reason this is so is simply that, in the latter case, brain activity changes rapidly. Another reason is because fMRI brain images are built statistically, not recorded photographically. In the typical fMRI case, hundreds of recordings are made of each voxel in the brain, at slightly different times (e.g., every two seconds). Each recording of each voxel within a given trial is analogous to a single frame in a movie. Learning what happens within each voxel, over time, is akin to watching motion seem to emerge from the observation of successive snapshots that comprise a moving picture. But that metaphor only captures part of the fMRI technique, because there are subsequently many repeat recordings of that voxel, under similar conditions, on many consecutive trials—the results of which are typically then averaged across trials. Complicating matters further is that there are about one hundred thousand voxels within the brain, and what typically matters is how neural activity within those voxels is varying over time, in relation to some task the subject(s) undertake while being scanned. Furthermore, within each voxel are millions of neurons of different types, interacting in ways that could be mechanistically different but indistinguishable from the measure of fMRI. In the end, fMRI brain images lay the result of any one of many possible statistical tests overtop of an anatomical image of a selected slice of the brain. That is, an fMRI image is a composite of an anatomical image, of the researcher's choosing, and a statistical representation of the brain activity in that image, also of the researcher's choosing.

2. Functional brain imaging is not mind reading.

¶31 There is more to a thought than blood flow and oxygen. fMRI is very good at discovering where brain tissue is active (commonly by highlighting differences between brain activations during different cognitive tasks). But differences are not thoughts. fMRI can show differences in brain activation across locations, across time, and across tasks. But that often does not enable any reliable conclusion about precisely what a person is thinking.[30]

3. Scanners don't create fMRI brain images; people create fMRI brain images.

¶32 Images are only as good as the manner in which the researcher designed the specific task or experiment, deployed the machine, collected the data, analyzed the results, and generated the images. It is important to remember that fMRI images are the result of a process about a process. Multiple choices and multiple steps go into determining exactly what data will be collected, how, and when—as well as into how the data will be analyzed and how it will be presented.

4. Group-averaged and individual brain images are importantly different.

¶33 Most brain imaging research is directed toward understanding how *the average brain*, within a subject population, is activated during different tasks. This is not at all the same thing as saying either that all brains performing the same task activate in the average way, or saying that the activation of a single brain can tell us anything meaningful about the operation of the average brain. Consequently:

Do not assume that the scan of any individual is necessarily representative of any group.

Do not assume that the averaged scan of any group will necessarily be representative of any individual.

---

[30] There appear to be some exceptions. *See, e.g.*, John-Dylan Haynes et al., *Reading Hidden Intentions in the Human Brain,* 17 CURRENT BIOLOGY 323 (2007) (determining through brain imaging, with up to 71% accuracy, which of two tasks a person is covertly intending to perform); Y. Kamitani & F. Tong, *Decoding the Visual and Subjective Contents of the Human Brain*, 8 NATURE NEUROSCIENCE 679 (2005) (determining through brain imaging, with near 80% accuracy, which of two overlapping visual patterns a person is paying attention to); S. A. Harrison & F. Tong, *Decoding Reveals the Contents of Visual Working Memory in Early Visual Areas,*458 NATURE 632-35 (2009) (determining through brain imaging, with 83-86% accuracy, which of two visual patterns a person is actively maintaining in memory).

5. There is no inherent meaning to the color on an fMRI brain image.

¶34 fMRI does not detect colors in the brain. fMRI images use colors—of whatever segment of the rainbow the researcher prefers—to signify *the result of a statistical test*. By convention, the brighter the color (say, yellow compared to orange) the greater the statistical significance of the differences in brain activity between two conditions. Put another way, the brighter the color, the less likely it is that the differences in brain activity in that voxel or region, between two different cognitive tasks, was due to chance alone. As with any color-coded representation, accurate interpretation requires knowing exactly what each color represents in absolute terms. The researcher specifies what each color will represent, and this matters. Yellow might mean that there is only one chance in one thousand that the difference between brain activations in this voxel, between condition, is due to random chance. Or, yellow might mean that there is one chance in twenty that the difference is due to random chance.[31]

6. fMRI brain images do not speak for themselves.

¶35 No fMRI brain image has automatic, self-evident significance. Even well-designed, well-executed, properly analyzed, properly generated images must have their import, in context, interpreted.

7. Classification of an anatomical or behavioral feature of the brain as normal or abnormal is not a simple thing.

¶36 Because we have learned a great deal about the brain, from dissection, imaging, and the like, we have some confidence about what a typical brain looks like, and how a typical brain functions. But even without full anatomical scans of everyone on the planet, we know there is considerable variation—both anatomically and functionally—within some general parameters. That means that it can be (with some exceptions, such as a bullet lodged in the brain) difficult to say with precision *how* uncommon a given feature or functional pattern may be, even if it appears to be atypical. Base rates for anatomical or functional conditions are often unknown. For example: suppose brain images show that a defendant has an abnormal brain feature. We often do not have any idea how many people with nearly identical abnormalities do not behave as the defendant did. How, then, to make a reasonable conclusion about the causal effect of the brain condition?

8. Even when an atypical feature of function is identified, understanding the meaning of that is considerably complex.

¶37 Brain images can show unique features and functions of a person's brain. But the meaning of them is rarely self-evident. Determining which of those are important, and how, depends not only on the legal context for which the images are offered, but also on expert analysis of what the images do and do not mean. For example, suppose that measurement of the fMRI-detected signal during a given cognitive task indicates that a person has less neural activity in a given region than does the average person. Does that mean that the person is somehow cognitively impaired in that region? Or might it alternatively indicate that the person has more expertise or experience than average, requiring less cognitive effort?

9. Correlation is (still) not causation.

¶38 The fact that two things vary in parallel tells us little about whether the two are necessarily causally related and, if so, which causes which. For example, suppose brain imaging reveals that

---

[31] Consider this quote from a popular account:

> With PET, for example, a depressed brain will show up in cold, brain-inactive deep blues, dark purples, and hunter greens; the same brain when hypomanic however, is lit up like a Christmas tree, with vivid patches of bright reds and yellows and oranges. Never has the color and structure of science so completely captured the cold inward deadness of depression or the vibrant, active engagement of mania.

KAY REDFIELD JAMISON, AN UNQUIET MIND: A MEMOIR OF MOODS AND MADNESS 196 (1995). Our point here is that the colors used are arbitrary, and may have been represented in this way to create precisely this impression.

seventy percent of inmates on death row for homicide have atypical brain activation in a given region, compared to normal, unincarcerated subjects. That statistic does not mean that the brain activation pattern causes homicidal behavior. It might mean that having murdered affects brain activations, or that being incarcerated for long periods of time affects brain activations, or something else entirely.

10. Today's brain is not yesterday's brain.

¶39 In all but the most fanciful of contexts, a brain scan likely takes place long after the behavior (such as criminal activity) that gives rise to the scan. Drawing causal inferences is therefore further complicated. People's brains change with age and experience. And some proportion of the population will develop atypical anatomical or functional conditions over time. If a defendant is scanned six months or six years after the act in question, and the scan detects an abnormality, it is not a simple matter to conclude with confidence that the same abnormality was present at the time in question or—even if one assumes so, arguendo—that it would have meaningfully affected behavior.

11. Scanners (in theory) detect what they are built, programmed, and instructed to detect, in the way they are built, programmed, and instructed to detect it.

¶40 Scanners are highly complex and often unique pieces of machinery. So (as in other areas of science) are the people who calibrate, program, operate, and interpret collected data. It is important to recognize that the product of these intersecting complexities may or may not be reliable, generalizable, and replicable.

12. fMRI brain imaging enables inferences about the mind, built on inferences about neural activity, built on the detection of physiological functions believed to be reliably associated with brain activity.

¶41 It is important to remember that fMRI does not provide a direct measure of neuronal activity—as do, for example, invasive techniques that measure single neuron recordings. fMRI detects fluctuations in oxygen concentrations thought to be reliably associated with neuronal activity. But the precise relationship between metabolic demands and neuronal function remains poorly understood.

¶42 Even if regional activations in brain images reflect true neural activity, it should also be kept in mind that our ability to confidently infer the cognitive process that must have led to such regional activation is highly constrained. This is because neuroscientists still understand so little about what the various regions of the human brain contribute to a particular cognitive function.

CONCLUSION

¶43 We have provided above a very brief introduction to the intersection of brain imaging and law (and provide in the Appendix a step-by-step tour of a neurolaw brain-imaging study) principally intended for those relatively new to this interdisciplinary intersection.

¶44 Courts are already frequently confronted with issues concerning the admissibility and proper interpretation of brain images. And all present indicators suggest that brain images will be proffered by more lawyers in more cases in more contexts for more purposes in the future.

¶45 On one hand, the issues for the legal system are simply the same as they long have been: What might the proffered evidence tell us that may help us to answer legally identified questions in fair, effective, and efficient ways? Brain imaging is simply the latest high-tech tool to be offered for its potential assistance in this age-old enterprise.

¶46 On the other hand, brain imaging represents a perfect storm of power, to be used or abused. It combines the authoritative patina enjoyed by scientific evidence generally, and the allure of all-modern brain science specifically, with the seductive power of visual images.

¶47 How the legal system will ultimately deal with the exogenous shock of such technologically, rhetorically, and visually powerful information remains to be seen. To deal with it well, however, the

legal system will need the combined efforts and advice of many legal and neuroscientific scholars,[32] such as those populating the MacArthur Foundation Law and Neuroscience Project,[33] the Gruter Institute for Law and Behavioral Research,[34] and the Society for Evolutionary Analysis in Law (SEAL).[35] And, fortunately, many efforts are underway. In the meantime, legal thinkers likely to encounter brain images in their work would be well-advised to lay carefully constructed mental templates, on which to hang existing and future information emerging from brain-imaging communities. We hope that what we have discussed here will provide a useful means for doing so.

