FIAT POENA RUAT IUSTITIAM? INVESTIGATING ENDORSEMENT OF RETRIBUTION AND ITS ROLE IN *MENS REA* ATTRIBUTION

by

Roland Charles Nadler

A.B., Harvard University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies

(Interdisciplinary Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2012

Abstract

In the past decade, a proliferation of discussion at the intersection of law and neuroscience has highlighted the significance of public attitudes toward punishment, with claims frequently made regarding the popularity or prospects of retributivism – the position in punishment theory that privileges desert as the basis of punitive action. However, no well-validated instrument for measuring endorsement of retribution has been available to ground the discussion in empirical data, and little attention has been paid to the possibility that an individual's views on retributivism may interact with judgments about intention and knowledge underwriting the imputation of *mens rea* ("guilty mind"). In Chapter 1, I construct and validate a new Endorsement of Retribution scale. In Chapter 2, I detail the design and results of a study that employs the new scale, investigating the relationship between participants' Endorsement of Retribution scores and their likelihood of judging that a hypothetical defendant's actions met a specific standard of guilt. The data from this study provided no support for the hypothesis that Endorsement of Retribution score is associated with an increased tendency to convict for legally irrelevant reasons. Moreover, highly retributive respondents were no more likely than other respondents to vote guilty for any reason, legally relevant or not. However, respondents were vastly more likely to convict an ostensibly nefarious character than an ostensibly morally upstanding one, regardless of retributive inclinations. These results highlight a previously acknowledged need to address the problem of cognitive biases in the reasoning that jurors are called upon to perform, while also serving as a reminder that the causal roots of such biases defy simple single-factor explanations, and partly dispelling the worry that attitudes about punishment constitute a major contributor in this regard.

Preface

All of the studies in this thesis involving human subjects received approval from the University of British Columbia's Behavioural Research Ethics Board. The Certificate Number for the Ethics Certificate obtained was H11-02821. This Certificate covers the study detailed in Chapter 1, and was extended to cover the studies detailed in Chapter 2 via the Post-Approval Amendment system.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Scores of colleagues, mentors, friends, family, and expert correspondents have made contributions to the creation of this thesis, ranging in scope from single points of information to extensive shaping of the research and writing. What follows is a necessarily partial listing of credits and gratitude. My apologies to anyone I have omitted, either unwittingly or for the sake of brevity.

- My advisor, supervisor, committee chair, professor, and PI, Dr. Peter Reiner, for countless hours of discussion, guidance, feedback, critique, strategizing, and (not least) motivational prodding.
- Dr. Judy Illes, the other half of my supervisory committee, for cultivating a marvelous garden of opportunity in the National Core for Neuroethics, and for fitting thesis committee duties into a jam-packed schedule.
- The R. Howard Webster Foundation and the Province of British Columbia, for endowing the R. Howard Webster Fellowship, which supported my graduate studies and thesis.
- The University of British Columbia, for the International Partial Tuition Scholarship.
- Several statistical consultants - Aline Tabet in the UBC Department of Statistics, Nick Fishbane from UBC Statistics' Short Term Consultancy Service, and my dear friend and former roommate Daniel Boada - without whose expertise my analytic techniques would have proven less than adequate.
- The fourteen law professors who helped establish face validity for the retributivism scale by taking the time to provide feedback on my preliminary scale items, with special gratitude to those who generously wrote back with comments.

- Jasmine Carey and Taylor Davis (and by extension, Del Paulhus) at UBC, for their generous assistance in navigating the ins and outs of scale construction.
- My parents, for providing me with unending support at every turn - and for helpful edits.
- The excellent people running the Interdisciplinary Studies Graduate Program, for creating an extraordinary platform for students to engage in cross-disciplinary work unhindered by departmental delineations, and for scholarship support.
- My colleagues and co-workers at the National Core for Neuroethics, for being such a stimulating and brilliant group and for providing useful feedback on presentations as this thesis developed.
- My colleagues and fellow residents at Green College, for making my home life as intellectually rich as my work life.
- Prof. Steve Wexler in the UBC Faculty of Law, for teaching - and allowing me to join in on, despite not being a law student - the thought-provoking seminar course that inspired me to choose this topic for my thesis.
- The many superb scholars from whose work this project is descended: Owen Jones, Francis Shen, Morris Hoffman, Rene Marois, Joshua Greene, Elizabeth Loftus, Fiery Cushman, Joshua Knobe, Thomas Nadelhoffer, Shaun Nichols, Adina Roskies, Jonathan Haidt, Hank Greely, Neil Levy, Bertram Malle, and Sarah Nelson, among many others.
- Last but far from least, the many hundreds of Amazon Mechanical Turk-recruited respondents who participated in the studies upon which this thesis reports, for their patience and interest.

Introduction

"*Distrust all in whom the impulse to punish is powerful.*" – Friedrich Nietzsche

The past decade has seen a profusion of scholarly work interrogating the conceptual underpinnings of punishment as the practice manifests in the legal and criminal justice system (e.g., Dolinko, 2003; Rubin, 2003; Huigens, 2005; Dingwall, 2008). A significant fraction of this work stems from a burgeoning interest in issues at the crossroads of law and cognitive science, with Joshua Greene and Jonathan Cohen's 2004 paper "For the Law, Neuroscience Changes Nothing and Everything" standing out as a notable catalyst of discussion for its provocative claim that the percolation of neuroscientific knowledge into society at large will reshape common perceptions of punishment and its proper role (Greene & Cohen, 2004).

Alongside such topics as the impact of neuroimages as courtroom evidence (Schweitzer, Saks, Murphy, Roskies, Sinnott-Armstrong, & Gaudet, 2011) and the applicability of neuroscientific investigative methods to familiar problems in juror psychology (Mobbs, Lau, Jones, & Frith, 2007), the question of whether and how a brain-based understanding of behaviour will – or should – reshape notions of criminal responsibility, desert, and punishment theory has become part of the stock-in-trade for neurolaw.

Frequently, commentators in neurolaw frame this topic in terms of what cognitive science will mean for the particular position in punishment theory known as retributivism[1] (e.g., Gazzaniga, 2008; Erickson, 2009; Buller, 2010). Russ Shafer-Landau characterizes retributivism as the notion that "the point of legal punishment … is that the guilty be given their just deserts" (Shafer-Landau, 2000, p. 189); Bagaric and Amarasekara similarly identify as a necessary condition that "all retributive theories assert that offenders deserve to suffer and that the institution of punishment should inflict the suffering they deserve …" (Bagaric & Amarasekara, 2000, p. 127). Retributivism is frequently articulated in contradistinction to models of justice that are broadly consequentialist – the latter focusing variously upon utilitarian, restorative, deterrence-based, or rehabilitative ends (e.g., Strauss, 2001; Rubin, 2003; Whitman, 2003).

The specific challenge posed by the brain sciences to retributivism is summarized by O. Carter Snead, with heavy citation of Greene and Cohen 2004, as follows:

> Greene and Cohen argue that advances in cognitive neuroscience—enabled by neuroimaging—will ultimately demonstrate that "ordinary conceptions of human action and responsibility" are false. "[A]s a result, the legal principles we have devised to reflect these conceptions may be flawed" and must be radically overhauled and replaced with principles that are grounded in a neuroscientific view of the truth about free will and human agency. The primary focus of their critique is

[1] Despite my choice to highlight the papers that discuss this topic in terms of retributivism, it must be noted that many do not employ this terminology, instead simply speaking of criminal or legal "responsibility" (e.g., Gilbert 2004, Farah 2005, Roskies 2006). I avoid dwelling on this terminological inconsistency primarily because it seems to me that many papers written about neuroscience and criminal responsibility are essentially using "responsibility" as a proxy for desert, and hence may be treated as discussions of retributivism without loss of accuracy.

> the principle of retributive justice—which, they assert, "depends on an intuitive, libertarian notion of free will that is undermined by science." (Snead, 2010, p. 8)

Snead goes on to note that Greene and Cohen's claim regarding the philosophical tenability of retributivism comes alongside a factual claim regarding public sentiment:

> Greene and Cohen argue that when and if the notion of human agency is shown to be illusory, societal attitudes may well change … once society internalizes the lessons of cognitive neuroscience as they bear on moral (and thus criminal) responsibility, retribution—relying as it does on a false understanding of human agency—will be eliminated as a legitimate general or distributive justification for punishment. (Snead, 2010, p. 10-11)

The tension between retributivism and neuroscience has produced a small library's worth of academic back-and-forth, with much (though not all) of it stemming from Greene and Cohen's philosophical-cum-neuroscientific broadside and accompanying sociological prediction. This state of affairs serves as the point of departure for the two projects chronicled in this thesis.

Background for Chapter 1

The first of these two projects begins with the observation that the topic in question has largely been addressed via the clash of arguments in the empyrean of pure theory. Considering the philosophical nature of the subject matter, this is hardly surprising or objectionable. However, few debates can play out entirely without making reference to how the world actually is; as anyone with a healthy sense of empiricism will readily aver, such references are a dicey business when made from the armchair without data to ground them. Indeed, the discipline of philosophy – traditionally the armchair tradition *par*

excellence – has recently experienced an empirical upheaval in the form of "experimental philosophy," illustrating that the scientific method can be profitably applied even to the most theoretical of subjects (e.g., Nadelhoffer, 2005; Bengson, Moffett, & Wright, 2009; de Brigard, 2010).

Prior Research on Public Attitudes Toward Retribution

Relative to the level of attention that Greene and Cohen's deflationary prediction has attracted, very few scholars have taken an empirical approach to understanding public support for retribution. Even when the counter-claim – i.e., "Neuroscientific evidence about the links between brain dysfunction and criminal behaviour seems … unlikely to change our lay views about the demands of justice …" (Greely, 2008, p. 1104) – is articulated in a convincing manner, readers are left without much in the way of quantifiable truth-conditions. Examples of empirical inquiry in this vein are sparse and seldom capture exactly what I am referring to: Dominic Johnson's 2005 analysis entitled "God's punishment and public goods," for example, investigates belief in supernatural, rather than juridical, punishment (Johnson, 2005). Kevin Carlsmith's 2006 study "The roles of retribution and utility in determining punishment" notes that punishment as a behaviour is well-studied, with a focus on "on the characteristics of situations and perpetrators that lead to greater perceived guilt and punishment" (Carlsmith, 2006, p. 438); his own experiment in that paper probes "how people make punishment decisions by looking at the type of information they seek, the order in which the information is sought, and the resulting confidence that people have in the appropriateness of the assigned punishment," (*ibid.*, p. 439) none of which bears on the measurement of individuals' support for retributivism (or

anti-retributive consequentialism) *qua* substantive punishment theory. There exists a rich body of research in experimental philosophy devoted to generating empirical data on the relationship between intuitions about moral / criminal responsibility and deterministic / neuroscientific understandings of behaviour (e.g., Nahmias, 2006; Nichols & Knobe, 2007; Roskies & Nichols, 2008; and especially de Brigard, Mandelbaum, & Ripley, 2009), but these findings do not necessarily imply anything about participants' opinions on the proper rationale for punishment. Gavin Dingwall has analyzed how retributivism informs the sentencing of adult offenders (Dingwall, 2008), but the general public is usually not involved in sentencing, except occasionally indirectly via democratically enacted policy.

The closest that empirically-oriented researchers have come to shedding light directly on the question of public support for retributivist punishment theory can be found in a 1994 paper by Felicia Pratto and colleagues introducing Social Dominance Orientation (SDO) as a variable in social psychology (Pratto, Sidanius, Stallworth, & Malle, 1994). The appendix to their paper includes a listing of short scales used in experiments validating the SDO construct; one of these scales, which correlated positively with SDO, is entitled "Belief in Retribution" and includes the following five items:

- Society does not have the right to get revenge for murder.
- For a terrible crime, there should be a terrible punishment.
- Even the worst criminal should be considered for mercy.
- Those who hurt others deserve to be hurt in return.
- Punishment should fit the crime.

This scale represents a concrete first attempt to track belief in retribution as a distinct and measurable entity. However, while it undoubtedly fulfilled its role within the context of the

1994 paper, it unfortunately cannot stand on its own as a valid instrument for gauging retributivist sentiment. The scale items are too general and too few, and no independent validation procedures appear to have been undertaken. Considering the need for such an instrument as outlined in the previous paragraphs, one might conclude that a scale doing the same work as Pratto et al.'s deserves dedicated treatment in a project all its own. In Chapter 1 of this thesis, I report on my efforts to construct and validate a 14-item Endorsement of Retribution scale that delivers on this idea.

Which Retributivism Should the Scale Assess?

The task of Chapter 1 is complicated by the fact that several distinguishable versions of retributivism have been staked out in the literature. Moreover, it is easy for articulations of retributivism to run together differentiable versions of the theory, each distinguishable from the others with sufficient attention to detail.

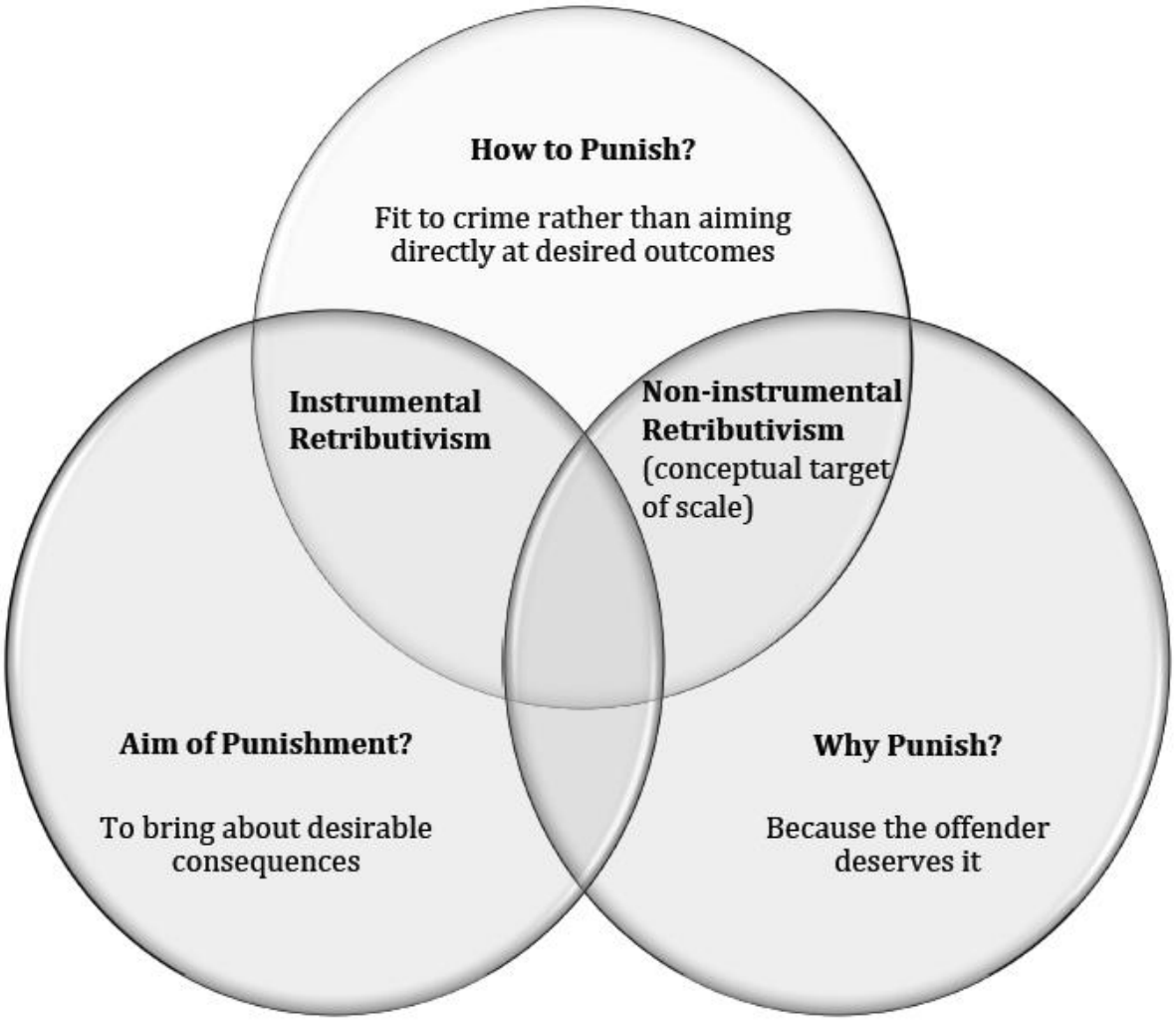

Figure 1. Concept mapping for questions answered by retributivism.

The most relevant variation in the specifics of retributivist theory pertains to the relationship between retributivism and consequentialism. Many commentators (e.g., Carlsmith, 2006; Bronsteen, 2007; Johnson, 2008) frame the pursuit of retribution and the promotion of desirable outcomes as mutually exclusive; however, some theorists (Robinson & Darley, 1996; Huigens, 2005; Robinson, 2008) argue that desert-reflecting punishment is a crucial and indispensable feature of a consequentially optimized criminal justice system. This is a view I label instrumental retributivism. Instrumental retributivism is consequentialist in its justificatory texture, but differs from more explicitly anti-

retributive consequentialist positions in terms of a key factual prediction: namely, that the best outcomes are reliably brought about by punishing offenders as though the goal and justification of punishment is the delivery of just deserts. The primary exponents of instrumental retributivism – Robinson and Darley in their 1996 paper "The Utility of Desert" – explain that desert-reflecting punishment sustains public perceptions of the moral legitimacy of the criminal justice system, without which everyone would be far worse off.

By contrast, non-instrumental retributivism holds that criminal offenders are properly subject to punishment only because the commission of their misdeed leaves them deserving of retaliation[2]. According to the non-instrumental retributivist, the type and severity of punishment is to be reckoned solely by reference to the type and severity of the misdeed that occasions the punishment. The downstream consequences of such retaliation, if they are to be considered whatsoever, may only figure in secondarily and do not constitute a justification to deliver anything less (or more) than the deserved punishment[3].

[2] For further explication of this view, see Bagaric & Amarasekara 2000; there, the authors refer to it as "intrinsic retributivism."

[3] A third variation on retributivism, less relevant to this discussion, is limiting retributivism, identified by commentator Richard Frase as "the consensus model of criminal punishment" (Frase, 2004). Limiting retributivism is a compromise position that relies on a desert-derived notion of proportionality to set the eponymous limits within which punishment for a given crime may fall, but within these limits permits consequentialist goals to influence the fine-tuning of a sentence. Because limiting retributivism is a hybrid model that borrows its tools from two ideologically pure positions, the Endorsement of Retribution scale does not deal with it directly.

The conceptual target of the Endorsement of Retribution scale whose construction and validation are detailed in the following pages is specifically non-instrumental retributivism[4], to the careful exclusion of instrumental retributivism.

Background for Chapter 2

Chapter 2 of this thesis covers the application of the Endorsement of Retributivism Scale to an investigation of the interaction of retributive sentiments and perceptions of criminal guilt. The point of departure for such an inquiry is ultimately tied to the aforementioned debates over the implications of cognitive science for criminal responsibility and punishment theory. Stephen Morse – here summarized by Tom Buller – has argued:

> … the law is not interested in free will in any deep "metaphysical" sense—whether we are "genuinely" or "ultimately" free. Rather, the law assumes that we have free will, and that, in our day-to-day lives, our actions are intentional and voluntary. (Buller, 2010, p. 197)

This attitude resonates with the lessons of my own personal experience in discussing this topic with legal scholars: the most frequent rejoinder I encounter is that "the law does not work that way" – that only someone ignorant of how the law actually functions could believe that Greene and Cohen have raised a genuine problem for the legal system. The law,

[4] By way of contrast, I note the "Comprehensive Justice Scale" under development by Jonathan Haidt, John Darley, and Dena Gromet (viewable at http://www.yourmorals.org/justice.php with details at http://www.yourmorals.org/justice_process.php). Although it may appear quite similar to my Endorsement of Retribution scale at first glance, the Comprehensive Justice Scale was designed with much broader and more ambitious goals, aiming to take stock of "the broadest possible range of ideas and intuitions about what you think should happen to the offender, and the victim" (*ibid.*, at URL) As such, this scale can serve as a complement to Haidt et al.'s, and should not be understood as a direct competitor.

it is alleged, simply does not care about what neuroscience is telling us about human agency – at least not in any deep or fundamental way.

I will refrain from commenting here on whether this represents a compelling argument, instead merely noting that, in my view, the continuation of this debate in neurolaw holds considerable merit, whatever the outcome. Rather than any kind of direct challenge, my response to the dismissive impulse of Morse and others is to propose that inquiry should begin, instead, with the things the law indisputably does care about, and examine how retributivism and its competitors serve those ends.

It seems uncontroversial to suggest, for example, that the law has a crucial stake in promoting accurate and conceptually self-consistent attributions of guilty mental states; at the very least, there would be a substantial *de facto* burden of argumentative proof on anyone wishing to argue the opposite. Researchers in philosophy and psychology who have studied guilty mental state attribution have explicitly recognized this *desideratum*, as illustrated by a 2003 publication by Bertram Malle and Sarah Nelson identifying "the valid and precise use of the concepts of mental states in reasoning about the defendant's actions and in assigning responsibility, blame, and punishment" as a central challenge in creating a system of criminal adjudication (Malle & Nelson, 2003, p. 564). Similarly, the project detailed in "Sorting Guilty Minds" – an empirical juror psychology study of formidable scale undertaken by Francis Shen and colleagues (Shen, Jones, Hoffman, Greene, & Marois, 2011) – is largely premised on the notion that it is, indeed, a highly undesirable state of affairs when defendants are inconsistently classified into the categories of "knowing" or "reckless" guilt – carrying, as it sometimes does, a sentencing difference of ten years in prison.

The motivation for my own study, then, is as follows: even if retributivism truly is the right punishment theory in terms of offering correct philosophical backing for commonly accepted practices, it would prove, at a minimum, worthwhile to investigate whether (as a matter of psychological fact) an affinity for retribution interferes with the attempts of the legal system to realize the goals extolled above.

Why might one think that such interference is possible or likely? To rehearse the argument in short, I mean to suggest the possibility that a preference for retribution may create an interference effect as an individual goes through the process of inferring whether a defendant exhibited a culpable mental state. The suspicion driving this suggestion is as follows: concepts such as intention, knowledge, luck, and causation are both components of criminal guilt (about which jurors must often make inferences) and concepts whose attribution is subject to cognitive bias. Such an interference effect, if it exists, could – at best – make jurors more sensitive to the components of criminal guilt, or – at worst – introduce a prejudice-driven bias into their inference-making, thereby creating unjustifiable inconsistencies in conviction outcomes.

The Conceptual Structure of *Mens Rea*

Formally, guilty mental states manifest in the law under the conceptual rubric of *mens rea* - Latin for "guilty mind" or "wicked mind." The conceptual structure of *mens rea* is such that, in large part, it is an amalgam of several more intuitively accessible concepts, most prominently intention, causation, knowledge, and (arguably) luck. These concepts are all philosophically fraught – indeed, an entire subfield is devoted to pinning down the nature of knowledge – and must be made tractable for everyday legal use. This is, *inter alia*,

the task of a jurisdictional criminal code. Since the American Legal Institute drafted it in 1962, the Model Penal Code (MPC), the brainchild of Herbert Wechsler, has been adopted as the basis for the criminal code of 38 states, and its influence is even more pervasive than that number indicates. The MPC sets out, among other things, requirements for the varieties of culpable mental states that would come to embody the concept of *mens rea* in American criminal courts.

The MPC identifies exactly four culpable mental states, in order of severity such that each state is a subset of a less severe one, and requires that laws specify at least the minimum-severity mental state required for the act proscribed in the law to be treated as a crime. The four states are "purposeful," "knowing," "reckless," and "negligent." Any act that does not at least qualify as negligent cannot be considered a crime, except for particular types of actions that fall under a category called strict liability. The exact definitions for *mens rea* states in the MPC are as follows (summarized in Shen, Jones, Hoffman, Greene, & Marois, 2011, p. 10-11):

> A person acts purposefully [with respect to a result] if it is his conscious object . . . to cause such a result. A person acts knowingly [with respect to a result] if he is aware that it is practically certain that his conduct will cause such a result. A person acts recklessly [with respect to a result] when he consciously disregards a substantial and unjustifiable risk that [his conduct will cause the result]. A person acts negligently [with respect to a result] when he should be aware of a substantial and unjustifiable risk that [his conduct will cause the result].

In each *mens rea* category, one can identify some combination of: an intentional element ("conscious object … to cause," "consciously disregards"), an epistemic element ("aware that it is practically certain," "should be aware"), a causal element, or a luck/risk/certainty

element. Perhaps not surprisingly given the philosophical contestability of knowledge, intention, causation, and luck, empirical evidence indicates that "laypeople do not comprehend mental state distinctions that are differentiated in legal doctrine" (Severance, Goodman, & Loftus, 1992, p. 107) – the aforementioned "Sorting Guilty Minds" study, for instance, uncovered that even when respondents are provided with definitions of guilty mental states for reference, they nonetheless persist in miscategorizing hypothetical defendants as having acted "knowingly" when the facts of the case supported an ascription of "recklessly," and vice versa (Shen, Jones, Hoffman, Greene, & Marois, 2011).

Knobe Effects in the Concepts Comprising *Mens Rea*

A confluence of theory and evidence from other disciplines suggests that this predicament may well extend beyond the specific parameters of the above-cited studies. Returning to the purview of experimental philosophy, it may not come as a surprise to note that the most-studied and foundational topic in the subfield is the nature of intentional action (e.g., Knobe, 2003; Nadelhoffer, 2005; Malle, 2006), with studies on the conceptual structure of knowledge constituting a notable offshoot from this vein (Beebe & Buckwalter, 2010), and other entries in the experimental philosophy literature addressing causal judgments in general (Knobe & Fraser, 2010), action (as contrasted with omission) in general (Cushman, Knobe, & Sinnott-Armstrong, 2008), and luck (Malle & Nelson, 2003). The fact that there has not been extensive cross-pollination between experimental philosophers and legal theorists interested in *mens rea* – considering the extensive and readily identifiable conceptual overlap – may strike readers as odd. Admittedly, more than a few experimental philosophers – and psychologists – have engaged with the concept of

mens rea (Nadelhoffer, 2006; Guglielmo & Malle, 2010; Young & Saxe, 2011), but the interest has not been noticeably reciprocal.

Experimental philosophy studies frequently aim to empirically detect a type of phenomenon involving inconsistent concept usage that has variously been termed the Knobe Effect (after Joshua Knobe, the philosopher whose study of such effects served as a flashpoint for further research) or sometimes the side-effect effect. For my purposes here, I define a phenomenon involving inconsistent conceptual usage as a Knobe Effect if and only if it involves a change in people's willingness to impute a given factual state or quality based on apparently irrelevant influence from normative, moral, or otherwise evaluative features of the situation. More plainly, a Knobe Effect is a type of double standard driven by moral feelings of disapproval or approval, in which a person's concept usage is influenced by his or her evaluative sentiments in a logically unjustifiable manner. It follows that there are many types of Knobe Effects, since many different concepts can be skewed by morally charged considerations. The Knobe Effect will seem familiar to psychologists, who will recognize it as a close cousin (perhaps even a subtype) of motivated reasoning.

The classic example of the Knobe Effect pertains to intentional action: ordinary folk are much more likely to say that the side effect of an action was brought about intentionally when that side effect has a negative moral valence than when it has a positive moral valence. An example with relevance to criminal law comes from a Thomas Nadelhoffer's 2006 paper "Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality," a study that investigated how participants understood the actions of a driver who brought about the death of an individual hanging on to the side of his vehicle. In particular, Nadelhoffer asked if the driver caused the victim's death intentionally and / or

knowingly. He found that such judgments were significantly modulated by an apparently irrelevant detail – namely, in one vignette, the driver was a thief and the hanger-on a policeman, whereas in the other vignette, the driver was a civilian (with no further characterization) and the hanger-on a carjacker (Nadelhoffer, 2006). Participants in Nadelhoffer's experiments were significantly more likely to say that the thief brought about the policeman's death knowingly and intentionally (and more willing to say that the thief deserved blame for this) than they were to make the same ascriptions of knowledge and intention and blame in the case where it was the carjacker who died and the civilian behind the wheel. I will eventually touch upon the limitations of the Nadelhoffer study, but for now, in combination with other works cited above, it serves to illustrate the Knobe Effect.

Results like these are significant because they indicate that ordinary people attribute intentionality and knowledge – key conceptual components of *mens rea* – in an inconsistent manner across structurally identical situations. It is natural to think of intention and knowledge as necessary for underwriting any kind of moral blame, but here one can see unsettling evidence of potential circularity – of moral blame, in turn, underwriting ascriptions of intention and knowledge. The notion that "it's only wrong if you meant to do it" becomes "you only meant to do it if it was wrong (or: if you seem like a nasty individual; or: if the outcome was significantly appalling, et cetera, et cetera)." This inverts the legal ethos of placing factual determination wholly prior to evaluative judgment.

Taking into account the widespread evidence that the conceptual building blocks of *mens rea* are highly susceptible to Knobe Effects, the natural conclusion seems to be that attributions of *mens rea* itself may similarly prove susceptible to such problematic inconsistency. This hypothesis is empirically testable. Additionally, though, my study

investigates the relationship between Knobe Effects and people's endorsement of retribution. In the next section, I explain what might lead one to hypothesize that these two variables are connected.

Retribution and Blame-Validation

In brief, the reasoning behind my hypothesis is that the evaluative stance adopted by an individual engaged in retributive thinking may fuel the Knobe Effects exhibited in the 2006 Nadelhoffer study. Nadelhoffer himself aptly fleshes out this rationale by citing, as the basis for his own study design, Mark Alicke's "Culpable Control Model" of the psychology of blame. He quotes Alicke:

> When blame-validation mode is engaged, observers review structural linkage evidence in a biased manner by exaggerating the actor's volitional or causal control, by lowering their evidential standards for blame, or by seeking information to support their blame attribution. In addition to spontaneous evaluation influences, blame-validation processing is facilitated by factors such as the tendencies to over ascribe control to human agency and to confirm unfavorable expectations. (Alicke, 2000, p. 558)

The notion of a blame-validation mode, as sketched by Alicke, supplies the theoretical point of interface between retributivism and *mens rea* attribution: the idea is that, cognitively speaking, retributive thinking and blame-validation mode are intrinsically linked.

One might be inclined to think that simply replicating Nadelhoffer's 2006 methods with an additional measure of participants' endorsement of retribution would be sufficient to test this idea. However, several features of the vignettes in the Nadelhoffer study limit applicability to an explicitly legal context. Part of my goal in undertaking the studies

described in Part II of the thesis was to fine-tune the vignettes in an effort to make the results speak more directly to legal theorists interested in the design of the criminal justice system.

Finally, I supply some preemptive clarification on the intention and motivation of this project. Suppose that these experiments were to uncover a significant effect whereby an affinity for retribution exacerbates a Knobe Effect for *mens rea* attributions: how might the law appropriately respond? I do not mean to suggest that either panic or drastic revision of the legal system would make for warranted reactions. The criminal justice system cannot summarily throw out *mens rea*, and this study is not presented as empirical ammunition for a radically revisionary project. Indeed, no behavioural experiment can settle the question of whether retributivism is the correct grounding for criminal punishment *sub specie aeternitatis.* For that matter, no experiment can reveal whether the idea of a correct punishment theory "from the point of view of the universe" is itself a sensible concept or merely product of confused thinking. Any changes to the criminal justice system that this study could properly occasion would surely need to take the form of cautious, pragmatic adjustments.

Chapter 1: Constructing and Validating a New Scale Measuring Endorsement of Retribution

Rationale

This Chapter details the theoretical and empirical steps I followed to construct and validate the Endorsement of Retribution scale, which is aimed at measuring the extent to which an individual sympathizes with non-instrumental retributivism. The following flow chart illustrates the validation process.

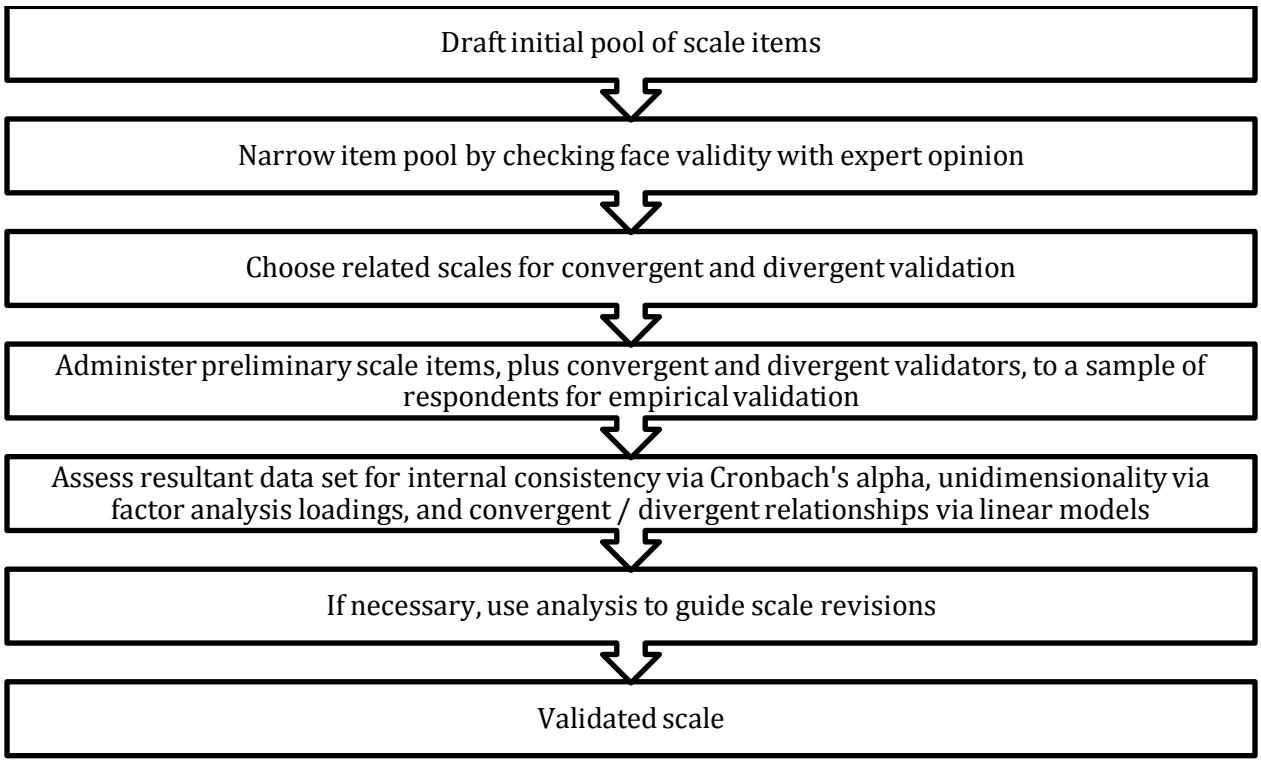

Figure 2. Flow chart for scale validation process.

Theoretical Validation Procedures

Scales like Social Dominance Orientation or Belief in a Just World are constructs for measuring underlying traits. Generally, the underlying trait is a familiar personality feature with a relatively uncontested definition; hence, although it requires a scale to validly

operationalize a trait like empathy into a quantitatively tractable variable, people's shared folk understanding of the nature of empathy furnishes relatively easy consensus regarding what exactly that scale measures.

Considerably more difficult, by comparison, is the construction and validation of scales that measure beliefs or attitudes of a more philosophical nature. While such philosophical attitude scales still involve attitudinal measurements, a unique challenge is introduced: those constructing the scale must deal with the conceptual trickiness of zeroing in on what, exactly, counts as the attitude they are trying to measure. Researchers endeavouring, for example, to construct a scale measuring belief in free will – as my colleagues Jasmine Carey and Del Paulhus have been (Paulhus & Carey, 2011) – must devise a way to navigate amidst ongoing philosophical debates about what, precisely, it means to believe in free will. Considering that the philosophical literature on free will can largely be understood as a discursive competition to furnish and defend a definition of "free will" that withstands the critical scrutiny of other philosophers – and that this competition continues unsettled to the present day – settling on a defensible working definition for the purpose of scale construction amounts to a rather tall order. The same is true of retributivism.

Because the actual definition of what is being measured can be nebulous, scales must compensate for any conceptual shakiness or grey areas by demonstrating validity, and also by making use of simple, clear, direct scale items[5] (Clark & Watson, 1995). With

[5] The conceptual specificity of non-instrumental retributivism posed a considerable difficulty in realizing this goal. Scale items need be as accessible and non-confusing as possible to lay readers while still maintaining conceptual accuracy. Constructing a scale to measure as technical and unfamiliar (to lay readers) an attitude as retributivism presents special challenges in this regard. For instance, consider the following scale item:

these considerations taken into account, and with an extensive review of the literature on retributive punishment fresh in mind, I developed a pool of potential scale items. From this pool, a subset of items would ultimately be selected to constitute the final scale. Of the 22 items in this initial pool, 13 were developed to express characteristically and specifically retributive sentiments, e.g.: "Putting convicted persons into therapy or counseling for their criminal tendencies, rather than punishing them, defeats the main point of criminal justice, because justice is about giving people what they deserve." In the interest of balancing out acquiescence bias – the tendency of survey respondents to agree with propositions suggested to them more than they would agree in a non-survey setting (Messick & Jackson, 1961) – I also developed 9 reverse-coded items[6].

"It doesn't matter whether imposing a deserved punishment is useful; punishing those who deserve it is intrinsically worthwhile even when it does not accomplish anything else."

The language in this item is admittedly at a higher reading level than that found in many psychological scales. Additionally, it could be argued that the item is somewhat "double-barreled" (i.e., responses could reflect more than one rationale or line of reasoning because the item joins two propositions together). However, remedying any of these issues would come at the expense of conceptual accuracy. Since non-instrumental retributivism is a position directly concerned with the intrinsic worth of punishment, irrespective of consequences, the sesquipedalian phrase "intrinsically worthwhile" must remain intact.

Moreover, asking either the half before the semicolon or the half after it in isolation would leave open the possibility of failing to highlight respondents who adhere to some manner of hybrid view (for instance, one on which punishment has no intrinsic worth, but its utility also does not matter). Only agreement or disagreement with the conjunction of the two statements bears directly on the underlying trait under investigation in the scale. Further supporting the design decision to prioritize conceptual accuracy over simplicity is the fact that some scale items that were eliminated in the face-validation process were ones specifically constructed for maximal simplicity.