[32] *See, e.g.,* LAW AND THE BRAIN (Semir Zeki & Oliver Goodenough, eds., 2006); LAW, MIND, AND BRAIN (Michael Freeman & Oliver R. Goodenough eds., 2009); NEUROSCIENCE AND THE LAW: BRAIN, MIND, AND THE SCALES OF JUSTICE (Brent Garland ed., 2004); George J. Annas, *Foreword: Imagining a New Era of Neuroimaging, Neuroethics, and Neurolaw*, 33 AM. J.L. & MED. 163 (2007); Bruce A. Arrigo, *Punishment, Freedom, and the Culture of Control: The Case of Brain Imaging and the Law*, 33 AM. J. L. & MED. 457 (2007); Abram S. Barth, Note and Comment, *A Double-Edged Sword: The Role of Neuroimaging in Federal Capital Sentencing*, 33 AM. J.L. & MED. 501 (2007); Nita Farahany, *Cruel and Unequal Punishments*, 86 WASH. U. L.R. 859 (2009); Neal Feigenson, *Brain Imaging and Courtroom Evidence: On the Admissibility and Persuasiveness of fMRI*, 2 INT'L J. L. CONTEXT 233 (2006); Brent Garland & Paul W. Glimcher, *Cognitive Neuroscience and the Law*, 16 CURRENT OPINION IN NEUROBIOLOGY 130 (2006); Steven Goldberg, *MRIs and the Perception of Risk*, 33 AM. J.L. & MED. 229 (2007); Oliver R. Goodenough, *Mapping Cortical Areas Associated with Legal Reasoning and Moral Intuition*, 41 JURIMETRICS J. 429 (2001); Henry T. Greely, *Neuroscience and Criminal Justice: Not Responsibility but Treatment*, 56 U. KAN. L. REV. 1103 (2008); Henry T. Greely, *Remarks on Human Biological Enhancement*, 56 U. KAN. L. REV. 1139 (2008); Henry T. Greely & Judy Illes, *Neuroscience-Based Lie Detection: The Urgent Need for Regulation*, 33 AM. J.L. & MED. 377 (2007); Joshua Greene & Jonathan Cohen, *For the Law, Neuroscience Changes Nothing and Everything*, 359 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON B: BIOLOGICAL SCI. 1775 (2004); Charles N. W. Keckler, *Cross Examining the Brain: A Legal Analysis of Neural Imaging for Credibility Impeachment*, 57 HASTINGS L.J. 509 (2006); Laura Stephens Khoshbin & Shahram Khoshbin, *Imaging the Mind, Minding the Image: An Historical Introduction to Brain Imaging and the Law*, 33 AM. J.L. & MED. 171 (2007); Adam J. Kolber, *Therapeutic Forgetting: The Legal and Ethical Implications of Memory Dampening*, 59 VAND. L. REV. 1561, 1623-24 (2006); Adam Kolber, *Pain Detection and the Privacy of Subjective Experience*, 33 AM. J.L. & MED. 433 (2007); Jennifer J. Kulynych, *The Regulation of MR Neuroimaging Research: Disentangling the Gordian Knot*, 33 AM. J.L. & MED. 295 (2007); Jonathan H. Marks, *Interrogational Neuroimaging in Counterterrorism: A "No-Brainer" or a Human Rights Hazard?*, 33 AM. J.L. & MED. 483 (2007); Terry A. Maroney, *The False Promise of Adolescent Brain Science in Juvenile Justice*, 85 NOTRE DAME L. REV. (forthcoming 2010); Dean Mobbs et al., *Law, Responsibility, and the Brain*, 5 PLOS BIOLOGY 693 (2007); Stephen Morse, *Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note*, 3 OHIO ST. J. CRIM. L. 397 (2006); Erin Ann O'Hara, *How Neuroscience Might Advance the Law*, 359 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON B: BIOLOGICAL SCI. 1677 (2004); Purvak Patel et al., *The Role of Imaging in United States Courtrooms*, 17 NEUROIMAGING CLINICS N. AM. 557 (2007); Mark Pettit, Jr., *FMRI and BF Meet FRE: Brain Imaging and the Federal Rules of Evidence*, 33 AM. J.L. & MED. 319 (2007); Richard E. Redding, *The Brain-Disordered Defendant: Neuroscience and Legal Insanity in the Twenty-First Century*, 56 AM. U. L. REV. 51 (2006); Robert Sapolsky, *The Frontal Cortex and the Criminal Justice System*, 359 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON B: BIOLOGICAL SCI.1787 (2004); Alexander McCall Smith, *Human Action, Neuroscience, and the Law*, in THE NEW BRAIN SCIENCES: PERILS AND PROSPECTS 103 (Dai Rees & Steven Rose eds., 2004); O. Carter Snead, *Neuroimaging and the "Complexity" of Capital Punishment*, 82 N.Y.U. L. REV. 1265 (2007); Sarah E. Stoller & Paul Root Wolpe, *Emerging Neurotechnologies for Lie Detection and the Fifth Amendment*, 33 AM. J.L. & MED. 359 (2007); Laurence R. Tancredi & Jonathan D. Brodie, *The Brain and Behavior: Limitations in the Legal Use of Functional Magnetic Resonance Imaging*, 33 AM. J.L. & MED. 271 (2007); Erich Taylor, Note, *A New Wave of Police Interrogation? "Brain Fingerprinting," The Constitutional Privilege Against Self-Incrimination, and Hearsay Jurisprudence*, 2006 U. ILL. J.L. TECH. & POL'Y 287 (2006); Sean Kevin Thompson, *A Brave New World of Interrogation Jurisprudence?*, 33 AM. J.L. & MED. 341 (2007); Stacey A. Tovino, *Functional Neuroimaging Information: A Case for Neuro Exceptionalism?*, 34 FLA. ST. U. L. REV. 415 (2007); Stacey A. Tovino, *Imaging Body Structure and Mapping Brain Function: A Historical Approach*, 33 AM. J.L. & MED. 193 (2007); Special Issue, *International Perspectives on Brain Imaging and the Law,* 26 BEHAV. SCI. & L. 1 (2008).

[33] The Law & Neuroscience Project, http://www.lawandneuroscienceproject.org (last visited Apr. 14, 2009).

[34] Gruter Institute, http://www.gruterinstitute.org (last visited Apr. 14, 2009).

[35] Society for Evolutionary Analysis in Law, http://www.sealsite.org (last visited Apr. 14, 2009).

Appendix:

The Neural Correlates of Third-Party Punishment

JOSHUA W. BUCKHOLTZ[1, 2], CHRISTOPHER L. ASPLUND[1, 2], PAUL E. DUX[1], DAVID H. ZALD[1,5], JOHN C. GORE[3,5,8,9], OWEN D. JONES,[5, 6, 7] AND RENÉ MAROIS[1, 4, 5] *

SUMMARY

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

* This Appendix contains an annotated version of the article orginally published at 60 NEURON 930 (2008).

Department of Psychology[1], Neuroscience Graduate Program[2], Institute of Imaging Science[3], Vision Research Center[4], Center for Integrative and Cognitive Neurosciences[5], Law School[6], and Department of Biological Sciences[7] and Departments of Radiology and Radiological Sciences[8] and Biomedical Engineering[9] of Vanderbilt University.

INTRODUCTION

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

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

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

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

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

¶6 In this study, we used event-related fMRI to reveal the neural circuitry supporting third-party decision-making about criminal responsibility and punishment. Given that these two legally distinct judgments are rendered on the basis of differing information and considerations (LaFave et al., 2007), we were

Comment [A1]: In this instance, "neural mechanisms" refers to the manner by which the brain encodes and processes information to enable a specific cognitive ability.

Comment [A2]: There are two basic experimental designs in fMRI, "block" and "event-related." In block designs subjects encounter long sequences (or "blocks") of the same kind of stimulus (e.g., pictures of various faces) interspersed with blocks of a control stimulus (e.g., pictures of shapes). Average brain activity in one block is then contrasted to average brain activity in the other block. In event-related designs, subjects encounter randomly intermixed stimuli (e.g. faces and shapes). The choice between designs depends on what is being investigated.

Comment [A3]: "fMRI" stand for "functional magnetic resonance imaging." By convention, the leading "f" is lowercase.

Comment [A4]: "Neural circuitry" refers to interconnected brain regions that interact, like a wired circuit, during information processing. Within the circuit, each brain region has a specialized function that contributes to the brain's information-processing task.

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

Comment [A5]: What governs study size? fMRI scan sessions are expensive, frequently extending 1.5 hours, at $300 to $600 per hour. In determining a suitable number of subjects, statistical power (the probability that a real experimental effect will be detected) trades against cost. As a general rule, studies with fewer than 10 subjects are treated with skepticism.

Comment [A6]: To prevent changes in subjects' brain responses that are due to variables not under the experimenter's control, researchers keep variations to a minimum. Here, the protagonist's name is kept constant across all scenarios, to avoid confounds that could follow if different names were used. Again, there are trade-offs: the possible confound of using the same name repeatedly (which risks subjects cumulating their reactions to John's behavior, despite instructions not to) was considered less problematic than that different brain activations could be caused by different subject associations with different names.

Comment [A7]: fMRI data are extremely "noisy," in the sense that a small but true brain "signal" of interest that changes with the experimental manipulation can be obscured by much larger but irrelevant brain activation differences between experimental conditions. Since noise is random, while the true signal is not, researchers can detect changes in true signal by averaging the signal across all trials (enabling noise to cancel out). Consequently, researchers aim to pack as many experimental trials as possible (here, 50) into a given 60-90 minute scan session. Averaging across a large number of trials increases the likelihood of detecting the experimentally manipulated signal.

Comment [A8]: It is common to hold constant other variables in an experiment (here, gravity/severity of the harm), to ensure that any changes in brain activity between two conditions are due to the variable being investigated, rather than other factors.

Comment [A9]: It is common to include control stimuli. Here, a "No-Crime" control was included to provide a baseline level of brain activity that is associated with subjects viewing a protagonist intentionally engaged in a relatively harmless act. Thus, the experimenter is able to disentangle brain activity associated with viewing intentional action per se from that associated with viewing intentional actions that are potentially criminal.

Comment [A10]: For the control condition to be maximally useful, it must be as similar as feasible to the main conditions (in length, subject task, and general format, for example).

Comment [A11]: Because even slight head motion interferes with accurate data collection, scanned subjects must generally indicate responses with their hands, by pressing buttons, moving a joystick, or rolling a trackball. Here, each finger had a separate button. The buttons corresponded to a relative (i.e., internal/subjective) scale of punishment, rather than to some absolute metric, because that enabled more meaningful comparisons between subjects (since subjects could differ widely in their personal upper limits of actual punishment).

Comment [A12]: In general, counterbalancing helps diminish the potential confounding effects of variables not being studied. For example, any effect of order of presentation, when encountering multiple stimulus types, can be neutralized or diminished by randomizing the presentation order of stimuli. The counterbalancing in this experiment ensured that equal numbers of participants saw each group of scenarios.

Comment [A13]: Here, the counterbalancing ensured that different brain activity between different scenarios was likely a function of the level of responsibility manipulated as a variable, rather than a function of some other difference (such as location, item stolen, etc.) between the two scenarios.

RESULTS

Behavioral Data

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

Comment [A14]: Behavioral data, in brain imaging contexts, are measurements of subject responses that are separate from the collection of brain images. Typically, these are not recorded by the MRI machine. Here, for example, behavioral data include the punishment rating each subject selected for each scenario, and the elapsed time between presentation of scenario and selection of punishment.

Comment [A15]: "Significant" is an important term of art in science. In scientific experiments, observed results can be due to three things: 1) the factor that that the experimenter thinks the results are due to (i.e., the experimental manipulation); 2) an unmanipulated factor that the experimenter hasn't thought of or controlled (i.e., a "confound"); or 3) random chance (i.e., a "false positive"). A claim of significance is ordinarily accompanied by a numerical representation (a "p" value) of the probability that the results arose by random chance. For example: $p < .05$ indicates that there is less than 5 chances in 100 that the result described could have arisen by chance alone. In setting a p-value to a given value, an investigator allows for the fact that there is a certain set probability that any effect is due to random chance, and it is near-universally agreed that $p < 0.05$ is a "reasonable" threshold. Statistical software outputs a p-value for each experimental comparison of interest. Thus, referring to something as "significant" in this context ordinarily means that the experimenter has submitted an experimental measure to a statistical test, and the outcome of this test allows the experimenter to be confident that the results have less than a 5% probability of being due to random chance. In some instances, a p-value may be set lower (e.g., to .01) to allow stricter control over the possibility of obtaining a false-positive.

Comment [A16]: The significant results from the statistical tests allow the authors to state that rating differences between conditions were due to the experimental manipulation. Briefly, in psychology and neuroscience, an independent variable is the factor that the experimenter manipulates to cause some effect on the dependent variable. When we talk about an effect of condition, we're talking about the effect of one or more independent variables on one or more dependent variables. Here, the dependent variable is punishment ratings and the independent variable is scenario category (which has three "conditions" or "levels"): Responsibility, Diminished Responsibility, and No-Crime.

Comment [A17]: These refer to the outcome of the statistical tests. In this case, an Analysis of Variance (ANOVA) test was used – the "F" value gives an indication of the strength of the experimental manipulation, and can be understood to represent the size of the difference in scores conditions, while the "p" value indicates that there is less than 1 chance in 1000 that these condition differences could have arisen by chance

Comment [A18]: S.E. stands for "standard error (of the mean)" which helps readers understand the estimated stability of the measurement across samples. Essentially, this indicates the likelihood that the mean value will "jump" around between different samples of subjects.

Comment [A19]: The paired t-test is a common statistical test used to test for the effects of a condition on a dependent measure.

Comment [A20]: This means the authors' key experimental manipulation (here, protagonist's criminal responsibility) affected how much punishment subjects gave to the protagonist.

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

Comment [A21]: The authors sought to quantify the subjects' emotional responses to the scenarios because they hypothesized that emotional responses could influence punishment decisions.

Comment [A22]: Subjects rated each of the 50 scenarios (presented in random order on a computer screen outside the scanner) on the basis of how emotionally aroused they felt following its presentation (0 = calm, 9 = extremely excited.