[6] Although care was taken to avoid articulating specific anti-retributive positions in these reverse-coded, this was only ever possible to a certain degree. For instance, the following item cannot avoid relying on a rehabilitative and consequentialist ideal for contrast:

"A theory of justice that focuses on striking back at deserving individuals for their misdeeds is shortsighted and inadequate compared to one that focuses on actively correcting and preventing the causes and effects of crime."

The careful exclusion of non-instrumental retributivism described in the Introduction was both necessary and non-trivial to implement; scales must measure a unitary underlying trait, and (unless further subscales are developed) the trait must vary unidimensionally. As such, scale items were carefully finessed to exclude the possibility of a respondent expressing an instrumental-retributivist view; all scale items involved, in one way or another, forced choices between desert-reflecting punishments with non-optimal consequences, and desert-ignoring punishments with optimal consequences.

Face Validation Process

The practice of validating scale items by ensuring that they genuinely assess the target concept is known as face validation. Although the notion of face validity is not explicitly mentioned in the 1995 Clark & Watson paper I otherwise utilized for procedural reference on scale validation, is widely attested in the literature (Nevo, 1985; Bornstein, 1996; Hardesty & Bearden, 2004). In the case of this particular scale, it proved possible to seek out and apply the wisdom of experts in order to establish face validity.

I identified 24 professors as experts on punishment theory and the nature of retributivism, based on their publications on these topics. All were either law or philosophy professors at English-speaking universities. I contacted them all, and 14 very graciously completed a short and informal questionnaire in which they assessed the face validity of a pool of potential scale items. Several offered useful feedback in the form of suggestions and commentary. Based on the input from these experts, the pool of potential scale items was narrowed from the original set of 13 pro-retributive items and 9 anti-retributive items to 7 of each. This winnowing process ensured that, relative to the opinions of a broad range of

experts, the questions in the scale truly succeeded at capturing the essence of retributive sentiment.

The final 14-item Endorsement of Retribution scale is as follows:

Pro-retribution items	**Anti-retribution items** (reverse-coded)
Even if a punishment is very costly and will not make anything better in society, it is important to impose it on lawbreakers who commit the crime that corresponds to that punishment.	I sympathize with reformers who say that it is barbaric for the legal system to punish people simply because they deserve it rather than punishing in a way that helps the greater good.
It doesn't matter whether imposing a deserved punishment is useful; punishing those who deserve it is intrinsically worthwhile even when it does not accomplish anything else.	When figuring out whether a punishment is too harsh or too lenient, we should only focus on whether it will successfully accomplish goals like rehabilitating the criminal or deterring future crime.
Even when we have very strong reasons to believe that a criminal offender is now fully committed to resuming life as a law-abiding citizen, we should punish that person anyway, not out of precaution but simply on principle.	In theory at least, there can be good reasons for a punishment that strikes us as cruel, provided such a punishment proves highly effective at deterring crime and reforming offenders.
When a person commits a terrible crime but suffers injuries in the process that will prevent them from ever posing a threat to others again, it is still appropriate to imprison them.	A theory of justice that focuses on striking back at deserving individuals for their misdeeds is shortsighted and inadequate compared to one that focuses on actively correcting and preventing the causes and effects of crime.
Putting convicted persons into therapy or counseling for their criminal tendencies, rather than punishing them, defeats the main point of criminal justice, because justice is about giving people what they deserve.	It seems pointless to me to deal harshly with someone who has committed an illegal act, unless doing so would reform the perpetrator or prevent other similar crimes from occurring.
As long as the legal system fits the punishment to the crime, justice has been served; paying attention to how our treatment of the criminal is helping or hurting society on the whole is a distraction.	The only way the death penalty could be justified is if research shows that the threat of capital punishment actually stops crime more effectively than the threat of imprisonment.
A judge who uses sentencing as a tool to reduce further crime or reform offenders is missing the point: after all, the purpose of sentencing is strictly about appropriately paying criminals back for their misdeeds.	In general, we as a society are too content to imprison criminals without actually considering whether doing so makes society concretely safer or more law-abiding in the long run.

Table 1: Endorsement of Retribution scale items.

Convergent and Divergent Validation

With the condition of face validity thus satisfied, the next step for scale validation involved assessing internal consistency, as well as convergent and divergent (discriminant) validity. This marked the first large-scale empirical undertaking for the project. I constructed an online questionnaire for this purpose that included basic demographic questions, the newly pared-down 14-item Endorsement of Retribution scale, and several other scales chosen as convergent and divergent validators.

The theory behind convergent / divergent validation is that certain relationships should exist between a given scale and other pre-existing well-validated scales, based on how the concepts logically relate (Smith & McCarthy, 1995). For instance, scores on a scale measuring empathy should be inversely related to scores on a scale measuring social dominance orientation (since the latter measures attitudes about how groups of people ought to engage in hostile interactions and out-compete one another); therefore, an investigator tasked with validating one would employ the other as a divergent validator. In theory, a given respondent's score on a validly constructed scale will be significantly – but not *too* strongly – predicted by that respondent's score on the convergent and divergent validators for the scale. If the predicted associations do not emerge, either the validators were chosen in a faulty way, or the scale is not appropriately constructed to measure the target concept. If the predicted associations emerge *too* strongly, then the scale as constructed is redundant, or at least not yet adequately conceptually separate from an existing scale.

For the retributivism scale, choosing convergent and divergent validation scales proved to be an involved task. Few extant scales measure so specific and philosophically

loaded a trait, and equally few measure traits that bear much conceptual relation to attitudes about punishment. Several candidate scales were identified, and after some consideration, four were selected for use in this phase of the validation process.

Scales Selected for Convergent and Divergent Validation

These scales were chosen to be evaluated for their correlation to scores on the Endorsement of Retribution scale in an empirical test of the scale's validity.

- **Belief in a Just World (BJW)**: The BJW scale (Rubin & Peplau, 1975) essentially measures the extent to which a respondent believes that, on balance, people get what they deserve and deserve what they get. Sample items include "I am convinced that in the long run, people will be compensated for injustices," and "I think people try to be fair when making important decisions." BJW has been suggested as an explanation for victim-blaming behaviours (Rubin & Peplau, 1975), in which people find fault with the victim of a harmful or disturbing act, often by suggesting that said victim's misfortune was in some way deserved; an underlying belief in a just world might motivate the perception that terrible things can only befall those who deserve it. Similarly, it seemed *prima facie* plausible that BJW scale scores should display a mild positive correlation with retributivism; it would make some sense that a belief in the eventual success of "karmic" justice and in the inevitability of comeuppance would be accompanied by a strong desire to see such retaliation delivered. BJW measurement options include the original 1975 scale by Rubin & Peplau, a revised 1991 version of the scale by Isaac Lipkus (Lipkus, 1991), and a 6-item version by

Dalbert and colleagues (Dalbert, Montada, & Schmitt, 1987). Owing to its brevity, I used the 6-item Dalbert et al. version.

- **Social Dominance Orientation (SDO)**: The SDO scale (Pratto, Sidanius, Stallworth, & Malle, 1994) examines the extent to which respondents hold a worldview that is equal parts dog-eat-dog, group-oriented, and accepting of inequality - a worldview characterized by the notion that, as one scale item puts it, "It's probably a good thing that certain groups are at the top and other groups are at the bottom." The scale also includes reverse-coded items, such as "No one group should dominate in society." The *prima facie* relationship between (SDO) and retributivism should be evident upon considering that criminal offenders may be viewed as a distinct group within society. For individuals who score high on SDO, any contingent that can be grouped can be out-grouped, and out-groups are to be kept down owing to their putative natural inferiority. It should, then, be the case that individuals keen on social dominance will be keen on retributive punishment - eager to retaliate against members of the outgroup and teach them a deserved lesson about their place in society. The reverse-coded anti-retributive items should clash with the sensibilities of high-SDO respondents, since sentiments expressed in those items eschew desert, humanize criminals, and idealize equality. Several versions of the SDO scale exist; for the purposes of scale validation, the version of choice was the 16-item SDO-6 (Pratto, Sidanius, Stallworth, & Malle, 1994). As mentioned previously, the creators of the SDO scale reported a positive correlation between SDO scores and "belief in retribution" (as measured by their proprietary scale) in their early work presenting

and establishing the SDO scale (Pratto, Sidanius, Stallworth, & Malle, 1994). This further bolsters the case for using SDO as a convergent validator.

- **Empathy**: Although empathy is a much-discussed factor in moral psychology (Pizarro, 2000; Bergman, 2002; Singer, Seymour, O'Doherty, Stephan, Dolan, & Frith, 2006), its role in the complicated context of punishment theory is convoluted and likely not unimodal. Individuals with high empathy might be expected to react with aversion to the unpleasantness of punishment; but they also might be expected to strongly empathize with the victims of crime, in which case they may actually be more supportive of retribution. Further complicating the picture is the fact that recent research (Bartels & Pizarro, 2011) indicates an inverse correlation between empathy and consequentialist moral judgment, with zero-empathy psychopaths reasoning like unflinching, utilitarian dogmatists. Ultimately, I made the decision to include empathy as a divergent validator, figuring that the most significant effect would prove to be increased aversion to pains and privations of criminal punishment. Many well-validated empathy scales exist; I used the one available at http://www.unh.edu/emotional_intelligence/ei%20Measuring%20Mood/mm%20Measuring%20empathy.htm
- **Moral Foundations Questionnaire (MFQ)**: Developed by Jonathan Haidt and colleagues, the MFQ measures the extent to which a respondent endorses each of five foundational moral concerns (as individuated by cross-cultural research on this topic) (Haidt & Graham, 2007). The moral foundations assessed are harm/care, justice/fairness, loyalty/in-group, authority/hierarchy, and purity/sanctity. Based on previous work by Haidt, Spassena Koleva, Jesse Graham, and others, I chose to

focus on Purity as a convergent validator; although the logical connection may seem somewhat opaque, purity/sanctity has proven to be the moral foundation most predictive of conservative stances on culture-war issues (Koleva, Graham, Haidt, Iyer, & Ditto, 2009). Following the same logic operative in the inclusion of an empathy scale, I further selected the Harm subscale as a divergent validator. Finally, owing to the intuitive conceptual linkage between authoritarianism and harsh punishment regimes, I selected Authority as a convergent validator. For the sake of completeness, I also included the scale items probing justice and loyalty, but my prediction was that they would not exhibit any clear relationship to retributivism. As Haidt and his colleagues encourage interested researchers to make use of the MFQ in their own experiments at http://faculty.virginia.edu/haidtlab/mft/index.php?t=questionnaires, I utilized the version made available at that link.

Scales Considered but Not Used as Convergent / Divergent Validators

These scales were included on an initial short-list of candidates for convergent and divergent validation, but were ultimately not included for reasons explained below.

- **Right-Wing Authoritarianism (RWA)**: Originally, I considered including this scale (by Altemeyer, 1981), measuring respondents' respect for authority, hostility toward out-groups, and support for traditional social norms. However, during the process of researching the scale, doubts emerged as to its utility and added value relative to other scales already being included. Despite its strong internal

consistency scores, I doubted that the RWA scale would reveal much over and above the information already gleaned from the SDO, BJW, and MFQ scales.

- **Behavioral Approach and Behavioral Inhibition Scales (BAS & BIS)**: Initial forays into the scale validation literature suggested that the widely-used BAS and BIS scales, measuring the strength of appetitive motives and of aversive motives respectively, might be appropriate as convergent or divergent validators. However, further investigation revealed that their conceptual relatedness to the target phenomenon was minimal.
- **Agreeableness**: The prominence of the "Five Factor Model" of personality traits (Agreeableness, Neuroticism, Extroversion, Conscientiousness, Openness) in social psychology led me to consider including at least one such measure. Of the five, Agreeableness seemed to be the least disconnected from people's attitudes toward punishment. However, this constituted a weak justification for lengthening an already time-consuming survey, so I left off Agreeableness.

Empirical Validation Procedures and Results

With convergent and divergent validators chosen, the design of the validation questionnaire was complete. All scale items were presented as 9-point Likert scales, with the left-most option labeled "strongly disagree" and the right-most option labeled "strongly agree." The order of scale items was randomized for all scales. For analytical purposes, the Endorsement of Retribution scale items corresponded to variables labeled PR1 through PR7 (for items that measure pro-retribution attitudes) and AR1 through AR7 (for items that measure anti-retribution attitudes, hence the reverse coding). A multiple-choice

comprehension check was included at the end to ensure that respondents were paying attention to the tasks at hand: respondents were asked to identify which of the comprehension check answer choices was not asked about in the scales they had just completed.

I obtained approval from the UBC Behavioural Research Ethics Board (BREB certificate number H11-02821) to administer the scale-validation questionnaire online. All participants gave their informed consent to participate and to the use of their data for the purposes of the study. I recruited my sample of respondents by offering 25 cents per completed survey using Amazon Mechanical Turk. This online labour market allows for expedient sampling of a subject pool that has proven no less representative of North American populations than traditional subject pools (Ipeirotis, 2010; Paolacci, Chandler, & Ipeirotis, 2010).

In the resultant data set of 259 completed responses, the results from 12 respondents who failed the comprehension check were discarded. As a set of additional measures against satisficing, 6 response sets were deleted due to their completion time clocking in at less than 4 minutes[7]. Now at n = 241, I conducted a series of statistical analyses on the resultant data set. All statistical analyses were carried out using R Project software.

[7] The 4-minute mark was chosen based on the distribution of completion times in the data set; a clear gap was found between respondents whose completion times indicated they were obviously satisficing (finishing the eight pages of the questionnaire in 1 or 2 minutes) and the rest of the respondents (the next lowest times were in the 5 minute range).

Sample Profile

Age	Mean = 36.5 years; SD = 12.9; median = 33
Gender	144 female, 93 male, 4 declined to specify
Religiosity (1 = lowest, 9 = highest)	Mean = 4.1; SD = 2.8; mode = 1
Most common religious affiliations	Christianity (*n* = 129); none (*n* = 62)
Most common education levels	Some college / university (*n* = 74); Bachelor's degree or equivalent (*n* = 61); high school diploma (*n* = 35)

Table 2: Demographic data for empirical scale validation.
n = 241. Note the mode for religiosity, indicating a strongly right-tailed distribution.

Factor Analysis to Diagnose Faulty Scale Items

I used factor analysis to investigate whether any of the Endorsement of Retribution scale items were faulty. In this context, a scale item is faulty if it fails to track the same underlying attitude as the rest of the scale items do. Factor analysis is a procedure for discerning whether variation among a set of variables can be more parsimoniously explained in terms of a few hidden underlying factors. It is useful for scale validation because – according to the logic of scales – the variance exhibited by each variable (i.e., answers to the scale items) ought to be explicable in terms of an underlying feature, namely the conceptual target of the scale. For the purposes of scale validation, factor analysis has been characterized as more of a diagnostic tool for assessing errant variables in a scale rather than as an actual indicator of the unitary nature of the scale (Grau, 2011). The most useful indicator for the latter feature is Cronbach's alpha, discussed in the next section.

To provide a preliminary indicator of how many factors to include in the factor analysis, I used a scree plot. It returned optimal coordinates of 2, so I ran factor analyses with 1, 2, and 3 factors. In each case, I re-ran each analysis using the varimax and promax rotations, as well as with no rotation. This method of diversely re-specifying the analyses

helped to ensure that significant results reflected actual effects and not merely artifacts of model specification. The next paragraph details the results of the analyses, grouped by how many factors were included in the formula.

When I specified two-factor or higher models, no matter which rotation I applied, the variables AR2, AR3, and AR4 tended to load on factors other than the main one (cf. Table 1 for the text of the scale items corresponding to these variables). Other variables loaded on the secondary and tertiary factors as well, but these loadings were weak and not consistent across various setups, whereas AR2, AR3, and AR4 reliably failed to load with the first, main factor. When I specified one-factor models, no matter which rotation I applied, only the variable AR3 fell below the factor loading display cutoff of 0.1. Hence, it appears that despite the failure of AR2 and AR4 to load on the main factor when other factors are available, the most problematic variable when specifying a one-factor setup is AR3. Since AR3 was the only variable to appear problematic in every factor analysis I ran, I noted it as a candidate for deletion from the scale.

These considerations led me to re-run the analyses with AR3 excluded. However, doing so actually did not noticeably improve the results. Re-running the factor analyses with AR3 removed caused many of the remaining variables to scatter to other factors, at least in two-factor models and higher. In one-factor models, removing AR3 causes all variables to be assigned factor loadings above 0.15. If AR3 were a truly faulty variable, removing it would have caused factor loadings to increase and to cluster further on one variable; instead, the opposite happened, complicating the case for deleting AR3.

Quantifying the Internal Consistency of the Scale Using Cronbach's Alpha

I calculated Cronbach's alpha for all the items in the Endorsement of Retribution scale. Subsequently, I calculated it for the scale with AR3 removed, and for the scale with AR2, AR3, and AR4 removed. Cronbach's alpha measures the internal consistency of a group of items – essentially, how strongly the responses to each item predict one another. It is a crucial feature of valid psychological scales, since scales must measure a unitary underlying trait. The alpha for the scale as a whole was .80, indicating high internal consistency. With the removal of AR3, the alpha rises to .83; with the removal of AR2, AR3, and AR4, the alpha is .84.

I also calculated Cronbach's alpha separately for PR1-7 and AR1-7, essentially treating the straight-coded and reverse-coded items as subscales of their own. The alpha for the pro-retribution items only was .82; the alpha for the anti-retribution items only was .61, suggesting that in general the reverse-coded scale items measure endorsement of retributivism with considerably lower reliability.

This finding raised the possibility that acquiescence bias may not have been a factor in the scale item responses. Calculations indicated that the mean response for the PR items was 5.07, and the mean response for the AR items was 5.14. A t-test revealed $p = .61$ for this difference, confirming that that acquiescence bias was not present in this sample.

Linear Models to Assess Convergent and Divergent Validity

For each of the other scales in the survey, I used linear regression models to evaluate whether average scores were correlated. I opted to use linear regression rather than generalized linear models because the 9-point Likert variables may legitimately be

treated as continuous, and because the distribution of average scale scores approximated a normal distribution in every case (except SDO, detailed below). I set up the regression models in a variety of ways, including and excluding different variables as needed to explore the data. Here I report on the conclusions that emerged from looking for results that remained robust across a broad array of model re-specifications – a statistical practice that guards against the pitfalls of model misspecification (Spanos, 2006). Listed from least correlated to most correlated, they are:

- **Empathy**: No correlation between empathy and retributivism emerged, no matter how the model was set up. In most models, p-values exceeded .5.
- **Loyalty**: This variable, from Haidt's MFQ, appeared positively correlated when modelled in isolation, but its variance was always absorbed by other variables across diverse model re-specifications, leaving it with no meaningful correlation.
- **BJW**: A weakly positive correlation of usually low significance appeared at times, especially when modelled in isolation, but it was absorbed by other variables in many models. These results make it unlikely that BJW is a useful predictor of retributivism, except possibly via intermediary variables.
- **Justice**: Modelled in isolation, this variable shows no correlation with Endorsement of Retribution score. Some putatively significant correlations emerge when the variable appears in more inclusive models. The most parsimonious explanation is that the correlations in the inclusive models are artifacts of the modelling rather than real effects.
- **Authority**: Although this variable showed no significant correlation with Endorsement of Retribution score in a kitchen-sink model including every variable,

it emerged as a significant predictor of retributivism in other, less inclusive models. The balance of evidence suggests a positive correlation, but not a commanding one: in most cases, especially the more inclusive models, the coefficient for the variable's slope is below .15.

- **Purity**: This variable displayed 2- or 3-sigma positive correlation in every model where it appeared, with slope coefficients ranging from roughly .15 to .25; while still not the single most influential variable in the data set, it represents a significant positive predictor, even taking a parsimonious statistical approach.
- **Harm**: For this variable, a robust inverse correlation, 2- or 3-sigma depending on the model, emerged from every setup. Its slope typically displayed coefficients of roughly -0.2. While modelling the Harm variable in isolation produced a very low R-squared value, its unusually high coefficient and significant p-value in every model that features it designate this variable a clear divergent validator.
- **SDO:** The distribution of this variable is non-normal, with a majority of average scores in the 1-3 range and a long, narrow right tail. However, transforming the variable by taking its logarithm does not substantively change any of the models, despite yielding a distribution relatively closer to normal. Given the apparently minimal impact of the normality violation, I report results here from models using the un-transformed variable. SDO consistently stood out as the variable with the strongest, most statistically significant relationship to Endorsement of Retribution score, frequently appearing in models as a 3-sigma variable with slope values clustering around 0.3.

Discussion of Empirical Validation Results

Based on these results, I argue that convergent and divergent validity have been adequately demonstrated. Exhibiting a high Social Dominance Orientation and endorsing purity as a moral value are positively correlated with endorsement of retribution. A case could also be made for including endorsement of authoritarian moral principles on this list of positively correlated predictors; especially when considered alongside the Purity variable, these results suggest a strong association between retributivism and political conservatism in general[8]. Complementing these observations is the inverse correlation between moral sensitivity to harm and retributivism: so-called "bleeding hearts," being averse to any harm visited on people for any reason, shy away from endorsing retribution and are more prevalent on the political left.

All of these relationships make theoretical sense. Retributive punishment is, indeed, a variety of social dominance: it involves one group (the perceived in-group of law-abiding citizens) figuratively stepping on another group (the perceived out-group of criminal offenders), in a particularly societally endorsed manner. Retributivism also requires and reinforces strong authority dynamics – indeed, a hallmark of authoritarianism is a tendency to hold that individuals who question or challenge authority deserve any rough treatment they may suffer as a result of doing so. The correlation between retributivism and emphasis on the moral salience of purity is predictable by way of the extent to which each appeals to categorical moral reasoning: an individual's underlying willingness to believe that particular acts may be justified or proscribed purely by reference to their nature, rather than their consequences, underwrites the ability to consider what is

[8] Even more evidence for the politicized nature of retributive attitudes can be found in Chapter 2.

unnatural to be wrong and what is deserved to be justified *ipso facto*. Conversely, a strong sensitivity to harm likely contributes to a stance on punishment that seeks the lowest levels of harm-infliction overall, especially rehabilitative approaches that clash with traditional retributivism.

The null-result variables also make sense. It might have seemed reasonable to expect an inverse relationship between empathy and retributivism for reasons mirroring those of the harm-retributivism relationship. However, some research (Bartels & Pizarro, 2011) has shown an inverse correlation between empathy and utilitarianism. It is not, then, surprising that no conclusive association emerges between the two variables, given the conflicting nature of the theoretical cues. With regard to the two MFQ items that displayed no correlation, namely Loyalty and Justice, there is no clear reason to suppose that the concepts of loyalty and punishment are connected in such a way that one reliably influences the other; one might be tempted to suspect, on the basis of terminological similarity alone, that scores on the Justice subscale would predict Endorsement of Retribution score, but indeed, the kind of justice probed by the MFQ items is more akin to a Rawlsian understanding grounded in principles of fairness[9]. What constitutes a fair punishment, then, is precisely what remains at issue between retributivists and consequentialists.

Finally, the lack of a significant correlation between Belief in a Just World and retributivism may seem most surprising and perhaps a threat to the claim that the retributivism scale displays convergent validity. However, note that, when modelled in isolation, the association between BJW and Endorsement of Retribution score is positive

[9] Haidt himself, in recent lectures, has moved away from the conception of fairness expressed in the version of the MFQ I used.

and significant; it simply recedes into non-significance when modelled alongside other variables. This is exactly what one might expect for a predictor whose relationship with the output variable is likely mediated by other factors. People inclined to believe that the world is a just place may tend towards retributivism, but the explanatory work is better done by variables like SDO.

The results from the factor analysis process do leave some room for concern that the reverse-coded variables AR2, AR3, and AR4 may be problematic. In particular, I flagged the variable AR3 as a potential candidate for deletion from the scale, but refrained from deleting it because further investigative factor analysis indicated that little if any benefit would follow from its removal. The results of the Cronbach's alpha calculations further support the decision to refrain from deleting AR3, and generally indicate – *contra* the results from the factor analysis process – that the scale successfully measures a unitary target. AR3 is also the only variable with a strong theoretical basis for exclusion. It corresponds to the following scale item: "In theory at least, there can be good reasons for a punishment that strikes us as cruel, provided such a punishment proves highly effective at deterring crime and reforming offenders." Disagreeing with this statement, according to the initial logic of the scale design, indicates a stronger endorsement of retributivism, whereas disagreeing indicates a position at odds with retributivism. Nor was the logic behind it idiosyncratic: the law professors who evaluated face validity for the scale items gave AR3 strong marks.

Lay participants, however, felt otherwise. Even though an allowance for consequentially ideal over-punishing is squarely at odds with retributivism, it is not surprising that people who are uncomfortable with retributive principles in general would

feel unease at the prospect of over-punishing. They almost certainly do not recognize that the most effective philosophical justification for restrictions on over-punishment is rooted in principles of desert and proportionality; anti-retributive respondents simply imagine a case of over-punishing and intuitively dislike it for the same reason that they dislike a retributive punishment when its consequentialist alternative is less harsh – namely, that harsh punishment feels barbaric and cruel to them. In that sense, then, their refusal to endorse consequentialism-supported over-punishment is not a useful indicator of their general sentiments toward more paradigmatic cases of retributive punishment[10].

Overall, statistical and theoretical considerations each supply reasons to remove AR3 from the scale and reasons to refrain from doing so. Without further data to elucidate matters, I must leave the resolution of this quandary to future investigators. Meanwhile, the experiments in Chapter 2 of this thesis required a pragmatic solution, one allowing the Endorsement of Retribution scale to be utilized. Owing to the lack of evidence for acquiescence bias, as well as the significantly higher Cronbach's alpha for the pro-retribution scale items, I chose to run all analyses in Chapter 2's studies twice – once using the entire scale, and a second time excluding all the reverse-coded items, effectively constituting a pro-Endorsement of Retribution subscale. This strategy is justified because the threat of acquiescence bias is what prompted the inclusion of reverse-coded items in the first place.

[10] This finding does, however, indicate that limiting retributivism may in fact be a theory that strongly accords with folk attitudes. After all, limiting retributivism incorporates consequentialist adjustments to punishment but restricts over- and under-punishment on desert-based grounds.

The balance of evidence indicates that this scale measures what it was designed to measure. Its internal consistency is strong, its convergent and divergent validators make theoretical sense, and the items have solid credentials for face validity. By virtue of the analyses detailed here, I find it justifiable to conclude that the Endorsement of Retribution scale meets all the requisite criteria for validity.

Chapter 2: Testing the Consistency of *Mens Rea* Attributions in Contrastively Parallel Cases

Rationale

The aims of the experiments in this Chapter were three-fold: to replicate the results of Thomas Nadelhoffer's 2006 "Bad Acts, Blameworthy Agents, and Intentional Actions" in a different context; to explore the general relationship between Endorsement of Retributivism and between-group response differences on *mens rea* attribution tasks; and to test the specific hypothesis that Endorsement of Retributivism would positively correlate with an increased likelihood for respondents to vote guilty when presented with the version of the task in which the defendant's character elicited negative moral reactions.

Design and Methods

Why Not Simply Duplicate Nadelhoffer's Experiments?

The vignettes in the 2006 Nadelhoffer study – in turn adapted from the details of the landmark 1961 case of *D. P. P. v. Smith* in Britain – represent the skeleton around which the vignettes I used for these experiments were built. However, several key features of the original vignettes undermined the option of simply adopting them unaltered for my purposes.

To see what these features were, it will be useful to refer to the text of Nadelhoffer's scenarios, as follows:

> (C1) Imagine that a thief is driving a car full of recently stolen goods. While he is waiting at a red light, a police officer comes up to the window of the car while brandishing a gun. When he sees the officer, the thief speeds off through the

> intersection. Amazingly, the officer manages to hold on to the side of the car as it speeds off. The thief swerves in a zigzag fashion in the hopes of escaping— knowing full well that doing so places the officer in grave danger. But the thief doesn't care; he just wants to get away. Unfortunately for the officer, the thief's attempt to shake him off is successful. As a result, the officer rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.
>
> (C2) Imagine that a man is waiting in his car at a red light. Suddenly, a car thief approaches his window while brandishing a gun. When he sees the thief, the driver panics and speeds off through the intersection. Amazingly, the thief manages to hold on to the side of the car as it speeds off. The driver swerves in a zigzag fashion in the hopes of escaping—knowing full well that doing so places the thief in grave danger. But the driver doesn't care; he just wants to get away. Unfortunately for the thief, the driver's attempt to shake him off is successful. As a result, the thief rolls into oncoming traffic and sustains fatal injuries. He dies minutes later. (Nadelhoffer, 2006, p. 13-14)

Crucially, note that the primary difference between these vignettes is the identity of the characters. Nadelhoffer characterizes the two cases as "structurally identical" (*ibid.*, p. 13) This is true – insofar as the questions being asked after the vignettes pertain to folk concepts (as distinct from formal legal concepts) of intentional action, knowledge, and blame (which, indeed, they do). However, properly speaking, these are not identical with the question of legal guilt. Nor are the vignettes indisputably structurally identical from a legal standpoint. In jurisdictions that employ the felony murder rule, the thief character's actions in C1 would almost certainly render him guilty of murder regardless of his mental

state. Meanwhile, depending on the jurisdiction, the driver's actions in C2 could be construed as meeting the criteria for justifiable homicide.

Updating the Vignette Design

In order to make these experiments more directly applicable to the legal realm, I removed the police officer element and downplayed the putative threat to the driver while maintaining plausibility to the greatest extent possible. The resultant vignettes, whose full versions are provided in the Appendix, feature two characters – a local doctor and a notorious shoplifter (who is not, for clarity, actively shoplifting anything at the time of the incident). Their roles are neatly reversed, with the cause of the driver's panic explained as follows: the doctor recognizes the shoplifter from a recent alert on the news, and in the case where the doctor is the driver, he fears being carjacked, whereas when the thief is the driver, the thief fears being apprehended. No information is provided in either case as to the motive of the individual who approaches and hangs on to the vehicle, but in both cases it is emphasized that he is clearly unarmed. It is explicitly stated that the driver of the vehicle swerves with the knowledge that doing so puts the hanger-on in great danger, but with the sole intention of getting away.

The vignettes as I have redesigned them are structurally identical even from a legal standpoint, with one exception: the most reasonable inference about the motive of the person approaching and hanging onto the vehicle is incongruent between the doctor (ostensibly he means to apprehend the shoplifter) and the shoplifter (ostensibly he is up to no good, of some sort). However, the person who approached the vehicle is not the one on

trial – the driver is, and his assailant's motives do not bear on the question of whether he met the "knowingly and intentionally" standard of guilt[11].

The question I asked respondents following the vignette also required extra specificity. Instead of relying directly on folk concepts of knowledge and intentional action, I specified a particular standard of guilt for homicide ("the defendant must have intentionally and knowingly brought about the death of the victim") and instructed respondents to vote guilty or not guilty based on the alignment between the facts of the case and the standard of guilt. Participants were instructed to bracket off concerns about justifiable homicide, with the explanation that such considerations (in the fictional, vaguely North American jurisdiction where this case transpires) are handled by the judge during sentencing. Respondents who voted not guilty were further queried as to which element of the standard of guilt they felt the facts of the case failed to meet: did the defendant not intentionally cause the victim's death, not knowingly cause the victim's death, or not cause the death in the first place?

As in Nadelhoffer's study, participants were randomized to one of the two vignettes and were not made aware of the existence of the other vignette, a typical design for between-subjects comparison. Here, the vignette in which the local doctor was the driver of the vehicle will be labeled DV (for doctor vignette) and the one in which the notorious shoplifter was driving will be labeled SV (for shoplifter vignette). In any case, when a

[11] In a broader sense, the probable motive of the vehicle-approaching party is indeed legally relevant, since it props up a case for this incident being regarded as a case of justifiable homicide. However, for it to be a justifiable homicide, it must be a case of homicide rather than of manslaughter in the first pace. As the judge's instructions in the vignette indicate, this determination is where the jurors come in, thereby ensuring that the broader relevance of the victim's motive does not properly enter into consideration of the question at hand.

character's name is used to refer to a vignette, this name always denotes the driver of the vehicle and hence the defendant on trial.

Structure of the Questionnaires

In addition to the vignettes, participants also filled out basic demographic items. In contrast to those from the previous Chapter, the demographic items did not include the question on religious affiliation (though the Likert scale for level of religiosity was kept), instead newly including a Likert scale question on political views. Also included were questions prompting respondents to report whether they had any legal education / legal work experience, and whether (to the best of their knowledge) they were eligible, at the time of their response, to serve on a jury. Finally, after the demographic items and juror task vignettes, all respondents completed the Endorsement of Retribution scale. At the end of the questionnaire, respondents were required to answer a multiple-choice comprehension check in which they were prompted to recall, from the vignette, what specific action the driver took that caused the hanger-on to fall from the car into traffic. This check was crucial since the interpretation of the results assumes that respondents paid close attention to the details of the case.

All experiments were conducted using Amazon Mechanical Turk, with approval granted by the UBC Behavioural Research Ethics Board as an extension of the work from Chapter 1 (thus also falling under BREB Certificate Number H11-02821). All respondents provided informed consent to participate and to the use of their data for the purposes of these studies. Participants, who were informed that the questionnaire would take roughly 10 minutes to complete, were offered compensation of 40 cents for completing it.

Rationale for Follow-up Studies

The choice to only present respondents with one vignette (rather than both sequentially) was premised on the assumption that significant order effects would plague the second vignette that any given respondent viewed (since the switch of characters is quite evident and would clue respondents in to the design of the experiment). In order to test this assumption, I conducted Follow-up Study 1 (FS1), in which both vignettes were presented sequentially to all participants (with the order randomized). The survey form, recruitment methods, ethics approval, and compensation were otherwise identical to those in the main experiment.

I also designed and carried out Follow-Up Study 2 (FS2), in which the "notorious shoplifter" was replaced with a "local telemarketer." The characters in the vignettes for the main study – the local doctor and the notorious shoplifter – are intended to establish a clear distinction between the sympathetic figure and the unsympathetic figure in the story. I maintain that these characterizations are appropriate to the goals of the study; the shoplifter's history of petty crime serves to elicit negative feelings toward him without introducing any factors that would make his case legally different (except perhaps from a sentencing standpoint, which is beyond the scope of this study). Nonetheless, I anticipated that critics might object to the stark difference between the two characters, perhaps contending that the clear demarcation between the good character and the bad character makes the discrepancy in conviction rates all but guaranteed.

I find this line of argument problematic for several reasons – principally that science with obvious results is still science and no less worth carrying out. Moreover, as the following section will make clear, FS1 demonstrated that people are, in fact, inclined to

treat the putative good guy and the putative bad guy equally on pain of inconsistency, undermining the notion that the characters are somehow too different. Nonetheless, I did consider it reasonable to worry that characterizing the unsympathetic character as a lifelong criminal, rather than as a merely unsavory (but law-abiding) citizen, would make for a less interesting result. For the sake of having more data to speak to these concerns, and in pursuit of an even more interesting result, decided to run FS2. The telemarketer character for this study was chosen because according to the 2011 Gallup "Honesty/Ethics in Professions" poll – available at http://www.gallup.com/poll/1654/honesty-ethics-professions.aspx – telemarketers are perceived as having among the lowest honesty and ethical standards among a wide away of professions.

The pretext for the driver's panic in FS2 is supplied by a case of mistaken identity: each man mistakes the other for a local shoplifter whose photograph has been broadcast recently on the local news. The pedestrian (in both cases, whether the doctor or the telemarketer) approaches the vehicle hoping to apprehend its driver, and the driver panics in fear of being carjacked. The survey form, recruitment methods, ethics approval, and compensation were otherwise identical to those in the main experiment. All statistical analyses were carried out using R Project software.

Results and Analysis

Combining Main Data Set and Data from Follow-Up Study 1

The main purpose of FS1 was to test the assumption that participants' votes of guilty or not guilty in the second vignette they viewed would be strongly affected by having viewed a prior vignette. The percentages of guilty votes in FS1 were:

% Voting Guilty (whole sample) (n = 387)	When Viewed First	When Viewed Second
Doctor	28.1	41.9
Shoplifter	52.9	28.1

% Voting Guilty (no legal educ.) (n = 361)	When Viewed First	When Viewed Second
Doctor	27.1	41.7
Shoplifter	52.8	30.4

% Voting Guilty (eligible for jury) (n = 340)	When Viewed First	When Viewed Second
Doctor	27.8	43.1
Shoplifter	52.1	29.5

Table 3. Conviction rates for Follow-Up Study 1.
Note the similarity between diagonally adjacent cells.

The dramatic switches in conviction rates based on whether the DV case or the SV case was viewed first or second, alongside the generally high levels of consistency between doctor-first and shoplifter-second rates (and vice versa), strongly indicate that participants' votes in the second vignette were largely driven by a desire to render the same verdict in each case, so as not to appear inconsistent. This result is telling in several ways. It suggests that, whether or not the law considers these cases structurally identical and hence deserving of an identical verdict, laypeople certainly do – though, they appear somewhat more willing to switch their vote when the shoplifter case is first and the doctor second than vice versa[12].

More to the point, this result validates the assumption that a between-groups contrastive vignette design is crucial to the experiment. Once respondents pick up on the

[12] This is an interesting framing effect in its own right; the result hints (but does not establish) that laypeople may generally be more comfortable with letting an unsavory character off the hook for reasons of consistency than with punishing a sympathetic character for the same reason.

contrast, their answers reflect their commitment to consistency rather than their raw attitudes about guilt. FS1 thus served its primary purpose.

Since the only difference between the main study and FS1 was the addition of the second vignette, the data sets from the main study and the first vignette of FS1 had no structural impediments to being combined into one larger data set. Statistically, this would be a defensible step only if there were no theoretical reason to expect any differences between the samples. This was indeed the case for the data sets in question; because the survey forms were so similar and the sampling method unchanged, the data collection for FS1 was effectively an extension of the main data set, as though I had simply left the main survey open until 800 responses had accumulated rather than 413. The combination of data sets is further justified when the general response trends in the samples are otherwise indistinguishable, and when reasonable measures are taken to filter out repeat respondents (which they were). I used t-tests to see if any significant differences obtained between the main data set and the follow-up data set.

	Mean (Main Data Set)	Mean (Follow-up Data)	P-value of Difference
Age	31.1	31.7	0.515
Gender	0.6	0.5	**0.002**
Political Views	3.9	4.1	0.146
Religiosity	3.5	4.2	**0.001**
Legal Education	0.1	0.1	0.973
Endorsement of Retribution score	4.9	4.8	0.684
Guilty / Not Guilty ratio	0.4	0.4	.916

Table 4: Demographic data for main and Follow-Up Study 1 data sets.
In this table and all others for this Chapter, statistically significant results are in bold font.