Comment [A23]: "Main effect" is a term of art referring to the effect of one experimentally manipulated factor (e.g. protagonist responsibility) on one experimental variable (e.g. reaction time). Often, investigators are interested in looking at the interactive effects of two or more conditions on a dependent variable. The term "main effect" is used to indicate that the influence of one independent variable was examined in isolation.

Comment [A24]: Reaction time is the length of time elapsing between the "onset" (when the subject was first presented with the scenario to consider) and the behavioral response (here, pressing a button to select a punishment level).

fMRI Data: Criminal Responsibility

¶9 To identify brain regions that were sensitive to information about criminal responsibility, we contrasted brain activity between Responsibility and Diminished-Responsibility scenarios. The resulting statistical parametric map (SPM) revealed an area of activation in the right dorsolateral prefrontal cortex (rDLPFC, Brodmann Area 46, peak at Talaraich coordinates 39, 37, 22 [x,y,z]; Figure 2a) that was significantly more activated in the Responsibility than in the Diminished-Responsibility condition. Time course analyses of peak activation differences confirmed that there was greater rDLPFC activity in Responsibility compared to Diminished-Responsibility or No-Crime conditions (R>NR, $p = 0.002$; R>NC, $p = 0.0004$; paired *t*-tests; see Figure 2b) and no difference between the Diminished-Responsibility and No-Crime conditions ($p = 0.19$). No effect of condition was found in the left DLPFC ($p > 0.2$ for all paired comparisons; see Methods), and the right DLPFC was significantly more engaged than the left DLPFC in the Responsibility condition ($p = 0.04$, paired *t*-test), suggesting that punishment-related prefrontal activation is confined to the right hemisphere. Bilateral anterior intraparietal sulcus (aIPS) demonstrated a pattern of responsibility-related activity that was similar to rDLPFC (Table 1, Supplementary Figure 1, Supplementary Results), whereas the temporo-parietal junction (TPJ) showed the reverse pattern, with more activity for the Diminished-Responsibility than the Responsibility condition (Table 1, Figure 3, see below).

Comment [A25]: All BOLD fMRI studies are based on comparisons of BOLD signal between two conditions. By subtracting BOLD signal (within a given brain region) during one condition from brain BOLD signal (within that same brain region) during another condition, the effect of the experimental manipulation on regional brain function can be estimated. Here, because the only factor that differs between experimental conditions is information about protagonist responsibility, this subtraction method allows fMRI investigators to remove brain activation that is not due to the independent variable. In this study, the authors took an average of the measured brain signal during each R scenario, an average of the measured brain signal during each DR scenario, an average of the measured brain signal during each NC scenario, and compared these condition-averaged signals for each subject. They then took an average across all subjects to see where in the brain the signal was significantly different between levels of the independent variable.

Comment [A26]: An "area of activation" is a region of the brain where the measured fMRI signal was significantly greater during one condition (e.g. R) compared to another (e.g. DR).

Comment [A27]: There are several ways to designate brain regions. One rather general way uses 45 "Brodmann's Areas." These are based on a classification scheme devised by Korbinian Brodmann (1868-1918), separating areas by neuron type and organization.

Comment [A28]: The "peak" refers, in this instance, to the specific region of the brain that demonstrated the strongest effect of condition.

Comment [A29]: Human brains vary widely in size and shape. Because fMRI investigators average condition differences in brain activity across subjects, it is imperative that a brain region on one subject correspond to the exact same brain region on another subject. Thus, before an fMRI investigator can compare brain activity between conditions, each subjects' brain must first be "normalized" into a common space. Neuroimagers therefore translate (or "warp") subjects' brains into a single, common brain space. The most frequently used template is that defined by the coordinate system of Talairach and Tourneaux. After warping into Talaraich space, which has a standardized three-dimensional coordinate system based on neuroanatomical landmarks, regional brain activation can be compared between subjects - and importantly, across studies. So, in this section of the paper, the authors are describing precisely where changes in brain function occurred. Roughly speaking, rDLPFC is like designating a city, Brodmann Area 46 is the street, and Talairach coordinate is the precise street number.

Comment [A30]: Time course analyses examine what is happening, over time, within a given region of the brain, during the cognitive task performed.

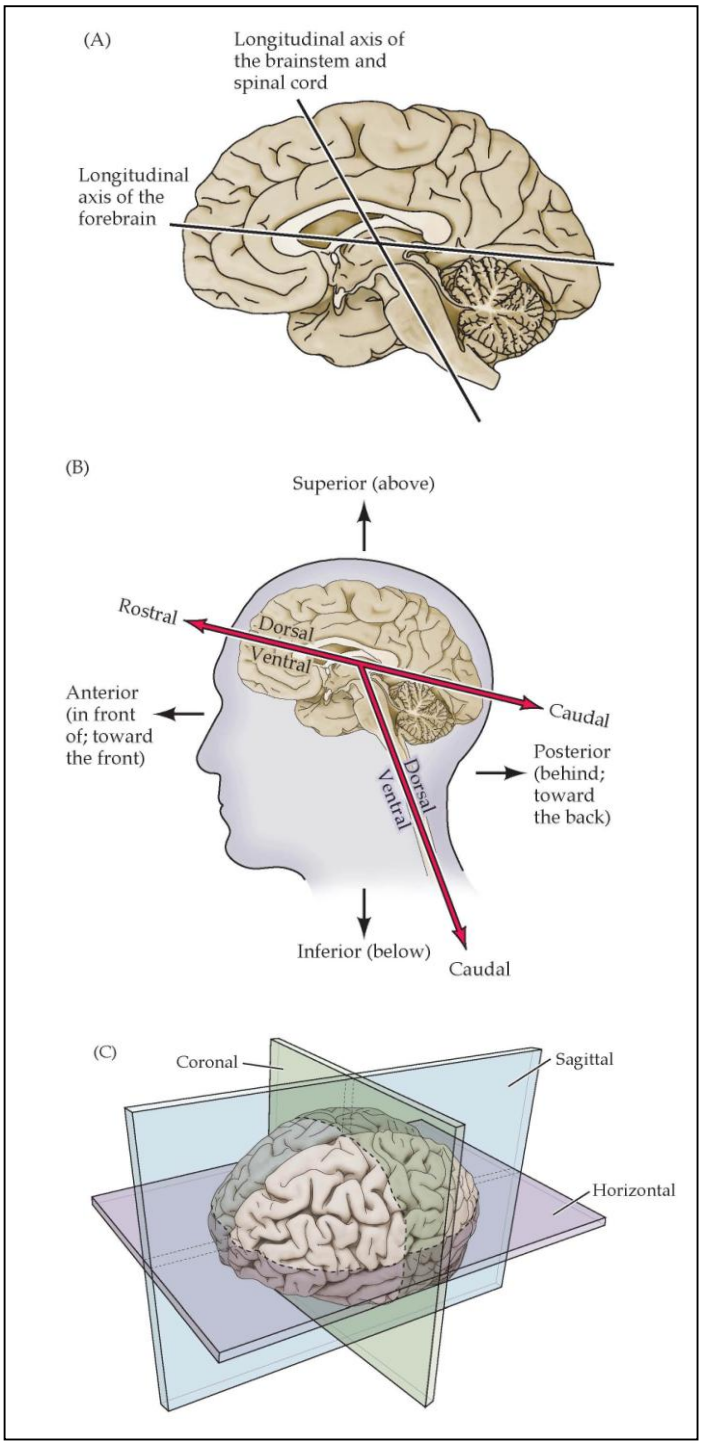
Comment [A31]: "Peak activation" refers to the maximum amplitude of BOLD signal (and hence, by inference, brain activity) within a given brain region. The sentence here describes an analysis of the different times at which, under different conditions, the maximum BOLD signal appeared within the brain region.

Comment [A32]: In articles describing experimental results, the experimental and analytical methods are ordinarily and carefully described in a separate "Methods" section.

Comment [A33]: The brain contains two largely independent hemispheres, left and right. Anatomical features (such as the amygdala) generally appear separately in both hemispheres.

Comment [A34]: For a depiction of brain orientation, see Box 1 (next page).

Box 1.
(Accompanies Comment 34)



Reproduced by permission of Sinauer Associates
from Dale Purves et al., Neuroscience, 2d edition 2001.

¶10 The greater rDLPFC activation in the Responsibility condition did not simply result from longer time-on-task: response times (RTs) to Responsibility scenarios were shorter than Diminished-Responsibility scenarios (p = 0.005, paired t-test), and the effect of condition on rDLPFC activity was still significant when response time was used as a covariate in an analysis of covariance (ANCOVA, $F(1,37)$ = 10.15, p = 0.003) or when response times were equated between conditions (see Methods; R>DR, p = 0.006; R>NC, p = 0.002; Supplementary Figure 2). In addition, rDLPFC activity was not correlated with reaction time (p = 0.09 in Responsibility scenarios, p = .12 in Diminished-Responsibility scenarios). We also assessed whether the activity pattern in rDLPFC might have been driven by between-condition differences in emotional arousal rather than by differences in criminal responsibility. To this end, we performed a peak activation difference analysis between the Responsibility and Diminished-Responsibility conditions after equating their mean arousal ratings (Responsibility = 3.62, Diminished-Responsibility = 3.50; p > 0.10, paired t-test; see Methods). The results still revealed greater rDLPFC activity in the Responsibility compared to the Diminished-Responsibility condition even in the absence of arousal differences (p = 0.0005, paired t-test).

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

¶12 In addition to the peak activation differences, the timecourse of rDLPFC activity revealed an early deactivation (negative percent signal change from baseline) around 8 s post-stimulus onset. Importantly, this early deactivation ('dip') does not account for the peak activation results outlined above: the activation differences between conditions at the dip do not predict corresponding activation differences at the peak (correlation of subjects' activity differences between the Responsibility and Diminished-Responsibility conditions at the dip and at the peak: ρ = -0.19, p = 0.49; Supplementary Figure 3; see Methods). Furthermore, rDLPFC activity during 'non-punished' Diminished-Responsibility and No-Crime trials strongly differed at the dip (p = 0.008) but not at the peak (p =

Comment [A35]: "Covariate" in this instance refers to a factor other than the independent measure that could contribute to condition differences in a dependent measure (in this case, fMRI signal). To ensure that this uncontrolled factor did not cause the observed condition differences in the dependent measure, the authors performed a test (called an ANCOVA) to see if condition differences remained even after taking that uncontrolled factor – the covariate – into account. A significant value for this test indicates that the covariate did not drive the observed differences in fMRI signal.

Comment [A36]: Response time differences between conditions might influence the observed pattern of brain results in a manner that was not anticipated or desired by the investigators. To control for this potential confound, an ANCOVA was employed.

Comment [A37]: A "peak activation difference analysis" is simply a test to see if the peak activation within a brain region (see comment 31) differed significantly between two experimental conditions (here, the Responsibility and Diminished Responsibility conditions).

Comment [A38]: "Deactivation" in this context refers to the fact that BOLD signal in this region, at this particular moment in time, was lower (i.e., comparatively deactivated) after showing subjects the experimental scenarios.

Comment [A39]: s = seconds

Comment [A40]: Post-stimulus onset means after presentation of the experimental stimulus (scenario)

Comment [A41]: That between-condition activation differences at peak and at dip were not related to each other suggests that this region of the brain might be involved in two distinct activities at these different points in time.

0.97), indicating that peak activation differences are not simply carry-over effects from differences during the dip.

fMRI Data: Punishment Magnitude

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

¶14 Although rDLPFC activity was not proportional to punishment amount, a linear relationship between peak BOLD amplitude and punishment magnitude was found in a set of brain regions that have been extensively linked to social and affective processing. To isolate such effects, we compared Responsibility scenarios with high punishment ratings to those with low ratings (median split by scenario across subjects; see Methods). The resulting SPM revealed activation in the right amygdala (peak Talairach coordinates 29, -7, -13; Fig 4; Supplementary Figure 5) as well as in other brain regions commonly associated with social and affective processing (LeDoux, 2000; Phelps, 2006; Phillips et al., 2003; Price, 2005), including the posterior cingulate**,** temporal pole, dorsomedial and ventromedial prefrontal cortex, and inferior frontal gyrus (Supplementary Table 2; Supplementary Figure 4, Supplementary Figure 5). The association between amygdala activity and punishment magnitude was further demonstrated by a strong correlation between amygdala BOLD signal and punishment ratings across Responsibility scenarios (ϱ =.70, p = 0.001; Fig. 4). However, punishment rating was not the only variable that correlated with amygdala function, as participants' arousal ratings yielded a similar correlation with amygdala activity (ϱ = 0.67, p = 0.001), and punishment and arousal ratings were themselves highly correlated (ϱ = 0.98, p = 0.000001). Correlations between peak BOLD signal and punishment ratings (and between peak BOLD signal and arousal ratings) also held for a number of the other affective regions, including ventromedial prefrontal cortex and posterior cingulate cortex (Supplementary Table 2; Supplementary Figures 4 and 5), indicating that the relationship between affective processing and punishment involved a distributed neural circuit.

¶15 Although the correlation between amygdala activity and punishment scores could be interpreted as evidence for a role of emotional arousal in the assignment of deserved punishment, it is also possible that such activity simply reflected subjects' emotional reaction to the graphical content

Comment [A42]: Signal amplitude here is a synonym for peak activation.

Comment [A43]: Correlation is a statistical test to see if two variables are related. A linear correlation means that as one variable increases in value, so does another (positive correlation). Alternatively, a negative correlation refers to a relationship wherein as one variable increases in value, another exhibits a commensurate decrease.

Comment [A44]: The greek letter *rho* refers to the value of a statistical test for correlation (the Spearman test).

Comment [A45]: I.e., represented in the brain by.

Comment [A46]: See comment 25 for explanation of BOLD.

Comment [A47]: "Affective" is a psychological term of art, meaning "emotional."

Comment [A48]: A median split divides a set of experimental observations (here, punishment ratings) into two groups split at the median, such that the higher half of the set is in one group, and the lower half of the set is in the other.

Comment [A49]: This correlation suggests that subjects' emotional responses to a scenario correlated positively with how much punishment subjects will assign to the protagonist in that scenario.

Comment [A50]: See comment 4 for description of neural circuits.

of the scenarios rather than its involvement in the decision-making process *per se*. To avoid the potential arousal confound inherent to an examination of criminal scenarios that differ in graphic content (as was the case for our comparison of high vs. low punishment scores within the Responsibility condition), we examined the relationship between punishment ratings and amygdala activity after controlling for the possible confounding effect of graphic arousal. Because Responsibility and Diminished-Responsibility scenarios were equated for graphic content and differed only by the presence of mitigating circumstances (see Methods), the potentially confounding contribution of graphic arousal to amygdala activity in the Responsibility scenarios can be controlled for by subtracting amygdala activity in the Diminished-Responsibility scenarios from that in the corresponding Responsibility scenarios. If amygdala activity appertains to punishment magnitude rather than, or in addition to, emotional arousal related to the graphic content of the scenarios, it should still track punishment ratings even after subtracting out graphic content differences in the scenarios. To this end, we created, for each pair of Responsibility and Diminished-Responsibility scenarios, punishment rating difference scores (Responsibility minus Diminished-Responsibility) and assessed whether these scores were correlated with the corresponding difference scores for peak amygdala BOLD signal. That correlation was significant ($\rho = 0.62$, $p = 0.001$; Figure 4), indicating that the magnitude of amygdala BOLD signal difference between Responsibility and Diminished-Responsibility conditions for a given scenario predicted a corresponding change in punishment rating for that scenario. Similar correlations were found in posterior cingulate and ventromedial prefrontal cortex (Supplementary Table 2). These findings suggest that activity within brain regions previously implicated in social and affective processing reflect third-party decisions about how much to punish, even after controlling for the potentially confounding arousal associated with the "graphic" content of the criminal scenarios.

Comment [A51]: It is incumbent on the investigator to prove that their effects are due to their experimental manipulation, and not to other uncontrolled factors that could explain their results just as well. Such an uncontrolled factor that could potentially explain the results better than the experimental manipulation is referred to as a confound.

DISCUSSION

¶16 The present findings suggest that the two fundamental components of third-party legal decision-making - determining responsibility and assigning an appropriate punishment magnitude - are not supported by a single neural system. In particular, the results reveal a key role for the right dorsolateral prefrontal cortex in third-party punishment. This brain region appears to be involved in deciding whether or not to punish based on an assessment of criminal responsibility. The only other brain region demonstrating a comparable pattern of responsibility-related activity (R>DR, R>NC, DR=NC) to rDLPFC was the anterior intraparietal sulcus (Supplementary Table 1, Supplementary Figure 1, Supplementary Results). This parietal region has been associated with a number of diverse cognitive functions including general response selection (Gobel et al., 2004) and quantitative numerical comparisons (Dehaene et al., 2003; Dehaene et al., 1999; Feigenson et al., 2004), which may hint at a role for this area in associating a specific action (punishment outcome) with a given scenario.

¶17 Our results also implicate neural substrates for social and affective processing (including amygdala, medial prefrontal cortex and posterior cingulate cortex) in third-party punishment, albeit in ways distinct from the rDLPFC. Specifically, while prefrontal activity was linked to a categorical aspect of legal decision-making (deciding whether or not to punish on the basis of criminal responsibility), the magnitude of assigned punishments for criminal transgressions parametrically modulated activity in affective brain regions, even after controlling for the potentially confounding arousal-related activity associated with the graphic content of the criminal scenarios. Our findings suggest that a set of brain regions (e.g. amygdala, medial prefrontal cortex, and posterior cingulate) consistently linked to social and emotional processing (Adolphs, 2002; Amodio and Frith, 2006; Barrett et al., 2007; Lieberman, 2007; Phelps, 2006; Phillips et al., 2003; Zald, 2003) is associated with the amount of assigned punishment during legal decision-making. As such, these results accord well with prior work pointing to social and emotional influences on economic decision-making and moral reasoning (De Martino et al., 2006; Delgado et al., 2005; Koenigs and Tranel, 2007) (Greene and Haidt, 2002; Greene et al., 2004; Greene et al., 2001; Haidt, 2001; Heekeren et al., 2003; Koenigs et al., 2007; Moll et al., 2002b; Moll et al., 2005), and provide preliminary neuroscientific support for a proposed role of emotions in legal decision-making (Arkush, 2008; Maroney, 2006). Our data concur with behavioral studies that have proposed a link between affect and punishment motivation in both second- and third-party contexts, and are consistent with the hypothesis that third-party sanctions are fueled by negative emotions towards norm violators (Darley

Comment [A52]: In the discussion section of scientific papers, the investigators comment on the significance of their findings, place these findings in the context of the current scientific literature, address possible shortcomings or limitations of the study, and make suggestions for future studies.

Comment [A53]: Neural substrates are brain regions that underlie a certain kind of information processing.

Comment [A54]: In this context, "parametrically" means that as the magnitude of punishment increases, so does brain activity in these regions.

Comment [A55]: Affect = emotion.

and Pittman, 2003; Fehr and Fischbacher, 2004a, b; Seymour et al., 2007). However, it must be acknowledged that the present conclusions rest exclusively on correlational data. Thus, additional research will be required to confidently determine the contributions of socio-affective brain regions to third-party punishment in the absence of any graphic arousal confound. In particular, it will be important in future experiments to fully dissociate the factors of crime severity and arousal by employing task conditions that manipulate arousal without affecting crime severity. Furthermore, future research should also focus on determining how these affective brain regions interact with dorsolateral prefrontal cortex during third-party punishment decisions.

Comment [A56]: A common criticism of fMRI is that it is inherently correlational. Brain activity changes are *correlated* with changes in the independent variable, but one cannot say definitively that the independent variable *caused* those brain activity changes. Nor can one definitively say that the regions identified by this correlational approach are necessary or sufficient for the kind of cognitive process under study (e.g. legal decision-making).

¶18 An additional concern in interpreting our findings, or any others based on simulated judgments, is whether they are relevant to real-world decision-making. After all, the punishment decisions made by our participants did not have direct, real-world consequences for real criminal defendants. Thus, it remains to be seen if our findings, generated by examining brain activation patterns during "hypothetical" judgments, will generalize to circumstances in which "real" punishments are made. However, there is some evidence suggesting that the hypothetical judgments made by our subjects may be a good proxy measure for real-world legal judgments. For example, post-scan debriefing of our subjects indicated that their punishment assessments were implicitly legal, with lower numbers corresponding to low prison sentences and higher numbers corresponding to high prison sentences (see Supplementary Table 3). Thus, participants appeared to adopt an internal punishment scale based on incarceration duration - a legal metric - when making their judgments, even in the absence of explicit instructions to do so. Further, we found that participants' decisions about punishment amount for each of the crimes depicted in the Responsibility scenarios were strongly correlated with the recommended prison sentences for those crimes, according to the benchmark sentencing guidelines of North Carolina, a model state penal code (ϱ = 0.8, $p<.0001$; Supplementary Figure 6; see Methods). Thus, although our subjects were not literally applying a criminal statute to an accused individual, these data suggest that subjects' punishment decisions were consistent with statutory legal reasoning. However, despite these suggestions, further empirical studies are required to confirm our supposition that neuroimaging studies of simulated third-party legal decision-making can be valid models for understanding the neural basis of real-world legal reasoning.

Relative contributions of Temporo-Parietal Junction (TPJ) and rDPLFC to Third-Party Punishment Decisions

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

Comment [A57]: "Temporally specific," in this context, refers to the fact that legal decision-making likely relies on different brain regions communicating in specific ways *at very specific times* throughout the legal decision-making process.

Comment [A58]: "Node," in this context, refers to one specific brain region that participates as part of a neural circuit. In general, a circuit refers to two or more brain regions that interact cooperatively to enable some kind of cognitive function.

Comment [A59]: Functional opposition means that as brain activity in one network increases, brain activity in another tends to decrease.

Comment [A60]: In the present study, TPJ is shown to be activated during a period when rDLPFC is deactivated, suggesting that they oppose each other. This opposition is referred to as "antagonistic."

Comment [A61]: Biphasic in this context refers to the fact that the early (deactivation) and late (activation) periods of the rDLPFC timecourse appear to be associated with different cognitive functions.

Moral versus Legal Decision-Making

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

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

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

Neural convergence of second-party and third-party punishment systems.

Comment [A62]: Neural convergence is when a common brain region underlies two distinct, but related, cognitive processes.

¶22 The prefrontal cortex area activated in the present third-party legal decision-making study corresponds well to an area that is involved in the implementation of norm enforcement behavior in two-party economic exchanges (peak Talairach coordinates of 39, 37, 22 [x,y,z] for (Knoch et al., 2006; Sanfey et al., 2003); vs 39, 38, 18 [x,y,z] for the present study), raising the possibility that rDLPFC serves a function common to both third-party legal and second-party economic decision-making. In this respect, it is noteworthy that this region of rDLPFC is recruited when participants decide whether or not to punish a partner by rejecting an unfair economic deal proposed by that partner (Sanfey et al., 2003); this result is analogous to our finding that rDLPFC is activated by the decision to punish the perpetrator of a criminal act. Furthermore, while disruptive

magnetic stimulation of this region impairs the ability to punish economic norm violations in dyadic exchanges (Knoch et al., 2006; van 't Wout et al., 2005), this manipulation has no effect on norm enforcement behavior when the unfair economic exchanges are randomly generated by a computer instead of a human agent (Knoch et al., 2006). This result accords well with our finding that rDLPFC was much less activated when the scenario protagonist was not criminally responsible for his behavior, and supports the notion that this prefrontal cortex area is primarily recruited when punishment can be assigned to a responsible agent (Knoch et al., 2006). Finally, we still observed greater rDLPFC activity in the Responsibility condition (compared to Diminished-Responsibility scenarios) when we restricted our analysis to scenarios that only contained physical harms ($p < 0.005$, paired t-test), suggesting that the overlap of rDLPFC activity between studies of economic decision-making and the present examination of legal decision-making is not solely driven by scenarios describing economic transgressions.