Although significant differences were detected in the two samples' gender proportions and mean religiosity, none of the important outcome variables were appreciably different. As such, I deemed it justifiable to combine the data sets from the main study and the follow-up to create one large data set with an *n* of 800. As a further check for the validity of this move, I tested some of the key findings in the paragraphs below using the original *n* = 414 data set and found no results suggesting that the combination of data sets altered the trends observed in the analysis. I also tagged the responses with a dummy variable encoding which data set they came from, and tested this variable for interaction effects with other variables in a random selection of the models used below; no significant interaction effects emerged from any such test.

Sample Profile

Age	Mean = 31.4; SD = 11.6; median = 27
Gender	370 male, 423 female, 7 declined to specify
Political views (1 = liberal, 9 = conservative)	Mean = 3.9; SD = 2.1; mode = 1
Religiosity (1 = lowest, 9 = highest)	Mean = 3.8; SD = 2.9; mode = 1
Endorsement of Retribution, full scale (1 = lowest, 9 = highest)	Mean = 4.8; SD = 1.2; median = 4.8
Endorsement of Retribution, half scale (1 = lowest, 9 = highest)	Mean = 4.8; SD = 1.6; median = 4.9
Most common education levels	some university/college (*n* = 245); "Bachelor's degree or equivalent" (*n* = 212); "high school diploma" (*n* = 107)

Table 5: Demographic data for combined data set (*n* = 800).
n = 800. Distributions for political views and religiosity were strongly right-tailed. Distributions for Endorsement of Retribution (including the subscale of pro-retribution items) were near normal.

Testing for Effects of Vignette Type on Retribution Score

All participants filled out the Endorsement of Retribution scale after completing the either the DV task or the SV task. Using a t-test, I investigated whether the type of vignette a respondent received (DV or SV) significantly influenced Endorsement of Retribution scores (EoR score). The type of vignette a respondent received can be treated as a variable – hereafter referred to as VT, for "vignette type" – that can take one of two values, DV or SV. The mean EoR score for participants in the DV condition was 4.83, and for the SV condition it was 4.84 ($p = .915$). This result indicates that participants' EoR score was not at all sensitive to which version of the vignette they were shown.

I ran the same t-test looking only at the impact of VT on the pro-retribution scale items. The means for the DV condition and the SV condition were 4.87 and 4.81 respectively – a difference whose p-value came out to .573, indicating no statistical significance[13]. Since both VT and EoR score were meant to be included as dependent variables, a significant interaction between them would have violated several important statistical assumptions later in the analysis. Verifying that no such effect was present in the data places the following analyses on surer footing.

Modelling the Relationship Between Variables of Interest

The design of the experiment focuses on respondents' vote of guilty (hereafter G) or not guilty (hereafter NG) as the outcome variable of interest – hereafter referred to as JD, for "juror decision." The main variables to model JD against were VT (i.e., whether a respondent was responding to the DV or SD condition) and EoR score, with the

[13] As an aside, running the same test on the anti-retribution items yielded means of 4.79 (DV) and 4.87 (SV), for a p-value of .357 – also not a statistically significant effect.

demographic variables as controls. I used binomial generalized linear models (GLMs) to investigate the main questions of interest. Binomial GLMs were the modeling technique of choice since the outcome variable, JD, is binary – it only takes the form of G or NG. To begin with, I ran a simple model with JD as the independent variable and the interaction effect between VT and EoR score as the dependent variable. I used analysis of deviance (type II ANOVA) to analyze the models. The results indicated a strong effect of VT on likelihood of voting guilty – a chi-squared value of 55.37 with $p < .00001$ – indicating that Nadelhoffer's results are overwhelmingly preserved within the context of this experiment.

The effect of EoR score on JD was palpable but not as dramatic as that of VT; the chi-squared value was 5.90, and the significance estimate .015; to be clear, this is the effect of retribution-endorsement on likelihood to vote guilty in either vignette. These data suggest that retributive individuals are generally more likely to infer a defendant is guilty no matter what the circumstances may be. However, further statistical modeling cast doubt on this result.

The variable of greatest interest for the primary hypothesis of this experiment fared poorly in this model. The interaction effect between EoR score and VT, which describes how much more likely to vote guilty in the SV condition (but, crucially, not the DV condition) an individual becomes as that individual's EoR score increases, displayed a chi-squared value of .798, but a significance estimate of .372 – indicating a null result. These data indicate that the degree to which a respondent endorses retribution makes that respondent no more likely to vote the shoplifter guilty than the doctor.

Results from a single model, however, are not conclusive in their own right. Hence, it was important to test several variations on the original model as checks against the results

from the most straightforward version. Crucially, this experiment is meant to approximate how actual North American jurors might behave in cases such as these, so I re-ran the models using subsets of the data – one including only participants who self-reported eligibility to serve on a jury, another including only participants who reported no legal education or work experience, and yet another including participants who met both these criteria. Additionally, as I concluded at the end of Chapter 1, all models that used the Endorsement of Retribution score were re-run using the score calculated from only the pro-retribution scale items. Finally, to ensure that the main conclusions from the analysis would hold up when controlling for the demographic factors I collected, I created inclusive versions of each binomial GLM, which modelled the JD variable against VT, EoR score (either using all scale items or only pro items), and the interaction effect between VT and EoR score, as well as age, gender, political views, and degree of religiosity[14].

The following tables contain the results from running analyses of deviance on the aforementioned models. The first number in each cell refers to the Chi^2 value, which describes the effect size – the extent to which the variable in question influenced the odds of voting guilty (JD). The second number in each cell refers to the probability estimate that the Chi^2 value emerged due to chance alone – the p-value. Note that for the table row covering VT in the simple models, two sets of values were obtained – one for the models using the full scale for EoR score, another for models using the pro-retribution subset of items for EoR score (hereafter, pro-EoR score). As the latter were never appreciably different, I report only the former.

[14] The variables for political views and religiosity were, as noted earlier, not normally distributed. However, the assumptions for generalized linear models do not include normal distribution for continuous dependent variables, so including these variables in the binomial GLMs here remains a valid option.

Effect on JD of:	Whole sample	Jury eligible	No legal exp	Eligible & no exp
VT	**55.37; ~0**	**46.49; ~0**	**54.90; ~0**	**44.70; ~0**
EoR score	**5.90; .015**	**5.30; .021**	2.85; .091	2.56; .110
EoR*VT interaction	0.80; .372	0.74; .391	1.23; .267	1.47; .226
Pro-EoR score	**10.73; .001**	**7.42; .006**	**7.46; .006**	**5.07; .024**
Pro-EoR*VT interaction	1.79; .181	1.22; .269	2.16; .141	1.75; .186

Table 6. Results from simple models of guilt.
Formulae: (JD ~ EoR*VT) & (JD ~ pro-EoR*VT)

Effect on JD of:	Whole sample	Jury eligible	No legal exp	Eligible & no exp
VT	**54.96; ~0**	**46.04; ~0**	**54.75; ~0**	**44.37; ~0**
EoR score	**5.61; .018**	**4.53; .033**	2.66; .103	1.91; .167
EoR*VT interaction	1.29; .257	1.28; .257	1.69; .194	1.97; .161
Age	**16.78; ~0**	**15.52; ~0**	**14.49; ~0**	**13.28; ~0**
Gender	1.83; .177	2.02; .156	2.97; .085	3.50; .061
Political views	0.84; .360	0.82; .365	0.78; .377	0.71; .399
Religiosity	1.47; .227	1.63; .202	0.89; .194	1.27; .260

Table 7. Results from inclusive models of guilt.
Formula: (JD ~ EoR*VT + age + gender + politics + religiosity)

Effect on JD of:	Whole sample	Jury eligible	No legal exp	Eligible & no exp
VT	**56.26; ~0**	**47.43; ~0**	**55.82; ~0**	**45.53; ~0**
Pro-EoR score	**8.65; .003**	**5.28; .022**	**6.18; .013**	3.45; .063
Pro-EoR*VT interaction	2.18; .140	1.67; .196	2.40; .121	2.09; .148
Age	**16.44; ~0**	**15.00; ~0**	**14.82; ~0**	**13.39; ~0**
Gender	1.74; .187	1.99; .158	2.74; .098	3.36; .067
Political views	0.45; .504	0.66; .417	0.29; .591	0.40; .527
Religiosity	1.14; .285	1.43; .232	0.66; .417	1.10; .294

Table 8. Results from inclusive models of guilt using half scale items.
Formula: (JD ~ pro-EoR*VT + age + gender + politics + religiosity)

Several clear conclusions emerge from these analyses. First, respondents were invariably susceptible to the switch in characters between vignette types. This upholds the conclusion from the first model discussed: everyone is more eager to convict the putative bad guy (that is, until they are forced to think about how they would respond if the roles were reversed, as Follow-Up Study 1 showed). The following data visualization helps illustrate this result:

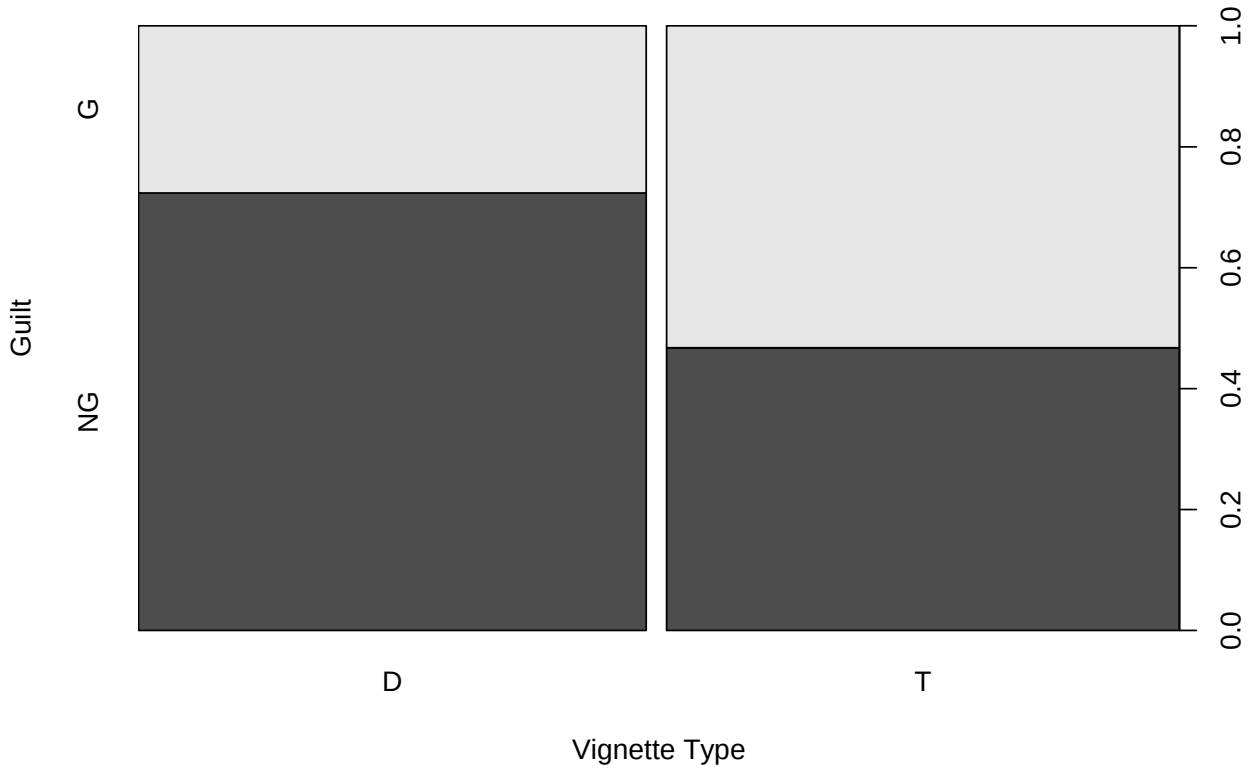

Figure 3. Graph of juror decisions by vignette type.
Full data set (*n* = 800). Darkened areas (labeled with NG) denote "not guilty," while lighter areas (labeled with G) denote "guilty." The DV condition is denoted by D, and the SV condition by T.

The second conclusion is that while endorsement of retribution seems to influence likelihood of voting guilty, this influence is weak and borderline statistically insignificant even in the best of cases. Paradoxically, the EoR score / juror decision relationship lapses

into non-significance as the data set is winnowed down to exclude respondents with legal education and/or those who are ineligible to serve on juries (or are unsure of their eligibility). The effect of legal education is particularly counterintuitive: it suggests that the inclusion of people who understand the workings of the law skews the data in favour of an effect whereby endorsement of retribution manifests in the form of a guilty vote. In general, the association between retributive sentiments and *mens rea* attribution was more significant when the analysis was restricted to the pro-retribution scale items (pro-EoR score); this makes sense considering the higher internal consistency scores for those items, discussed in Chapter I. Taken as a whole, the results for the retribution variables support the prior decision to treat the pro-retribution scale items as a more reliable indicator of the attitude I intended to track in these studies.

Overall, the data do not support the conclusion that endorsement of retribution operates as a factor in people's decisions about criminal guilt. These results do not rule the effect out, but if it exists, its impact is neither drastic nor consistent. The box plots below provide some graphical representations of the result. These box plots represent the distribution of responses in a given category, with the mean response indicated by the central bold line and the spread of the response distribution indicated by dotted lines.

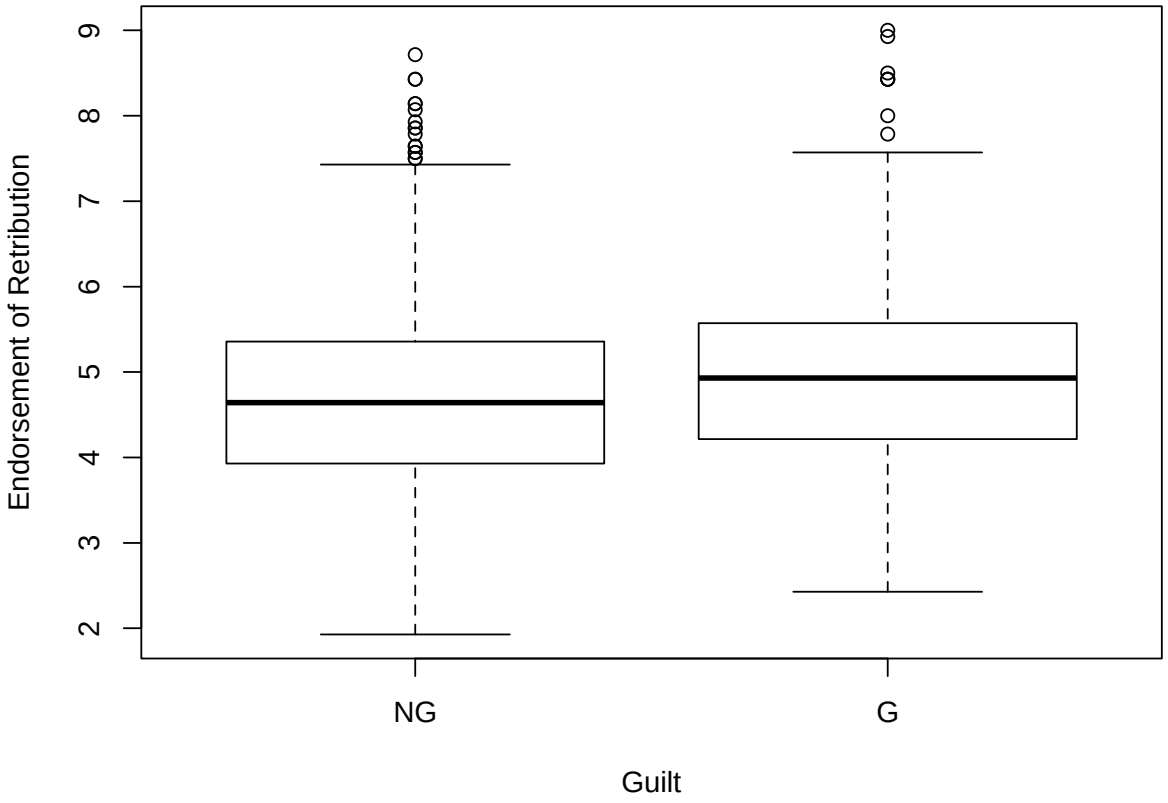

Figure 4. Box plot of Endorsement of Retribution scores.
Sorted by those who voted not guilty (left), and guilty (right), using the whole data set (*n* = 800).

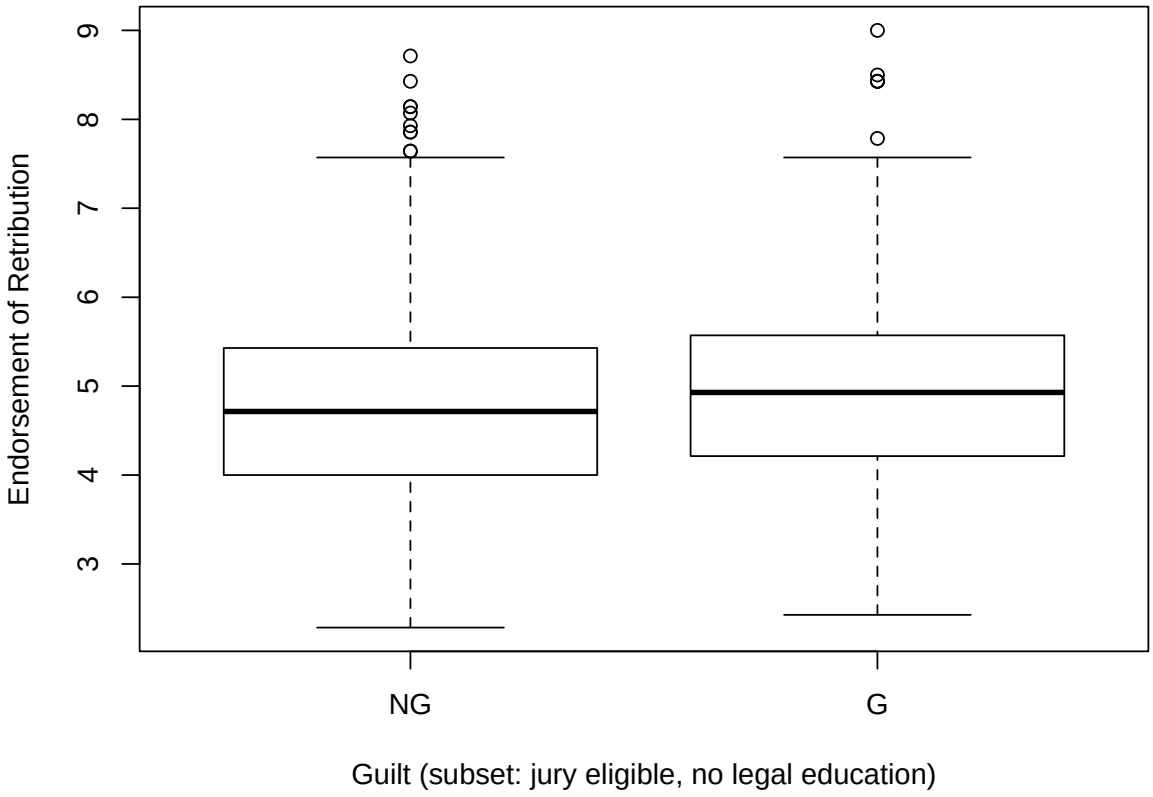

Figure 5. Box plot of Endorsement of Retribution scores for subset of respondents.
Sorted by those who voted not guilty (left), and guilty (right), only including respondents who self-reported eligibility to serve on a jury and no legal experience or education (n = 626).