Comment [A63]: Repetitive transcranial magnetic stimulation (rTMS) is a non-invasive technique that disrupts brain activity. In rTMS, a series of magnetic pulses are applied to a circumscribed region of the brain. These pulses temporarily interfere with brain activity in that region, creating a reversible "virtual lesion."

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

EXPERIMENTAL PROCEDURES

Subjects

¶24 Sixteen right-handed individuals (8 males, age 18-42) with normal or corrected-to-normal vision participated for financial compensation. The Vanderbilt University Institutional Review Board approved the experimental protocol, and informed consent was obtained from each subject after they were briefed on the nature and possible consequences of the study. A brief psychological survey was also administered to exclude individuals who may react adversely to the content of the criminal scenarios. Exclusion criteria included history of psychiatric illness, being the victim of or having witnessed a violent crime (including sexual abuse), and having experienced any trauma involving injury or threat of injury to the subject or a close friend/family member.

Comment [A64]: Because handedness can affect which side of the brain is used to process some kinds of information, group-averaged brain scan studies typically use subjects of one handedness or the other.

Comment [A65]: The Institutional Review Board (IRB) is charged with approving, monitoring, and reviewing human subjects research to make sure that subjects' rights are respected, and that research conforms with established ethical standards.

Comment [A66]: An experimental protocol is the specific "recipe" for executing a study. It details the process for recruiting subjects and running the experiment.

Comment [A67]: Subjects must provide "informed consent" to participate – that is, they must be fully informed about what to expect in an experiment, and what their rights are as a subject, before they are allowed to agree to participate.

Comment [A68]: Exclusion criteria are established reasons to exclude a subject from participating in the study.

Paradigm

¶25 In this experiment, subjects participated in a simulated third-party legal decision-making task in which they determined the appropriate level of punishment for the actions of a fictional protagonist described in short written scenarios. The principal goal of our study was to isolate the neural processes associated with the two fundamental processes of legal decision-making: deciding whether or not an accused individual is culpable for a given criminal act, and determining the appropriate punishment for that act (a parametric process based on the ordinal severity of a crime). Correspondingly, our design manipulated responsibility in a dichotomous fashion and crime severity in a continuous fashion. Each participant viewed 50 scenarios (some inspired by prior behavioral studies of relative blameworthiness (Robinson and Darley, 1995; Robinson and Kurzban, 2007)) depicting the actions of the protagonist named "John." The 50 scenarios were subdivided into three sets (complete scenario list available as Supplementary Methods). In the Responsibility set (N = 20), the scenarios described John intentionally committing a criminal action ranging from simple theft to rape and murder. The Diminished-Responsibility set (N = 20) included similar actions comparable in gravity to those in the Responsibility set, but contained circumstances that would often legally excuse or justify the otherwise criminal behavior of the protagonist. The No-Crime set (n=10) depicted John engaged in non-criminal actions that were otherwise structured similarly to the Responsibility and Diminished-Responsibility scenarios. The No-Crime scenarios were included to assist in interpreting activity differences between Responsibility and Diminished-Responsibility scenarios (see e.g. Fig. 2).

Comment [A69]: That is, the protagonist's responsibility was either full or diminished (dichotomous), but he was described committing a range of crimes ranging in severity from simple theft to rape and murder (continuous).

¶26 Two groups of 50 scenarios were constructed and their presentation counterbalanced across the 16 participants (8

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

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

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

Comment [A70]: Two sets of complementary scenarios were constructed so that no individual subject viewed the same core scenario details (e.g., unlawfully taking a book) twice – once as a responsibility scenario and then again as a diminished responsibility scenario. This arrangement would increase the likelihood that subjects would consciously detect our experimental manipulation. If subjects become conscious of an experimental manipulation, they tend to behave differently – an experimental artifact referred to as "demand characteristics."

Comment [A71]: Word length is a potential confound. If the subject reacts differently to two scenarios that not only have different content but also different length, then the differences in brain function might be caused by differences in reading time, instead of content.

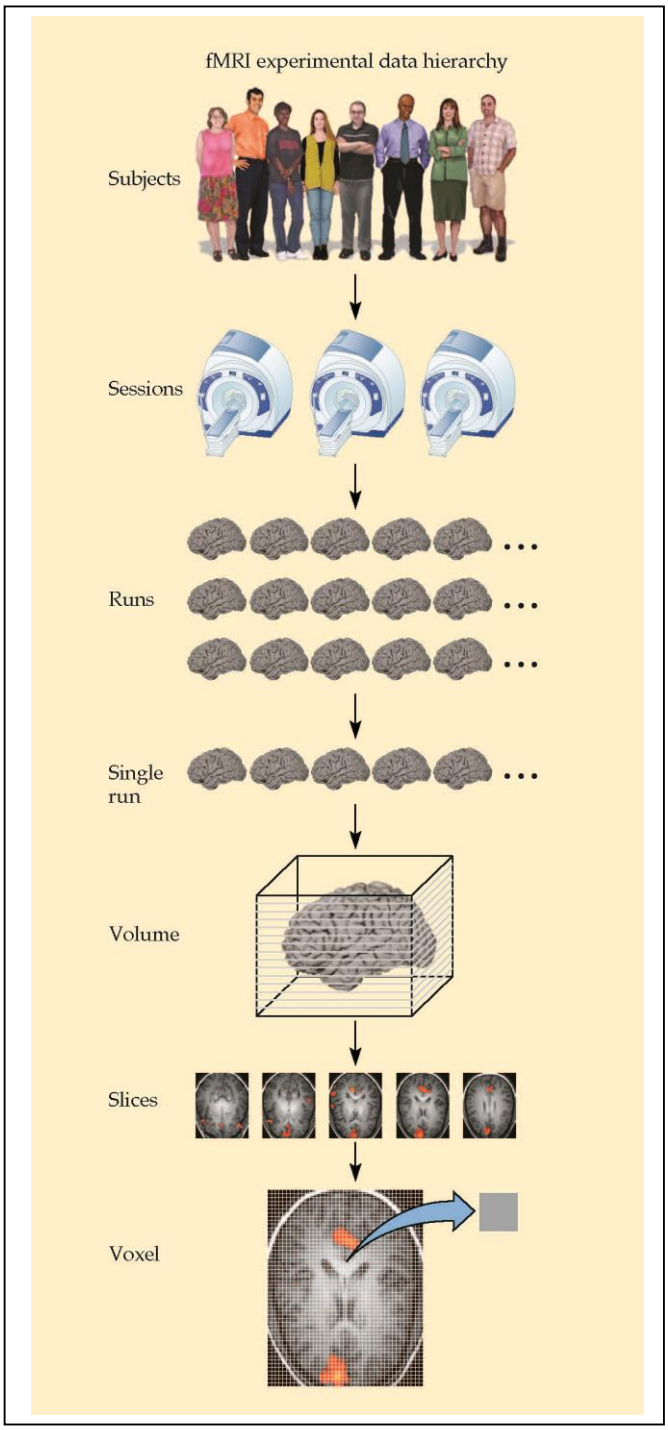
Comment [A72]: Functional scans refer to MRI scans that detect brain activity over time. In contrast, anatomical scans detect brain *structure* in detail, but without assessing brain *function*.

Comment [A73]: It is good practice in fMRI studies to give participants something to focus on in between trials in order to limit mind-wandering.

Comment [A74]: Researchers commonly want to avoid "order effects" – a confound whereby the order in which a condition type is presented (e.g. mostly R first or mostly NR first) creates response biases. It is also important to avoid "run effects" – a confound whereby certain trial types predominate in one run and are absent in another. To avoid both, we randomized trial types both within and between runs.

Comment [A75]: For a depiction of the differences between sessions, runs, volumes, slices, and voxels, see Box 2 (next page)

Box 2.
(Accompanies Comment 75)



Reproduced by permission of Sinauer Associates
from Scott A. Huettel et al., 2003.

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

Comment [A76]: Emotional *arousal* measures how emotionally excited a subject feels. Emotional *valence* describes the direction (positive or negative) of that arousal.

Comment [A77]: Multiple regression analyses are standard statistical processes for disentangling the multiple influences of multiple variables. More specifically, they examine independent and interactive influences of multiple *independent variables* on a given *dependent variable*. (See comments 16 and 35.)

Internal scale questionnaire

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

Comment [A78]: Debriefing is often an important feature of scientific studies involving human subjects, as it can help researchers understand the extent to which subjects' motivations and observed behavior matched the experimenter's expectations. In this case, the authors wanted to gauge precisely what subjects intended for a given punishment value, given that the scale (0-9) was by design subjective rather than objective and absolute.

Relationship between Punishment Ratings and Legal Statutes

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

¶32 For each responsibility scenario, we determined the crime(s) (such as larceny, involuntary manslaughter, or murder) with which John might reasonably be charged under the criminal code of North Carolina (2005 General Statutes of North Carolina, Chapter 14). We then determined, for each crime, the authorized presumptive sentencing range (such as 58 to 73 months in prison), assuming no aggravating or mitigating factors that could, under the statute, increase or decrease the authorized sentencing range (2005 General Statutes of North Carolina, Chapter 15A, Article 81). We then calculated and

assigned to each scenario the mean for this range, in months. As the distribution of sentence values was highly right-skewed, we log-transformed (natural log) to create a normal distribution of sentence values (we verified that non-transformed data produced similar correlations as transformed data). For scenarios with multiple crimes, the averages for each respective crime were summed (whether this summed value or simply the mean value for the most severe crime depicted in a given scenario was used in the correlation analysis did not significantly affect the results). Where the upper limit of the sentencing range was life in prison, it was coded as 29 years (which has been estimated as the average time likely to be served by lifers newly admitted in 1997)(Mauer et al., 2004). Similarly, where the upper limit of the sentencing range was death, it was also quantified as life in prison (29 years). The log-transformed mean sentences for each of the 20 scenarios were then correlated with the group-averaged punishment ratings for these scenarios.

Comment [A79]: To work properly, some statistical tests require that the data have certain features (e.g., a so-called "normal distribution" of values). The authors applied a standard mathematical transformation to the sentence scores to permit the use of these statistical tests.

Statistical Analysis

¶33 Mean punishment and arousal scores and reaction times were calculated for each subject for each condition (Responsibility, Diminished-Responsibility, and No-Crime) and entered into a repeated-measures Analysis of Variance (ANOVA) using SPSS 15 (SPSS Inc. Chicago, IL) to determine main effects and interactions. Data from 16 subjects were used for all analyses. Punishment, arousal scores and reaction times were compared between conditions and post-hoc tests were performed using Fisher's Least Significant Difference (LSD) measure using an alpha level of .05. Two-tailed tests were used in all cases. For correlational analyses, data from Responsibility scenarios (N = 20) were averaged across all (N = 16) subjects. Examination of scatterplots for the correlation of rDLPFC signal and punishment suggested the presence of outliers. As non-parametric correlations tend to be more robust to outliers, we used Spearman's ρ to measure correlations between fMRI signal, behavioral measures, and recommended sentences. All correlations that were significant using Spearman's ρ were also significant ($p < 0.05$) when we employed Pearson's r.

Comment [A80]: The alpha level defines an acceptable rate of "Type-I" error (the false positive rate). More specifically ".05" means, here, that the authors accept as significant only the statistical comparisons that they are confident have less than a 5% false positive rate.

Comment [A81]: This "Statistical Analysis" section details how the data were prepared before testing for relationships between fMRI signal, behavioral measures, and recommended sentence values. Such testing was accomplished via correlation analysis. (See comments 43 and 56.)

Comment [A82]: This section details the precise statistical test – a Spearman correlation, signified by the greek letter *rho* – used to examine relationships between fMRI signal, behavioral measures, and recommended sentence values. It also gives a rationale for using this test, as opposed to other common tests of correlation (like Pearson's).

Comment [A83]: Pearson's test – which uses the letter "r" to denote the value of its statistic – is another way to test for relationships between variables. It is more vulnerable to extreme values than the Spearman test.

fMRI Data Acquisition

Comment [A84]: This section of the paper describes the specific parameters that determine the characteristics of the signal that will be acquired from the brain using the fMRI scanner. Those parameters include such details as the number, thickness, and orientation of the brain slices from which this signal will be acquired.

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

Comment [A85]: 3D (three-dimensional). This refers to an image of brain structure, also known as *T1-weighted*.

Comment [A86]: Magnet strength is measured in units of Tesla (T), after Nicola Tesla (a prolific inventor and electrical engineer). For comparison, the Earth's magnetic field strength is around one twenty-thousandth of one Tesla.

Comment [A87]: Manual responses are typically recorded because speaking often moves the head (and hence brain) in subtle ways, interfering with the fMRI signal. Additionally, MRI scanners are very loud, making it impractical to record speech.

Cambridge MA). Functional (T2* weighted) images were acquired using a gradient-echo echoplanar imaging (EPI) pulse sequence with the following parameters: TR 2000 ms, TE 25 ms, flip angle 70°, FOV 220x220mm, 128x128 matrix with 34 axial slices (3 mm, 0.3 mm gap) oriented parallel to the gyrus rectus. These image parameters produced good T2* signal across the brain except in ventromedial frontal cortex, where some signal dropout was evident in all subjects (Brodmann area 11).

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

Comment [A88]: T2* (pronounced "tee-two-star") is a biophysical parameter. It describes an atomic phenomenon that is strongly influenced by the local physiological conditions within a small area of brain tissue. It is affected by the local magnetic environment. Changes in T2* between conditions allow the experimenter to contrast brain activity between conditions. With fMRI, what is really being measured is the biophysical phenomenon known as T2* relaxation. That is, fMRI detects differences in T2* relaxation between oxygenated blood (decreases T2*, increases fMRI signal) and deoxygenated blood (increases T2*, decreases fMRI signal).

Comment [A89]: A gradient echo pulse sequence describes the sequence of changes in the three smaller magnetic fields within the fMRI scanner. These smaller magnetic fields are used to create differences in magnetic field strength (gradients) between one end of the scanner and the other. These gradients are used in localizing brain activity. EPI is a customizable recipe that specifies the precise radiofrequency pulses to be used during the scan. It allows for the rapid collection of fMRI images.

Comment [A90]: The parameters detailed here are basically standard. These are values that the experimenter feeds to the MRI machine to tell it how the brain images are to be collected. For example, 34 axial slices, 3mm thick: this tells the scanner how the investigator wants to cut up the brain - specifically, into 3mm slices, oriented in a particular plane, 34 slices of which together comprise an entire brain image volume. The 0.33 mm gap defines the distance that separates slices, so they do not overlap. TR: specifies how long it takes the scanner to acquire an entire volume of 34 slices. TR is important because it defines the lower end of the temporal resolution of scans. For example, if it takes 2s to acquire one volume, the experimenter can't claim to detect changes that occur at a rate that is faster than 2s.

Comment [A91]: This informs the reader about how the slices were oriented, using a known structural brain feature (the gyrus rectus) as a reference.

fMRI Data Preprocessing

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

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

Comment [A92]: fMRI data are not immediately ready for analysis after being obtained from the scanner. Several image processing steps, known as "preprocessing," are required to make the images suitable for analysis (processing). This section describes those steps.

Comment [A93]: Random effects analyses allow for the generalization of results from one specific sample (e.g. the 16 subjects scanned in this study) to the population at large.

Comment [A94]: fMRI images are acquired in sequence over the course of many minutes, during which subjects often make slight head movements. These cause sequential images to become unaligned, with respect to the position of the head. Motion correction therefore reorients the images to account for slight head movements.

Comment [A95]: A single brain image (known as a "volume") is comprised of many sequentially obtained thin "slices" that are acquired over the course of anywhere from a few hundred milliseconds to several seconds (2 seconds in this study). While these slices are linked together and treated as though they were acquired at the same time, slight differences can exist between the slices due to minute differences in acquisition time. The slice time correction step "corrects" for these differences by slightly "blurring" the slices in a given volume over the total acquisition time.

Comment [A96]: fMRI signal changes that are unrelated to the experimental task can occur across the scan session. As one example, one scenario could trigger an emotionally arousing memory, which in turn elicits brain activity that is irrelevant to the task. These changes can obscure task-related signal, limiting the ability of the experimenter to detect an effect. Linear trend removal "removes" some of these task-independent signal changes.

Comment [A97]: To take into account anatomical differences between subjects that still remain after warping to a common space (see comment 29) each brain image volume is slightly blurred, a process known as smoothing. This prevents brain differences that are not meaningful from being interpreted as if they are.

Comment [A98]: This is a description of the degree of smoothing applied to the images. For fMRI, this typically ranges from 0-8 mm.

Comment [A99]: In a typical fMRI scan session, two types of images are acquired. "Structural" images provide a high resolution picture of brain anatomy, but give no information about brain activation over time. "Functional" images (also referred to as "T2*" or "epi" images due to technical details of how these images are acquired) provide a very good picture of changes in brain activation over time, but give very poor information about brain anatomy. *Coregistration* describes a process whereby a subject's functional and structural images are aligned or "registered" together. This aids the process of spatial normalization to a common space.

Responsibility Analysis

¶38 This analysis was performed to isolate brain regions that were sensitive to responsibility during punishment assessment. Signal values for each fMRI run were transformed into Z-scores representing a change from the signal mean for that run and corrected for serial autocorrelations. Design matrices for each run were constructed by convolving a model hemodynamic response function (double gamma, consisting of a positive γ function and a small, negative γ function reflecting the BOLD undershoot – SPM2, http://www.fil.ion.ucl.ac.uk/spm) with regressors specifying volumes acquired during the entire trial (stimulus onset to stimulus offset) for a given condition. These were entered into a general linear model with separate regressors created for each condition per subject (random effects analysis). We then contrasted the beta-weights of regressors using a t-test between conditions to create a statistical parametric map (SPM) showing voxels that demonstrated significantly increased activation in the Responsibility condition compared to the Diminished-Responsibility condition. Predictors for the No-Crime condition were weighted with a zero (i.e. not explicitly modeled). We applied a False-Discovery Rate (FDR) threshold of $q < .05$ (with $(c(V) = ln(V) + E)$) to correct for multiple comparisons. Only activations surviving this corrected threshold are reported.

Comment [A100]: The material in this paragraph describes how the investigators compared brain imaging volumes between the experimental conditions to create a statistical map of fMRI signal differences.

Comment [A101]: In this context, each scenario for which subjects determine punishment is one experimental "trial." 10 trials comprised one fMRI "run" of approximately 7 minutes. 5 runs (50 trials) comprised one complete experiment.

Comment [A102]: *"Statistical parametric map"* (SPM) is the more precise term for a brain image in fMRI. "Pictures" of brain images, resulting from fMRI studies, are not akin to direct, photographic snapshots of brain activity. They are instead generated by a computer, using parameters defined by the experimenters to perform statistical comparisons of measured fMRI signal between experimental conditions. These statistical comparisons are performed in every single "voxel" in the brain. A "voxel" (a contraction of "volume element") is the smallest unit of resolvable measurement of fMRI signal. Voxel size is determined by the investigator and programmed into the fMRI scanner at the start of each experiment. In the current study, voxels were 3mm (on a side) cubes - a typical size for fMRI studies. Thus, investigators divide their subjects' brains into tens of thousands of these voxels for the purposes of localizing changes in brain activity between conditions. This means that for each subject, tens of thousands of statistical comparisons were performed – one statistical test for each of these 3mm cubes. The colors in a "brain image" represent the value of the statistical test in each voxel, with brighter colors usually meaning higher statistical values, and thus a greater difference in brain activity between two conditions. Our brain map is thus really a statistical map or – more accurately – a statistical parameter map. In this case, the parameter referred to is a t-statistic, as the investigators are using t-tests to compare activity between conditions.

Comment [A103]: As stated above (see comment 102) investigators perform tens of thousands of statistical tests. As voxel sizes are quite small, the brain is thus divided into many many voxels. Say, for example, that a brain is divided into 60,000 such voxels: that means that 60,000 separate statistical tests will be performed. If an investigator sets the maximum probability of false positives error to 5%, and 5% of 60,000 is 3000, with 60,000 statistical tests, and a p-value of $p < 0.05$, that means that an investigator could potentially have an "activation" of 3000 voxels due to chance alone. 3000 voxels, at a voxel size of 3mm, is very large, meaning that "activation" in an entire brain region could be a false positive. This problem is referred to as the "multiple comparisons" or "multiple testing" problem, and statistical corrections have been devised to account for this issue. The "False Discovery Rate" (FDR) approach is one such correction, and has been implemented in the current study.

Comment [A104]: Surviving a corrected threshold, in this context, means meeting the criteria for rejecting the null hypothesis – i.e., that there are no between-condition differences in brain activation within a given region – even after invoking the correction for multiple comparisons described above (comment 103).

¶39 Volumes of interest (VOIs) were created from the suprathreshold clusters isolated in the above SPM at the conservative FDR threshold. The boundary of these VOIs were drawn from SPMs thresholded using a less conservative implementation of FDR ($q < .05$, $c(V) = 1$). The signal for each trial (event) included the time course from two TRs (four seconds) before stimulus onset to 13 TRs (26 seconds) after. Each event's signal was transformed to a percent-signal change (PSC) relative to the average of the first three TRs (0-4 seconds before stimulus onset). Event-related averages (ERAs) were created by averaging these PSC-adjusted event signals; separate ERAs were created for each combination of VOI, condition, and subject. These ERAs were then averaged across subjects for display purposes.

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

Arousal- and Reaction-Time Equated Analyses

¶41 To determine whether activation differences between the Responsibility and Diminished-Responsibility conditions were driven by punishment assessment rather than any differences in arousal, these two conditions were compared after equating for arousal ratings. This was accomplished by deleting the six trials with the highest arousal ratings from the Responsibility

Comment [A105]: In this context, VOI (volume of interest) and ROI (region of interest) both refer to regions of the brain within a statistical parametric map that the experimenters have selected for further, more detailed analysis.

Comment [A106]: "Clusters" are contiguous activated voxels. Suprathreshold clusters are clusters that survive a given corrected threshold (see comment 104).

Comment [A107]: This sentence describes the fact that the investigators have selected as VOIs, clusters that survive a certain kind of correction for multiple comparisons.

Comment [A108]: This is a method for quantifying the magnitude of experimentally-induced BOLD signal increase. It is the percentage of change in BOLD signal, between the two experimental conditions, to which researchers attend.

Comment [A109]: BOLD signal changes during each trial are averaged across each trial type (e.g., here, in Responsibility trials) for each suprathreshold cluster for each subject. These per-subject averages are then averaged across subjects.

Comment [A110]: There is a lag between changes in brain activity and changes in BOLD signal. BOLD signal (see comments 25 and 31) takes about 4-6 seconds to reach its maximum following brain activity.

Comment [A111]: Brain activity that occurs around the time a punishment decision is made.

Comment [A112]: This describes how the investigators defined the BOLD signal peak – the precise time of the maximum increase in BOLD signal from baseline.

condition for each subject. Time courses were extracted and peak differences were compared as above.

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

Comment [A113]: This term of art refers to the process of "extracting" the underlying statistical information from brain images. This is often useful for performing more in depth statistical analyses.

Dissociation of activation peak and deactivation dip

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

Comment [A114]: SPSS is a widely used software application for performing statistical comparisons.

Laterality Analyses

¶44 To confirm the lateral specificity of Responsibility-related activation in right DLPFC, we extracted BOLD signal from the corresponding left DLPFC volume of interest (i.e. 'x-mirrored' VOI, centered on talairach coordinate -39, 37, 22). We performed a two-way ANOVA with "Condition" (Responsibility, Diminished-Responsibility and No-Crime) and "Side" (Left and Right) as independent variables and BOLD signal as the dependent variable. Post-hoc comparisons between conditions in each hemisphere, and between hemispheres for the Responsibility condition, were performed using paired *t*-tests.

Comment [A115]: The same brain region as identified on the right, except on the left side (mirrored on the x-axis in the Talairach coordinate system, which separates left from right).