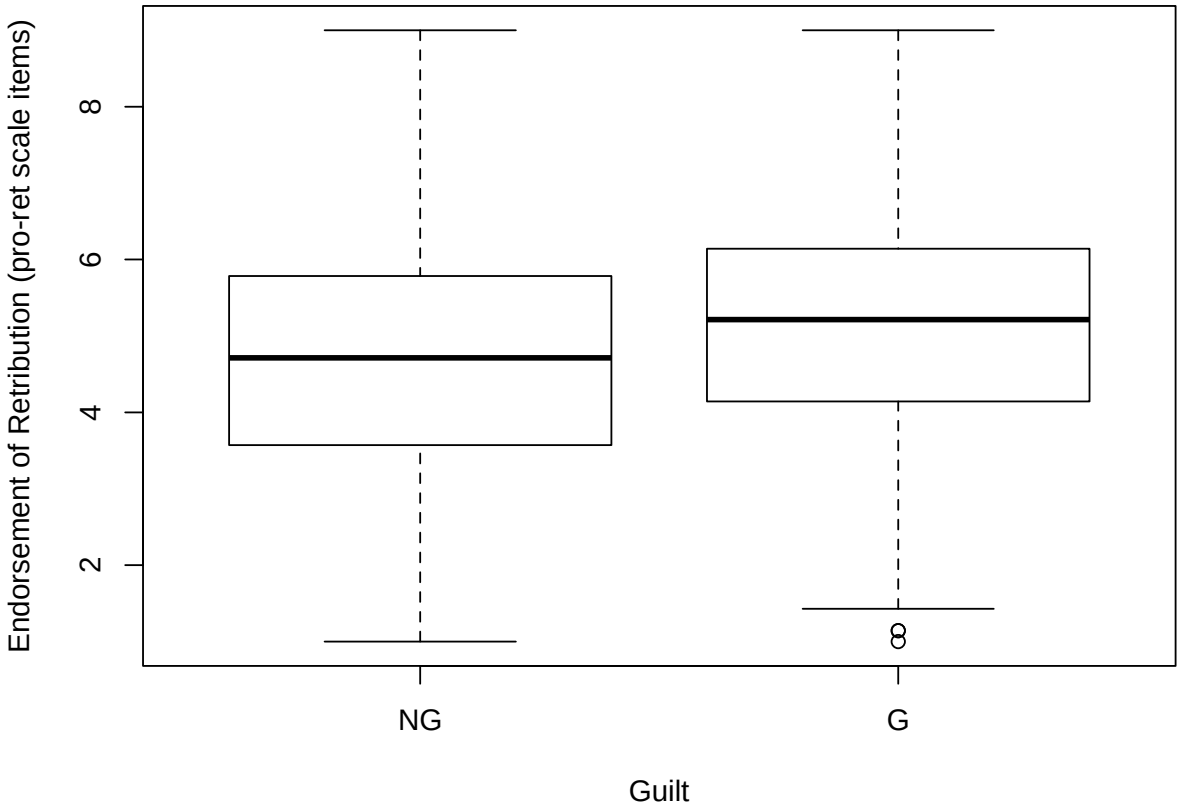

Figure 6. Box plot of pro-EoR score.
Sorted by those who voted not guilty (left), and guilty (right), using the entire data set (*n* = 800).

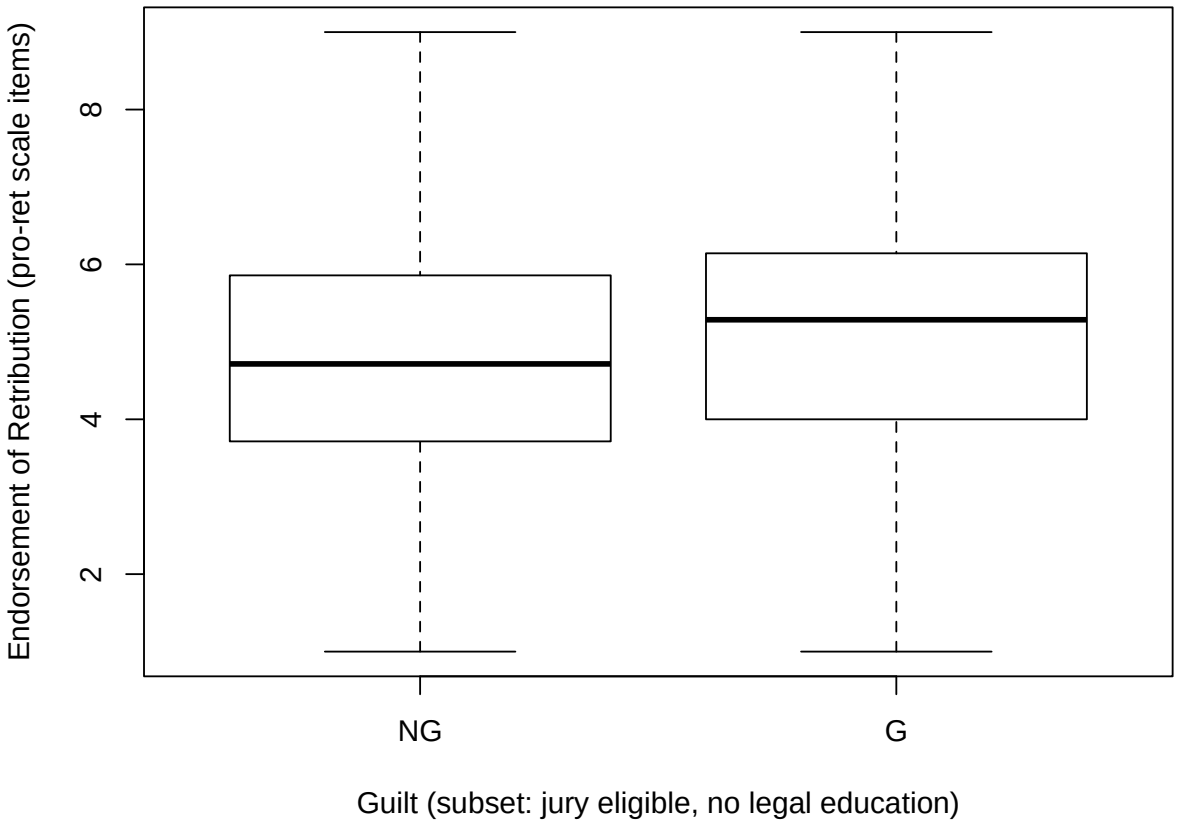

Figure 7. Box plot of pro-EoR score for subset of respondents.
Sorted by those who voted not guilty (left), and guilty (right), only including respondents who self-reported eligibility to serve on a jury and no legal experience or education (n = 626).

None of these interaction effects approached significance. As such, the hypothesis that highly retributive respondents would be disproportionately more susceptible to the effect of vignette type on guilt perception was conclusively nullified.

The lack of notable differences between the simple and inclusive models for the data indicate that these conclusions are robust when controlling for the effects of age, gender, political views, and religiosity. Although respondent age exhibited a strong effect on likelihood of voting guilty, no other demographic variables exerted consistent or significant influences on the outcome variable. In order to double-check the correlation between the

outcome variable and control variables, I ran a binomial linear regression modelling the variable for juror decision (JD) against the demographic items, this time excluding VT, EoR score, and their interaction. The only significant relationship in this model involved the variable age – a chi^2 value of 13.94 with $p = .0002$; the other three variables exhibited small effect sizes at $p > .1$ (notably, this included the relationship between political views and JD). While this association between age and JD was predictable from previous results, it is peculiar in terms of theoretical expectations. Why would age have a systematic effect on tendency to convict? Yet this effect emerged strongly across all of the models. Below is a visualization of this trend:

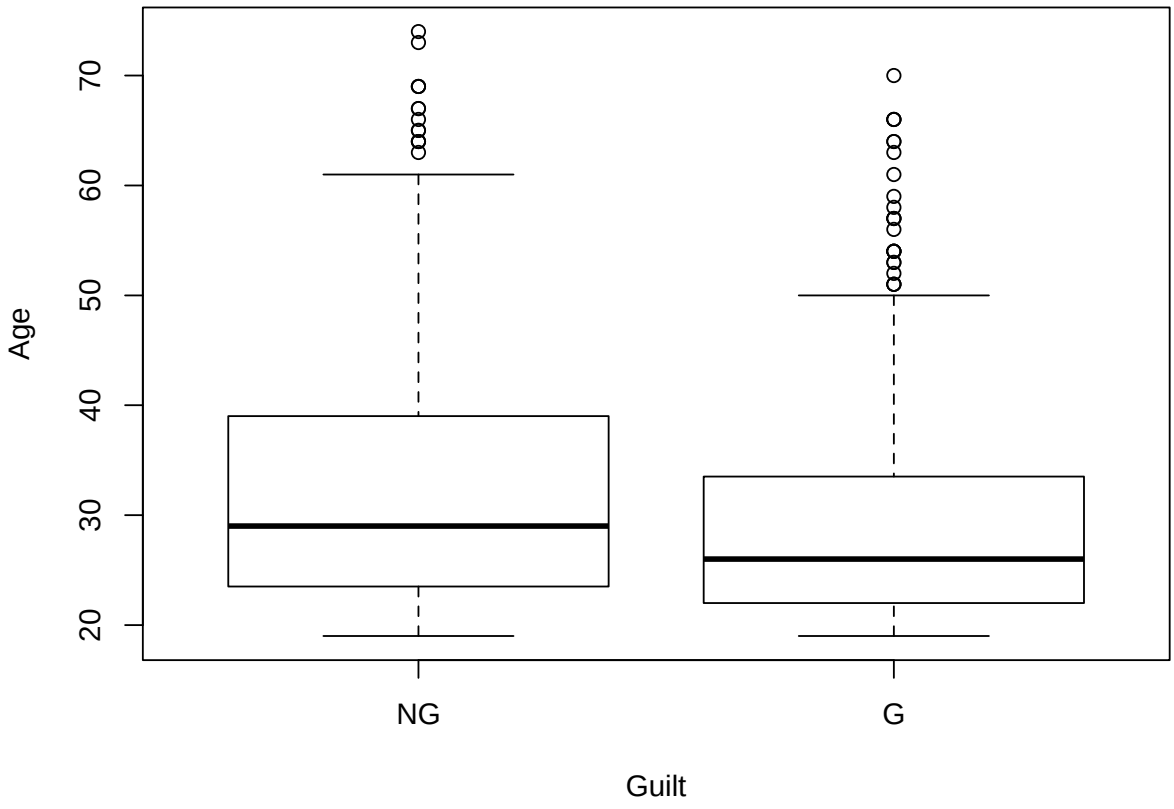

Figure 8. Box plot of age.
Sorted by those who voted not guilty (left), and guilty (right), using the entire data set (n = 800).

Multicollinearity Issues

Intuitively, it seemed likely that some correlations would emerge between a few of the dependent variables and the demographic variables – especially between EoR score and political views. In order to test for this, I created several models using simple linear regression, with EoR score as the outcome variable, and various subsets of the demographic items as dependent variables.

I note before listing the results that the use of linear regression for these models is by itself problematic, because the ordinary least squares method used in linear regression

models requires a normal distribution for all variables in order for the coefficient in the model to serve as the best linear unbiased estimator of the slope of the effect. This assumption is not met in these models due to the markedly non-normal distribution of the responses for political views and religiosity. Hence, the results from these models are not intrinsically trustworthy in isolation. However, when combined with scatter plots depicting the distributions of individual variables, the significant results from these linear models are useful in a confirmatory role. All of the regressions and figures below use the entire data set, except in the second and third tables, where 7 respondents who declined to specify their gender significantly skewed the results and were temporarily omitted.

	Age	Gender	Political Views	Religiosity
Coefficient	**0.02**	-0.12	**0.20**	0.01
p-value	**~0**	.158	**~0**	.736

Table 9. Regression of Endorsement of Retribution score on demographic variables.

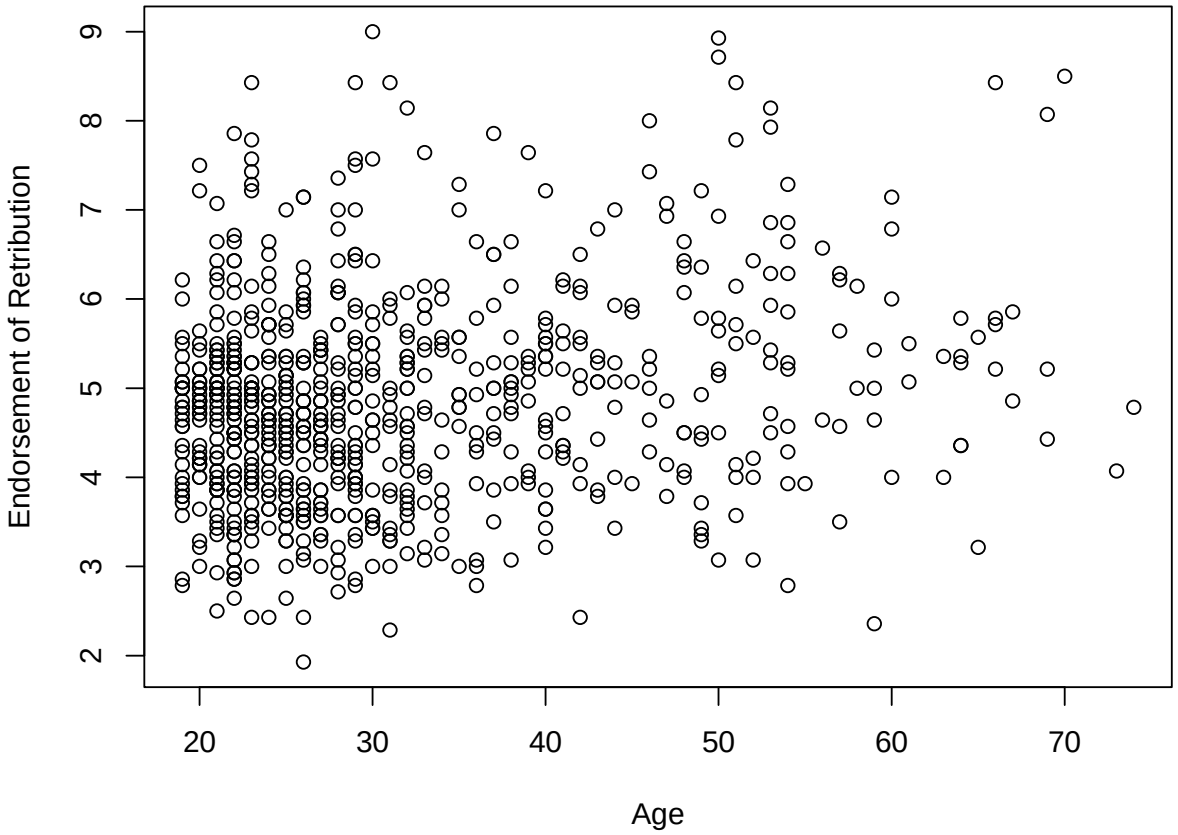

Figure 9. Scatter plot of age vs. Endorsement of Retribution.
Note the positive trend whereby older respondents tend to endorse retribution more.

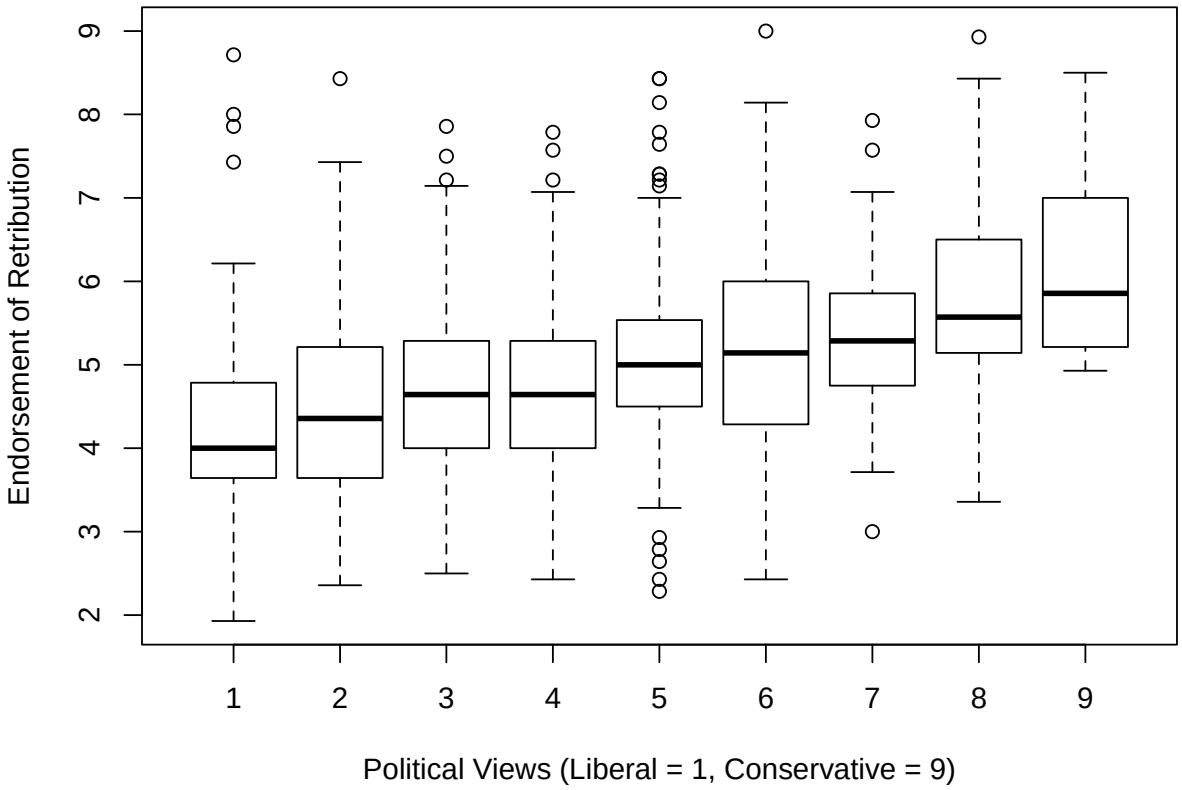

Figure 10. Box plot of political views vs. Endorsement of Retribution.
Note the positive trend indicated by the rising mean Endorsement of Retribution score for increasingly conservative respondents.

	Age	Gender	Religiosity
Coefficient	-0.002	**.435**	**.300**
p-value	.739	**.002**	**~0**

Table 10. Linear regression of political views on other demographic variables.