Punishment Rating Analysis

¶45 To identify brain regions that tracked the degree of punishment subjects assigned to a scenario, we performed a median split for punishment scores given during Responsibility scenarios. Based on the median punishment value for each

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

¶46 VOIs were created as described for the Responsibility analysis. The extracted peak activation values were used for a correlation analysis between punishment rating and BOLD response. Specifically, for each of the 20 Responsibility scenarios, the peak amplitude of the group-averaged ERA was computed, and the resulting value was correlated with the corresponding group-averaged punishment rating for that scenario. These peak values were also used in the between-condition difference score analyses.

ACKNOWLEDGMENTS

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

FIGURES

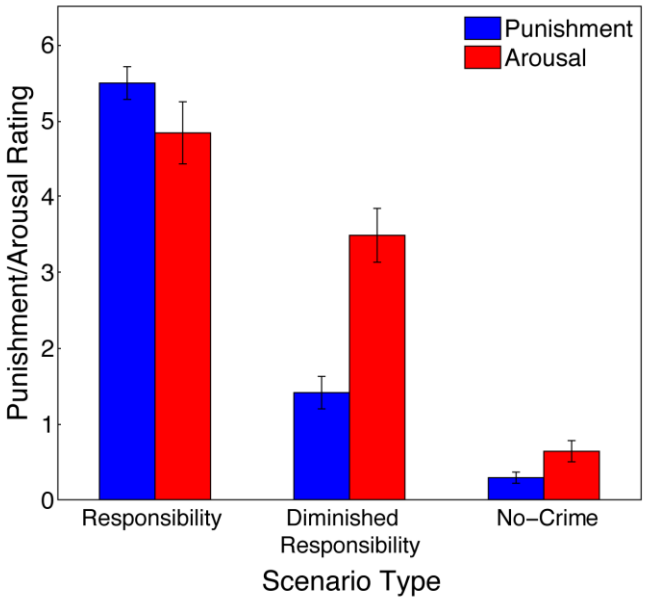


Figure 1. Punishment and arousal ratings for each scenario type.

While punishment and arousal scores were similar in the Responsibility condition, punishment scores were significantly lower than arousal scores in the Diminished-Responsibility condition.

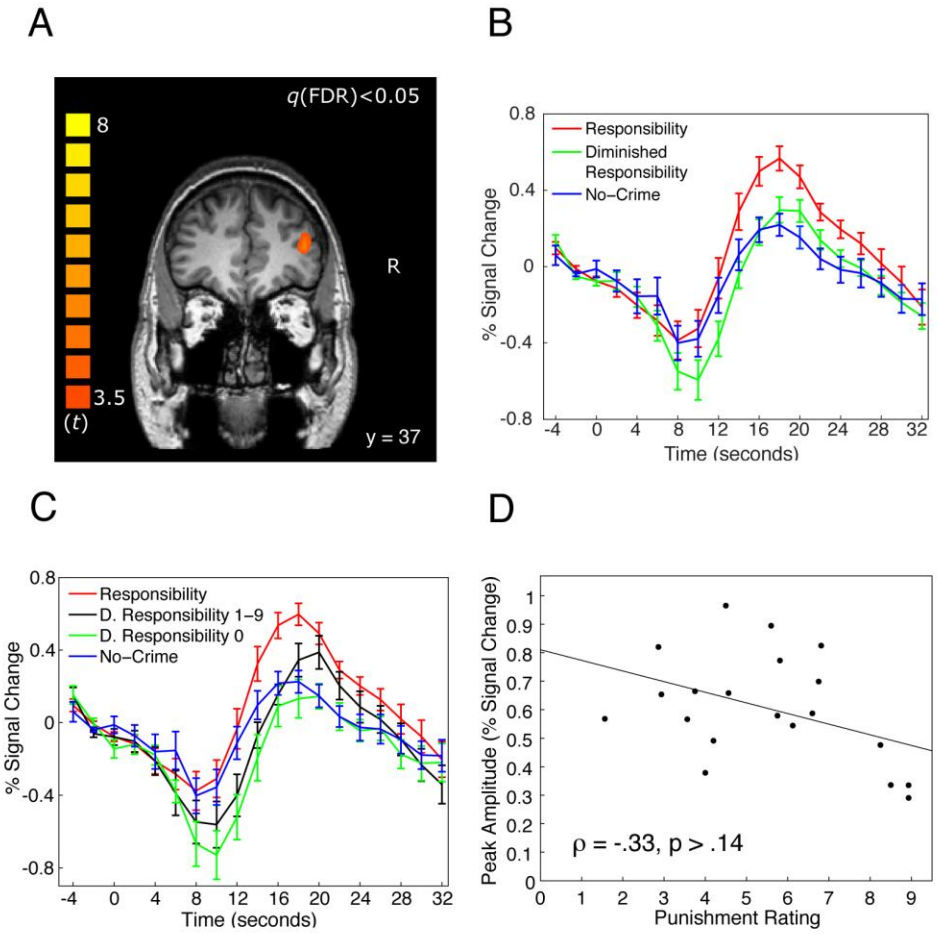


Figure 2. Relationship between responsibility assessment and right DLPFC activity.

A) SPM displaying the right DLPFC VOI (rendered on a single subject T1-weighted image), based on the contrast of BOLD activity in the Responsibility condition compared to the Diminished-Responsibility condition, $t(15) > 3.5$, $q < 0.05$, random effects analysis. R = Right Hemisphere. **B)** BOLD activity time courses in right DLPFC for the Responsibility, Diminished-Responsibility and No-Crime conditions. BOLD peak amplitude was significantly greater in the Responsibility condition compared to both the Diminished-Responsibility and No-Crime conditions ($p = 0.002$, $p = .0004$, respectively). Peak was defined as the single TR with maximal signal change from baseline within the first 13 volumes after scenario presentation onset. *t*-tests were performed on these peak volumes, which were defined separately for each condition and each subject. **C)** BOLD activity time courses in right DLPFC for Responsibility, "non-punished" Diminished-Responsibility (Diminished-Responsibility 0), "punished" Responsibility (Diminished-Responsibility 1-9) and No-Crime scenarios. BOLD peak amplitude was significantly greater in "punished" compared to "non-punished" Diminished-Responsibility scenarios ($p = 0.04$), while no difference was observed between "non-punished" Diminished-Responsibility and No-Crime scenarios ($p = 0.98$). **D)** Relationship between BOLD peak amplitude in right DLPFC and punishment ratings in the Responsibility condition. These two variables were not significantly correlated ($p > 0.15$).

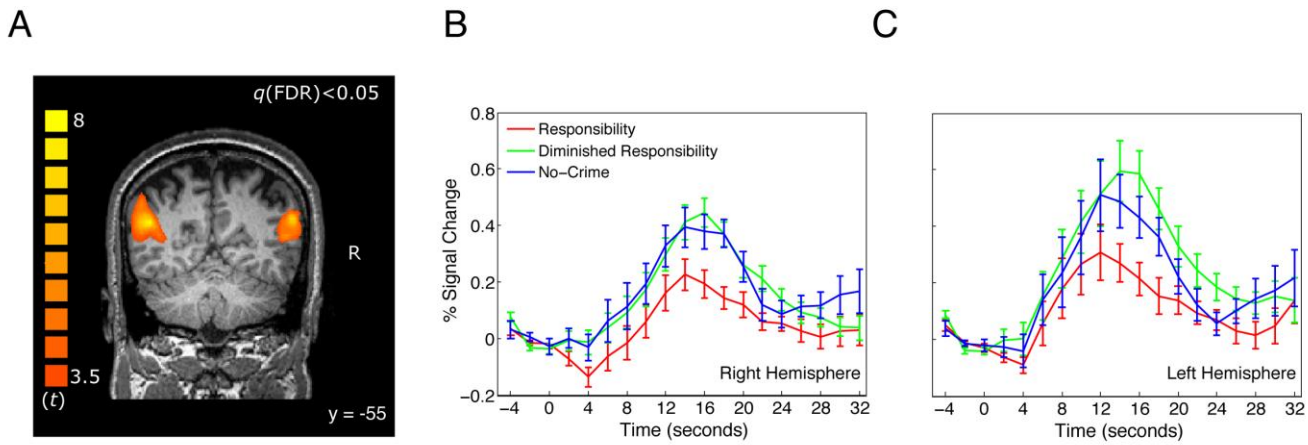


Figure 3. Relationship between responsibility assessment and bilateral temporo-parietal junction (TPJ) activity.

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

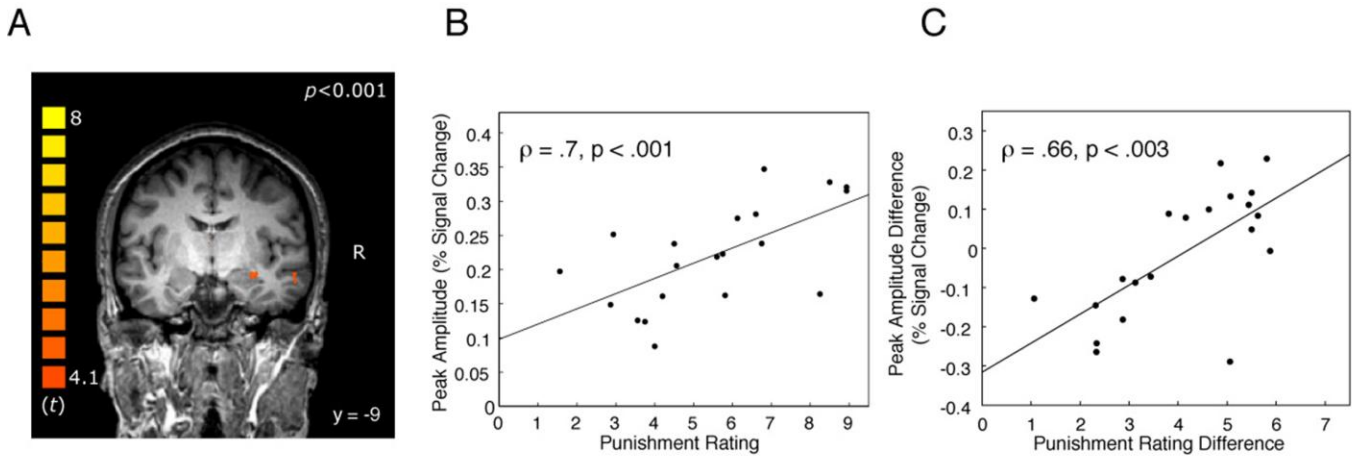


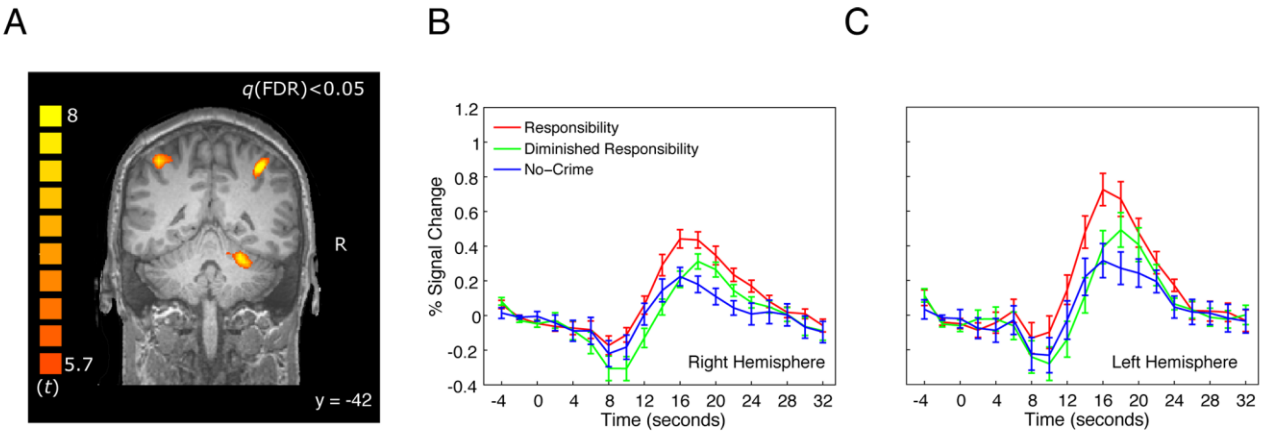
Figure 4. Relationship between punishment and right amygdala activity.

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

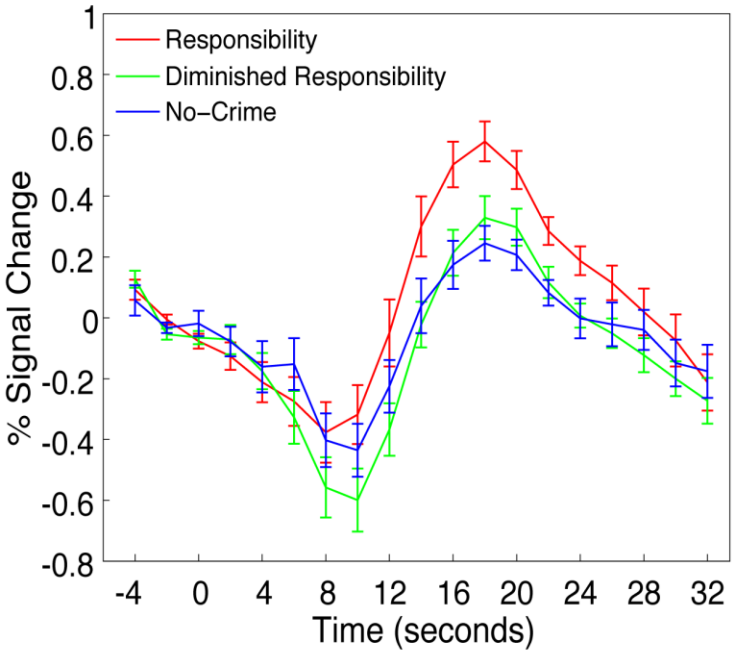
B) Relationship between BOLD peak amplitude in the right amygdala and punishment ratings in the Responsibility condition. These two variables were significantly positively correlated (p = 0.001).

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

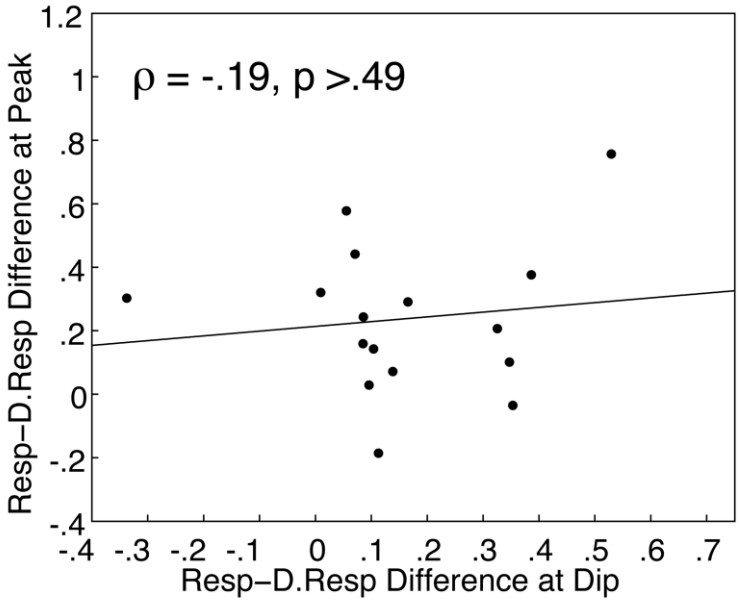
SUPPLEMENTARY FIGURES



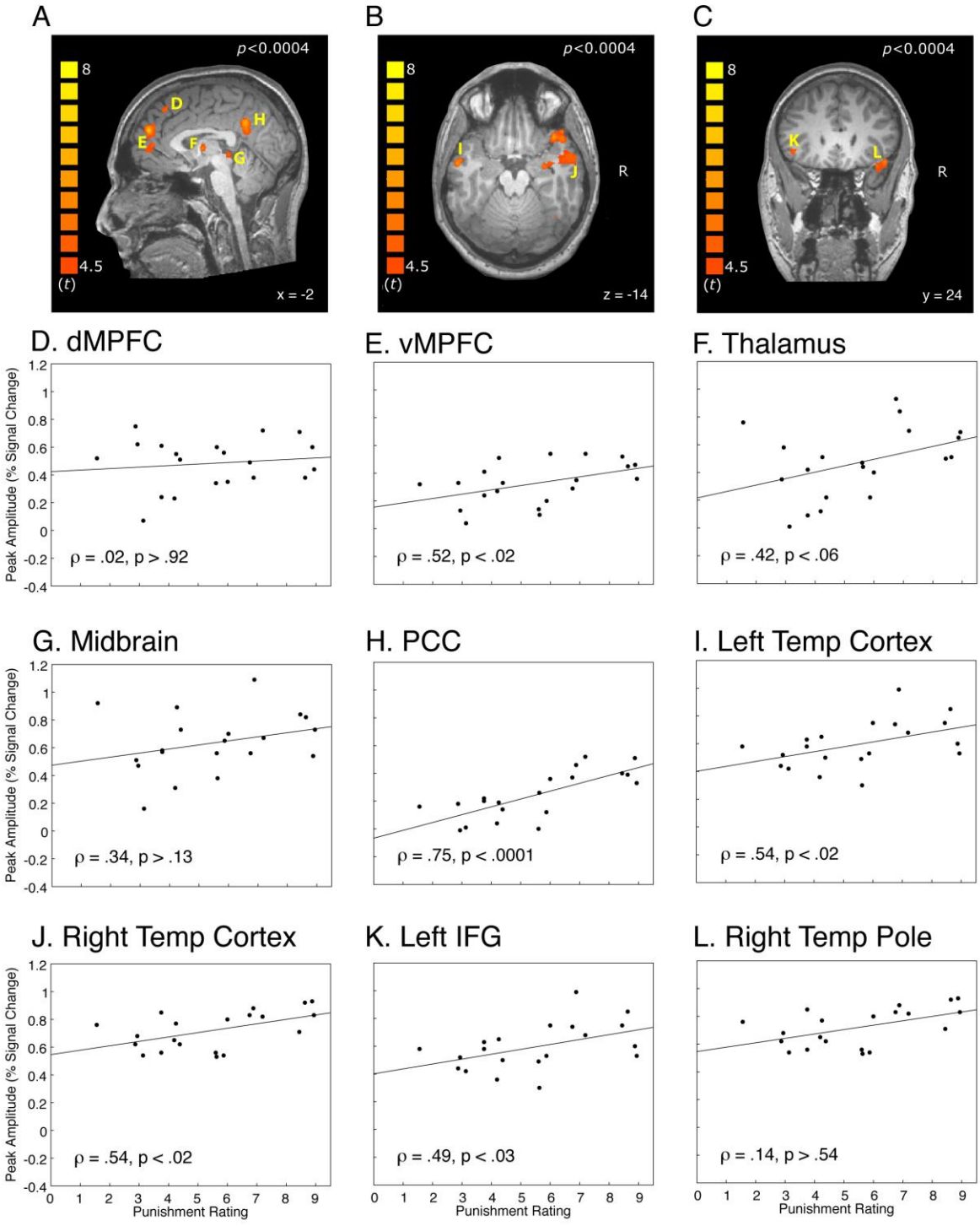
Supplementary Figure 1.



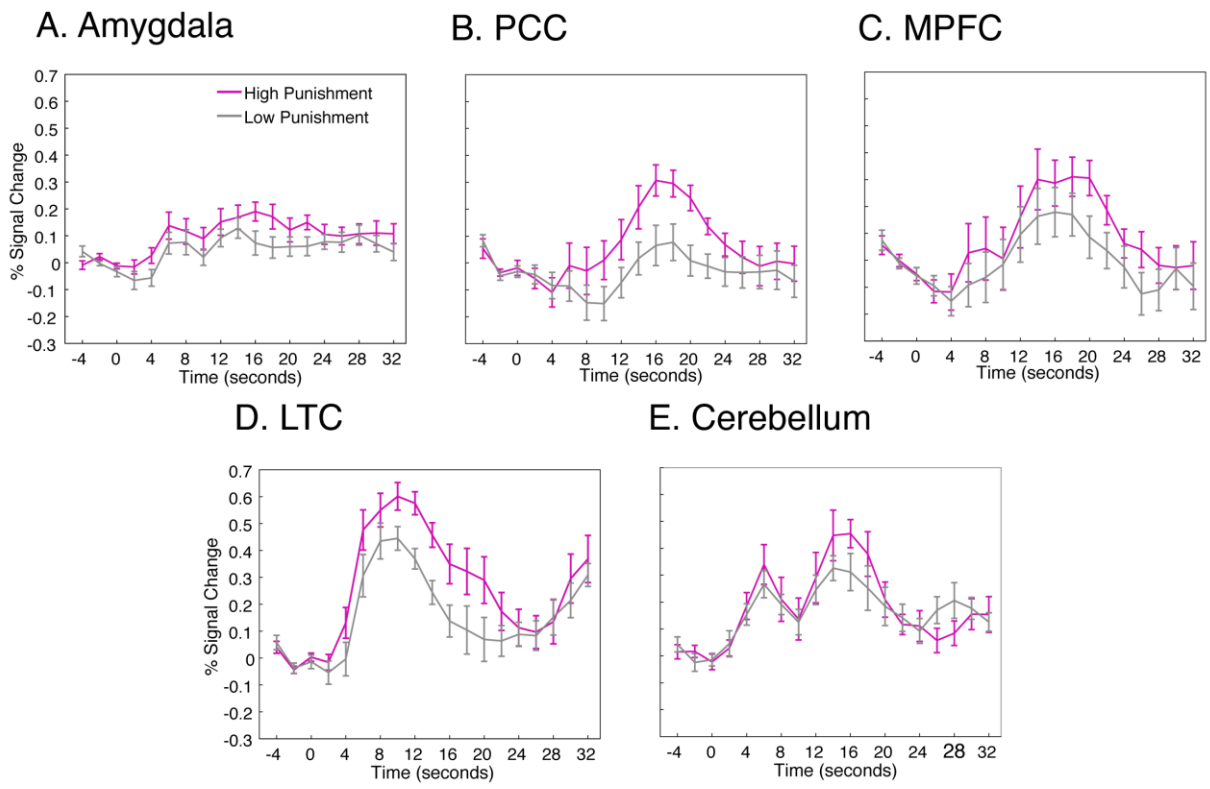
Supplementary Figure 2.



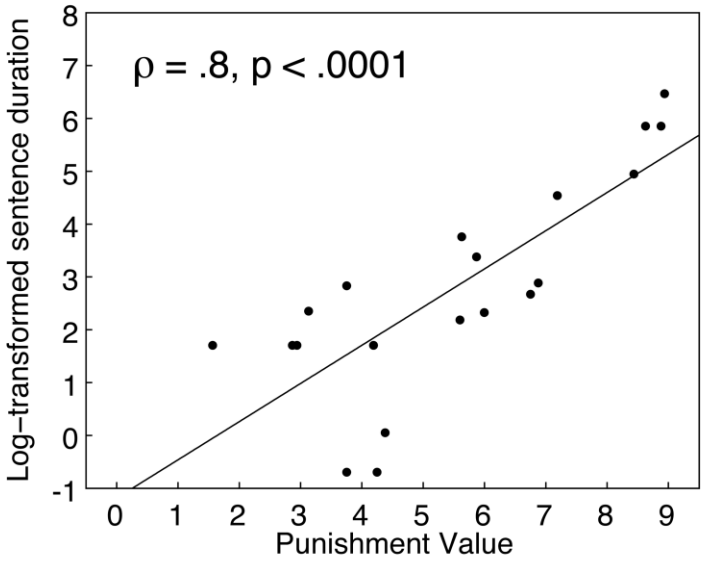
Supplementary Figure 3.



Supplementary Figure 4.



Supplementary Figure 5.



Supplementary Figure 6.