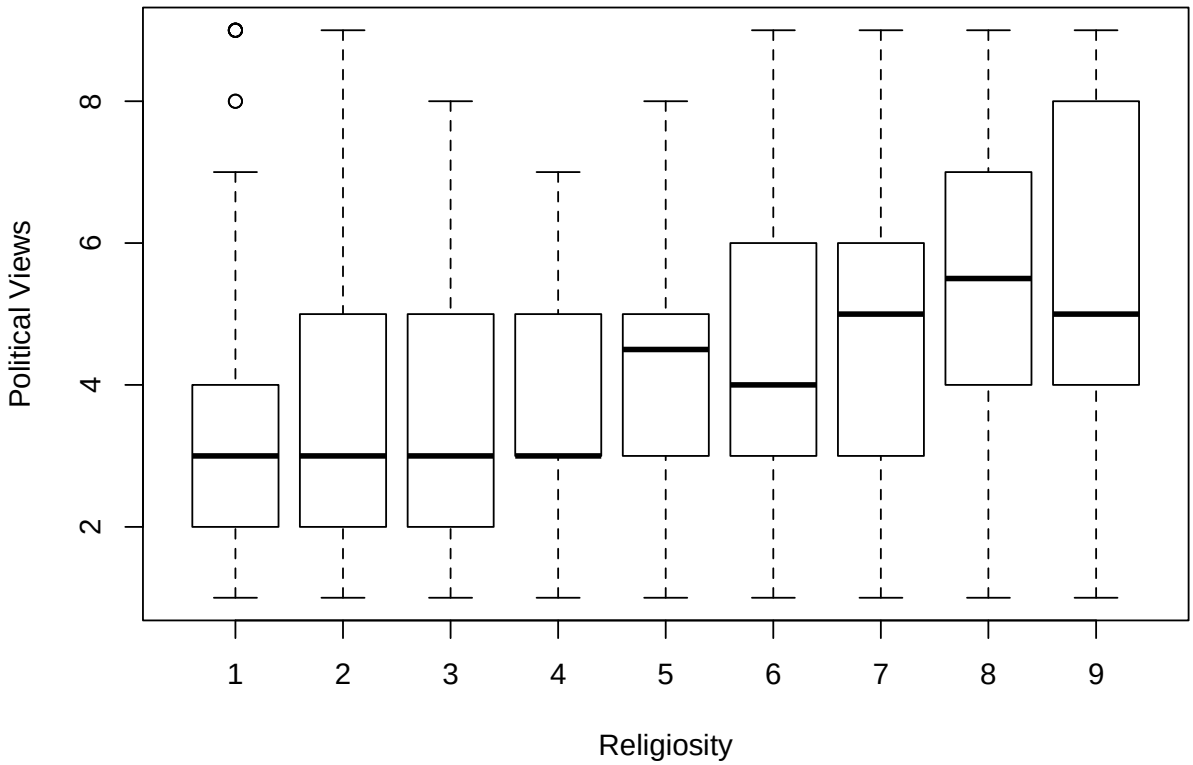

Figure 11. Box plot of religiosity vs. political views.
Note the positive trend indicated by the rising mean scores of political conservatism for increasingly religious respondents.

	Gender
Coefficient	**-2.873**
p-value	**.0005**

Table 11. Linear regression of age on gender.

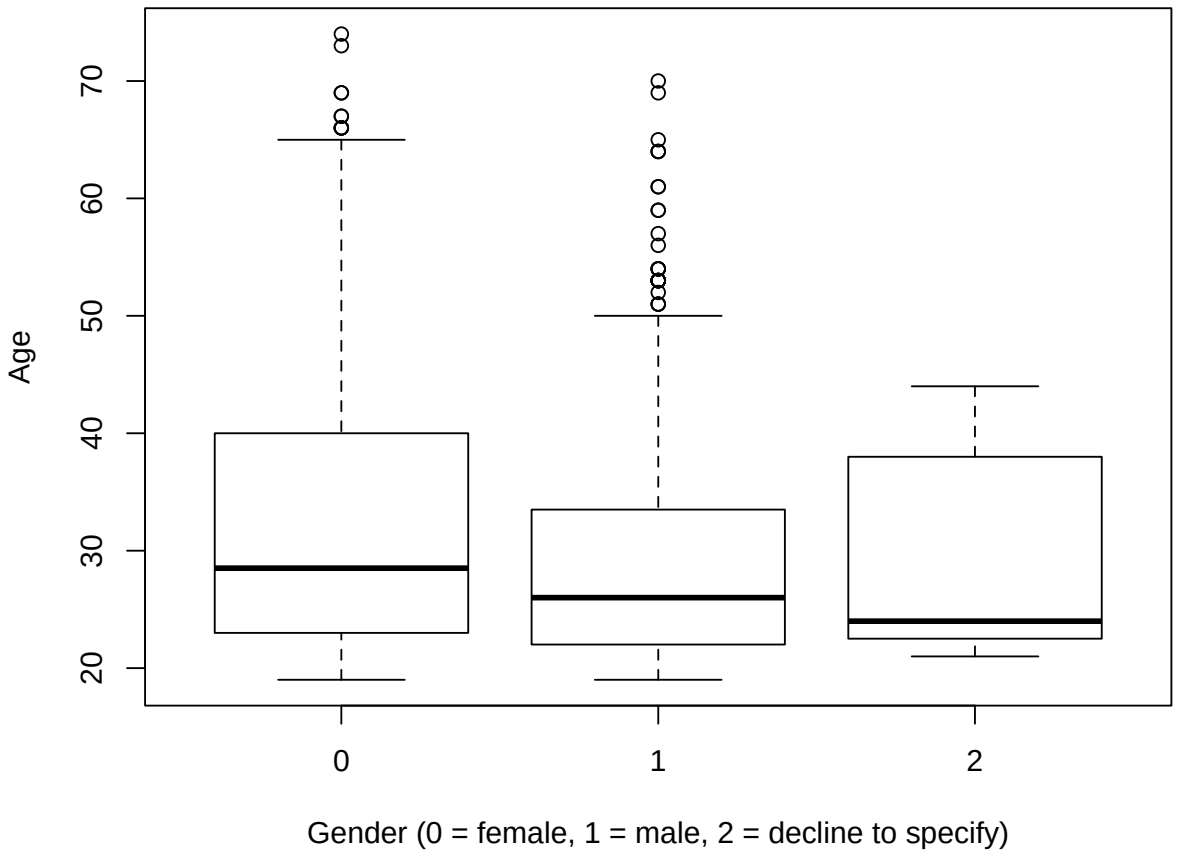

Figure 12. Box plot of age and gender.
Sorted by female respondents (left), male respondents (middle), & those declining to state (right).

The correlations between the variables included in these regressions are problematic for the analyses in the previous section. In general, these findings of multicollinearity diminish the level of confidence that can be placed in the estimates of the relative effects of several variables as given by a model. However, two of the three findings in the previous section are almost certainly not significantly undermined by the multicollinearity. The influence of VT on JD displayed large effect sizes and small p-values, and was highly robust to model re-specification; this finding is thus highly unlikely to have been an artifact of variable collinearity. Similarly, the null-result finding for the interaction effect (i.e., the effect of VT

on JD being differentially stronger depending on EoR score) is not credibly threatened by the fact that the elements of the interaction are collinear with the control variables in the model.

On the other hand, these findings do make the effect of EoR score on juror decision even murkier than previously concluded. The results indicating that endorsement of retribution had a significant effect on attributions of guilt gradually lapsed into insignificance as the models were pared down via subsetting; but the multicollinearity introduces the possibility that the associations between, *inter alia*, age and retribution, or politics and religiosity, created a swamping-out effect for the guilt-retribution relationship that was amplified by the reduction in sample size as the data set was pared down. In other words, it is uncertain whether the insignificance of the guilt-retribution correlation in some of the models from the previous section signals a real effect, or merely represents the multicollinear nature of the other variables surfacing thanks to the smaller *n* in those models.

Results from Follow-Up Study 2

The aim of this follow-up experiment was to double-check whether the conclusions from the main experiment were preserved despite the changes in wording described in the Design & Methods section – particularly concerning the shift from "notorious shoplifter" to "local telemarketer." As it turned out, the trends in this data set (n = 360) were drastically different.

Most strikingly, the difference in conviction rates between the doctor and the shoplifter vignettes disappeared in this version of the study. 36.76% of respondents voted

guilty in the doctor condition, and 29.71% voted guilty in the telemarketer condition; a t-test revealed that the difference in the mean guilty-vote rate between conditions was not statistically significant, at p = .157. The following graph visualizes this result:

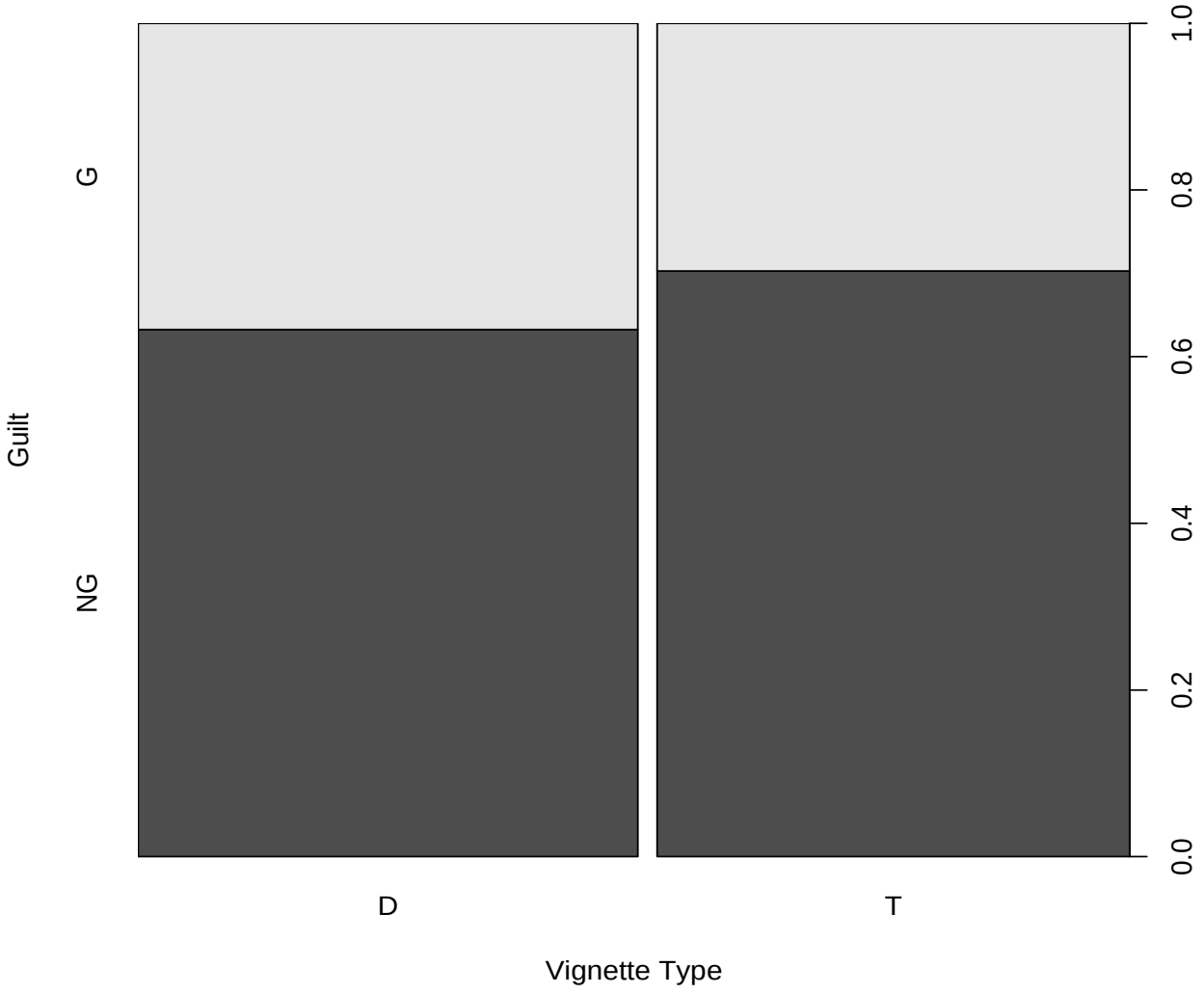

Figure 13. Graph of juror decisions by vignette type, follow-up data set.
Full data set (*n* = 360). Darkened areas (labeled with NG) denote not guilty, while lighter areas (labeled with G) denote guilty. The Doctor Vignette condition is denoted by D; the Telemarketer Vignette condition is denoted by T.

I also ran a binomial GLM modelling JD against VT, EoR score, and their interaction effect. Analysis showed that all of these variables had rather small Chi^2 effect sizes (< 2), and none were statistically significant.

Nor were these effects simply caused by the use of a demographically non-equivalent sample:

Mean (SD)	Age	Gender	Political Views	Religiosity	Retribution
Main Data Set	31.38 (11.6)	.48	3.95 (2.09)	3.80 (2.88)	4.84 (1.16)
Follow-Up 2	32.44 (12.1)	.46	3.90 (2.09)	3.42 (2.86)	4.83 (1.15)

Table 12. Demographic variables in the main and Follow-Up 2 data sets.
Listed as mean and standard deviation (where applicable).

These striking null results obviated the need for further analysis; even if a particular model re-specification were to yield a positive result, the likelihood of its being a veridical effect and not a fluke would be extremely low against the backdrop of these results.

Reasons for Voting Not Guilty

As a tertiary investigation to further contextualize these results and provide a point of departure for researchers interested in exploring further, I provide the results from the survey item that queried those who voted NG for their reasoning as to which element of the guilt standard was not met in the case. The standard of guilt that respondents were instructed to use was: "to be considered guilty of homicide, the defendant must have intentionally and knowingly brought about the death of the victim."

The results from the main data set (n = 800) indicate that most participants couched their rationale in a claim about intention:

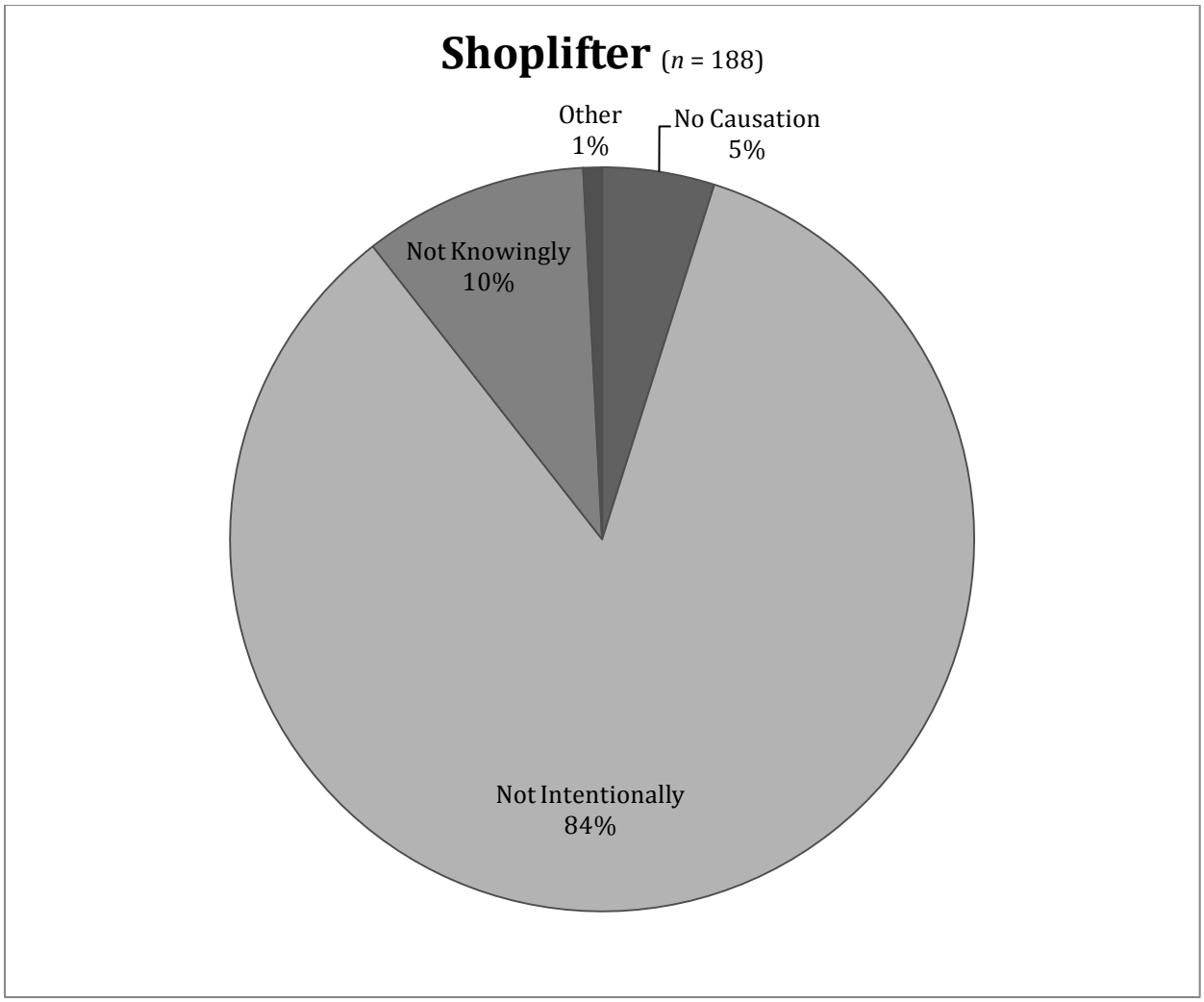

Figure 14. Reasoning for not guilty votes, shoplifter case, main data set.

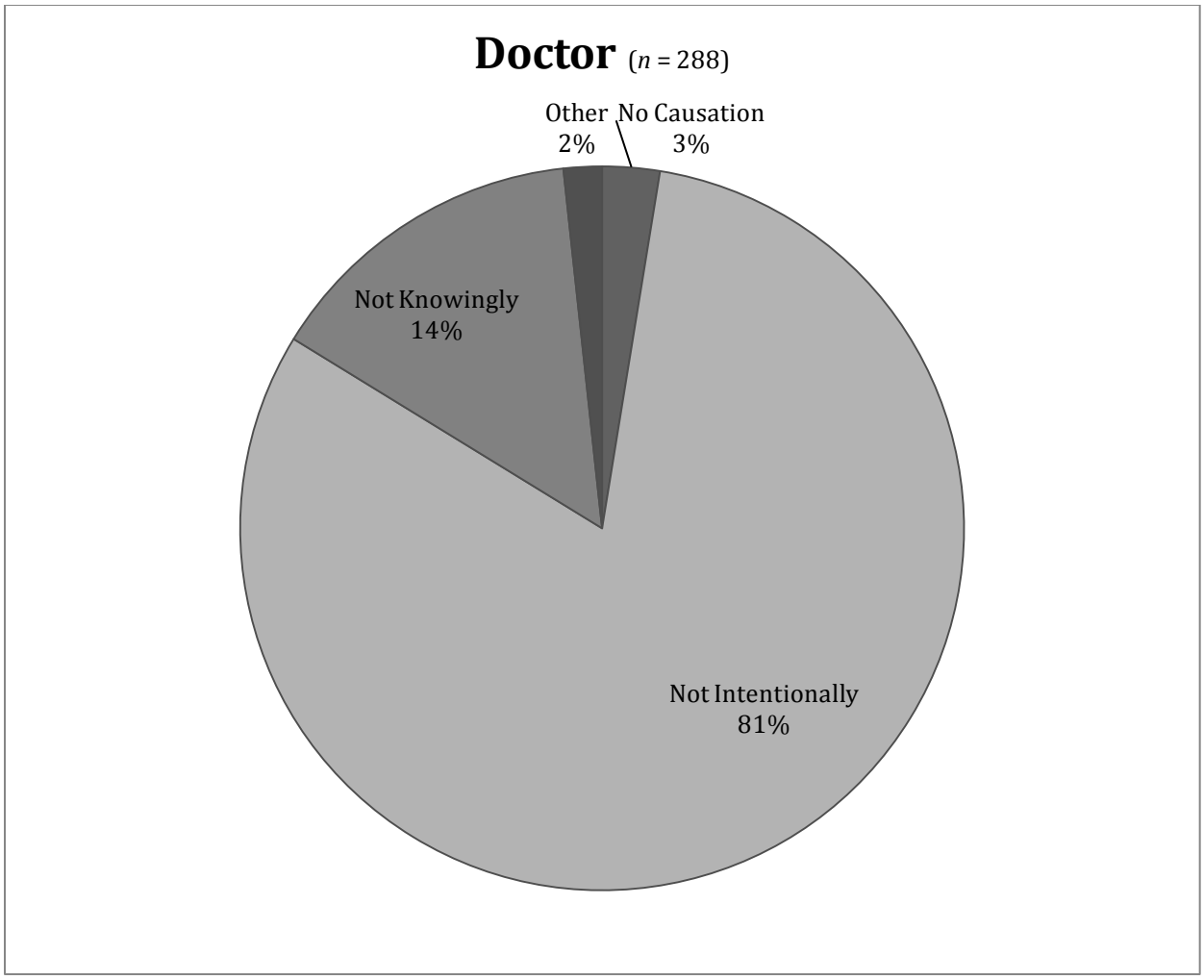

Figure 15. Reasoning for not guilty votes, doctor case, main data set.

As these charts make evident, no appreciable difference emerged between the DV and SV conditions. Most of the responses in the seldom-chosen "Other – please specify" category involved respondents directly disagreeing with the instructions (e.g., criticizing the directive to set aside concerns about justifiable homicide), suggesting that these responses were generally not indicative of any incompleteness in the array of answer choices[15].

[15] Two astute respondents did, however, explain their selection of this answer choice as a way of invoking their right to jury nullification.

A similar breakdown of responses was observed in the data for Follow-Up Experiment 2:

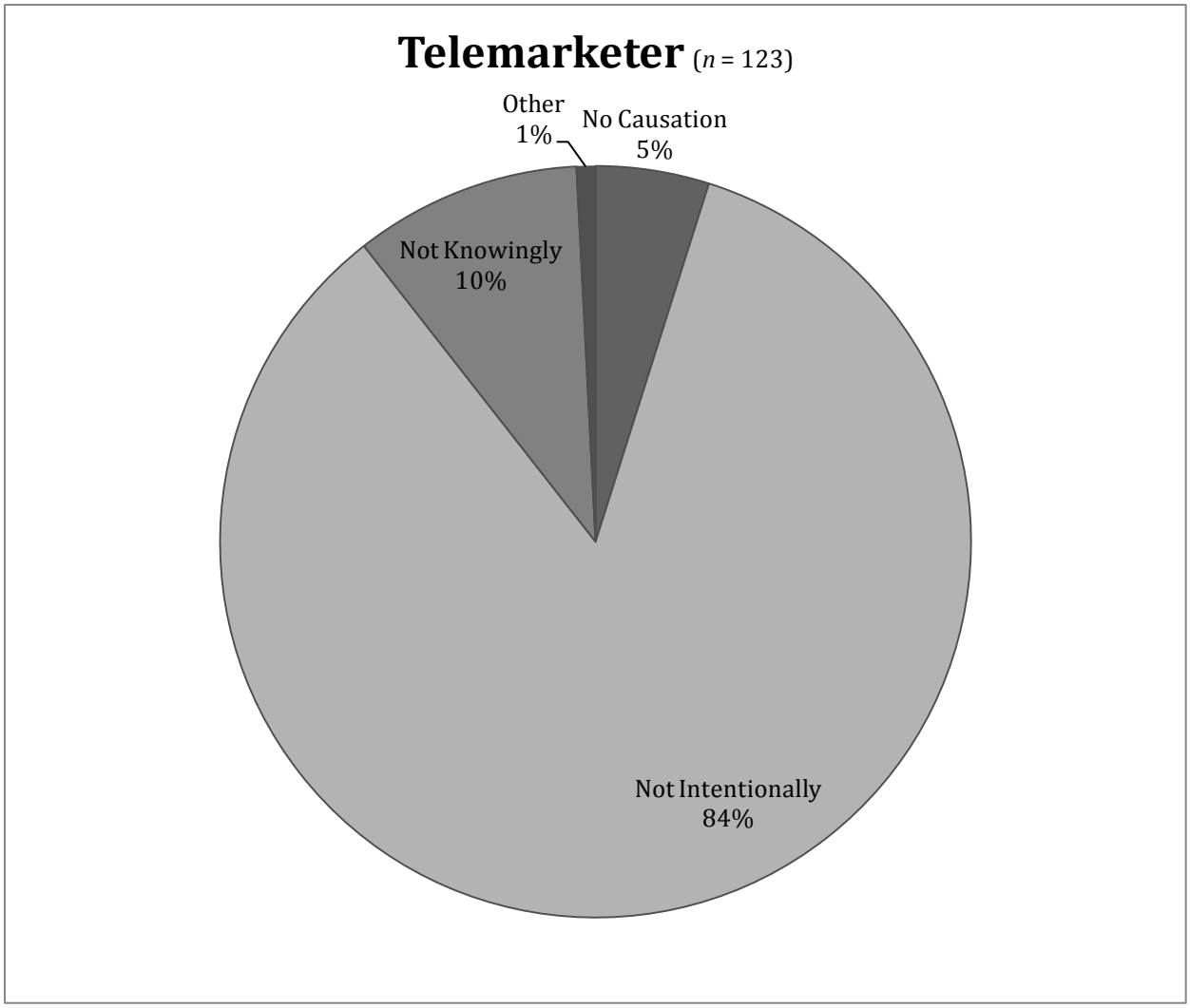

Figure 16. Reasoning for not guilty votes, telemarketer case, follow-up data set.

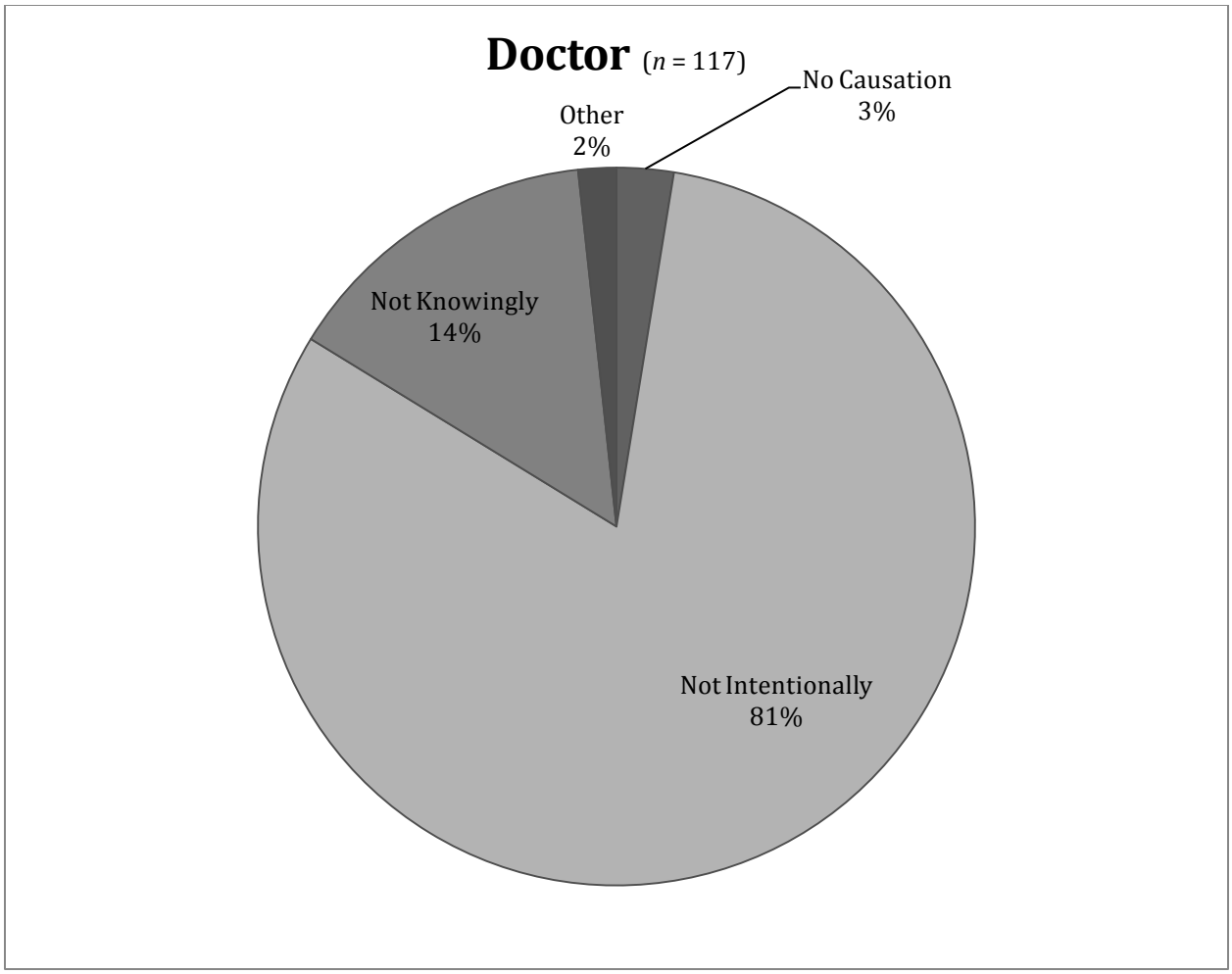

Figure 17. Reasoning for not guilty votes, doctor case, follow-up data set.

These results may serve as preliminary evidence that the concept of intention is more easily modulated by moral considerations than the concepts of knowledge and causation.

Discussion of Results

Taken as a whole, the results establish two key findings and leave one important question of interest unsettled. My specific hypothesis was that respondents who exhibited high endorsement of retribution would be comparatively more likely to vote the shoplifter guilty and less likely to vote the doctor guilty, whereas respondents with low endorsement of retribution would cast votes consistently across cases. This hypothesis was conclusively

nullified. Meanwhile, a hypothesis analogous to the one tested in Thomas Nadelhoffer's 2006 study – that respondents would be generally more likely to vote a morally suspect individual guilty than a morally upstanding one – was upheld, although the scope of its interpretation was constrained. In contrast to the clear and consistent results for these first two avenues of investigation, the data provide little guidance in understanding the general effect of endorsement of retribution on attributions of *mens rea*. Nonetheless, some positive conclusions from the data can be supported, especially insofar as they help identify questions for further research.

Implications for Interpreting Past Studies

In the data set for the main study, participants strongly favoured voting guilty in the case of the shoplifter as compared to the case of the doctor. This directly mirrors Nadelhoffer's finding that people judge the thief in his C1 vignette more blameworthy than the driver in his C2 vignette. In one way, the results presented here strengthen his conclusions: the vignettes in this study were less discrepant than those in Nadelhoffer's, thereby reducing the viability of any argument to the effect that the disparity in blameworthiness / guilt reflects a defensible and justifiable considered judgment rather than a cognitive bias in the same vein as the Knobe Effect. The results from FS1 also support this interpretation: once respondents were confronted with a very similar vignette that switched the positions of the characters, they generally kept their guilt votes consistent, telegraphing that they would consider a discrepancy between the verdicts in the two cases to be unjustified.

These results also bolster Nadelhoffer's conclusion that the change in vignette characters (and hence in moral valence of the cases) prompts a shift in how people apply concepts of knowledge and intentional action. Since the studies presented here were designed to link guilt to judgments of intention and knowledge, the results concerning the between-groups conviction rate do double duty to support this conclusion as well. Further mirroring Nadelhoffer's results are my own data tracking the rationales offered by respondents who voted "not guilty" – a large majority highlighted "intent" as the key domain in which the facts of the case failed to match up to the standard of guilt, consonant with Nadelhoffer's finding that participants were generally more willing to say that the driver had brought about the death intentionally than they were to say it was brought about knowingly (across all vignettes).

Despite all of the ways in which my findings align with Nadelhoffer's, the results from FS2 suggest that the scope of experiments like these is narrower than one might be inclined to assume. Simply by virtue of replacing an extremely unsavory character (the shoplifter) with a mildly unsavory one (the telemarketer), the entire array of effects noted in the main study vanished. If the moral valence of a scenario is supposed to generate predictable biases in the attribution of mental states like intention and knowledge, the natural expectation is for this biasing effect to be robust across a variety of scenarios. Instead, it would appear that the effects Nadelhoffer uncovered are markedly sensitive to the delicate specifics of the scenario deployed in the studies. Experimental philosophers would be well advised to take note of how this conclusion underlines the deceptively high stakes of vignette composition.

The discrepancy between the findings of Follow-Up Study 2 and the main experiment may trace back to the general social-psychological dynamics of outgroup "infrahumanization" (Leyens, Rodriguez-Perez, Rodriguez-Torres, Gaunt, Paladino, Vaes, & Demoulin, 2001) . As noted in Chapter 1, criminals are probably much easier to conceive of as a foreign Other – a neatly separable outgroup – than telemarketers. The latter are noisome and reviled (whether fairly or not), but they are still "us" and likely do not elicit feelings of hatred or indignation in the same way that lawbreakers can. Nonetheless, this explanation carries with it the implication that not all moral sentiments drive Knobe Effects with equal strength – a conclusion which, to my knowledge, has not previously been discussed or speculated about in the experimental philosophy literature.

Teasing Apart the Interplay of Endorsement of Retribution and *Mens Rea* Attribution

The experiments in this Chapter were designed not only for exploratory investigation and for comparability to the Nadelhoffer study, but to test the hypothesis that endorsement of retribution is associated with a pronounced susceptibility to the biasing effects of moral judgment (a.k.a. the Knobe Effect) on *mens rea* attribution. The concepts in this hypothesis were operationalized down to three variables in the study design: endorsement of retribution was represented as a numerical variable via the Endorsement of Retribution scale, *mens rea* attribution was tracked by respondents' decision to vote guilty or not guilty (the JD variable), and the biasing effect of moral judgment was supplied by the differences in the two vignettes (the VT variable). The association between endorsement of retribution and vignette type as they impacted the juror decision variable

was captured in models by the interaction effect of the Endorsement of Retribution and vignette type variables.

The statistical analyses performed on the data from these experiments clearly establish that this interaction effect played no significant role in *mens rea* attribution. Despite the limitations of the study, I can report with confidence that the hypothesis it was designed to test is conclusively nullified. What caused the results to turn out this way? Without further data, I can only offer speculative explanations. Perhaps the least revisionary way of accounting for this result involves revisiting the notion of a cognitive blame-validation mode. The hypothesis of this study was motivated by the idea that an individual with strongly retributive attitudes slips more easily into blame-validation mode, and that such a thinking style amplifies the biasing effect of moral considerations on factual judgments. Instead, the effect of a disposition toward blame-validation may have been much more straightforward – namely, an increased willingness to impute blame, no matter the moral valence of the situation. An even more parsimonious alternative possibility is that retributive attitudes simply do not interact with a blame-validating stance whatsoever. The determination of which of these two explanations is more likely depends on what, exactly, the data from these experiments had to say about the relationship between endorsement of retribution and jurors' willingness to convict in general. Unfortunately, this is where the results become murky.

Study Limitations and Yet-to-Be-Settled Interpretations

Even if the data set for this study had proven more tractable it would still be difficult to make any authoritative conclusions about the influence of EoR score on the JD variable.

As reported in the previous section, the statistical significance of this effect varied depending on how the statistical models were set up, with some indicating a borderline significant effect of modest size and others yielding results falling well short of the $p = .05$ threshold for statistical significance. The demographic dynamics of these inconsistencies were a source of further confusion – for instance, it appeared that much of the association between retributive attitudes and willingness to vote guilty was coming from the contingent of the sample that indicated having legal training. Strictly speaking, the lack of an obvious emergent result makes it most advisable to conclude that no effect exists; nonetheless, the results' sensitivity to various statistical modelling decisions and its general near-miss nature render this conclusion equivocal and uncertain.

Further complicating the picture are issues in the data set regarding the collinear interactions of its variables. As described in the previous section, it is plausible that the effect of multicollinearity in this data set is a swamping-out of dynamics that would otherwise have emerged from the statistical models more clearly (though more evidence would be required to positively affirm that this was the case). Although the other conclusions from this study – the null result for the aforementioned interaction effect, and the general biasing effect of the vignette switch – are almost certainly not threatened by the multicollinearity issues, the already-uncertain nature of the relationship between retributivism and *mens rea* was only intensified by these findings.

The discovery of unexpected correlations amidst the control variables in the data set also raises concerns about sample quality. Why should respondent age have been so significant a predictor of several other variables? What would the association between votes for guilt and endorsement of retribution had looked like if the latter had not been so

closely linked to respondents' political views, especially considering that political views were not a significant predictor of voting guilty when modelled in the absence of Endorsement of Retribution score?

It goes almost without saying that a more traditionally established (and hence more expensive and labor-intensive) sampling paradigm would have improved the quality of the resultant data set. Although Amazon Mechanical Turk has convenience in its favour – to say nothing of affordability – it ultimately can only sample as widely as its user-base. While evidence from studies in psychology indeed indicates that this is not a major limitation in terms of psychometric traits, empirical research in a more sociological vein is not thereby exempted from concerns about sampling bias in Mechanical Turk studies. That being said, I do not mean to imply that the relationships between control variables in this study necessarily indicate the data set is problematic. Indeed, it may well be the case that a larger data set with a more robust sampling procedure would have yielded comparable results.

Additionally, some of the evidence for multicollinearity can be viewed as constitutive of ancillary findings. The correlation between political views and Endorsement of Retribution, for instance, is a relationship that makes theoretical sense but was nonetheless useful to confirm empirically. It also suggests that, should neuroethical debate over the role of punishment grow more widespread, the issue has the potential to become a politicized point of culture war contention. Meanwhile, the strong association between likelihood of voting guilty and respondent age is a jarring result that demands further research. In a larger and more representative sample, would younger participants still have voted guilty more often on the whole? If so, how does this jibe with prior scholarship looking at jury statistics, especially in more ecologically valid settings?

Finally, a key limitation of this study is that it cannot provide any evidence for a causal relationship between any of the variables under investigation. The only effect that must be causal (since no third variable could have influenced it) is that of vignette type on likelihood of voting guilty; participants were randomly assigned to vignettes, so the specifics of the vignette must have caused the difference, rather than merely being correlated to it. The other results, though, must be considered correlative unless and until further studies supply evidence that they are causal.

In sum, despite issues in the data set that created inferential challenges, this study achieved several of its objectives. The experiment and analysis produced results that not only spoke clearly and directly to the primary hypothesis, but also shed useful light on previous studies. Although not all lines of inquiry ended in a place of uncontestable resolution, several new research questions were identified along the way, rendering the exercise productive even when inconclusive.

Conclusion

The studies presented here extend and improve upon previous work relevant to neuroethics and neurolaw. They supply a novel investigative instrument in the form of the Endorsement of Retribution scale and provide data that sheds new light on the interplay of *mens rea* attribution and moral considerations. Several avenues for further work now come into focus, as well as opportunities for translating the results into practical application.

Opportunities for Future Studies

Experimental replication of the validity metrics for the Endorsement of Retribution scale will benefit this new construct considerably. In particular, more research is needed to determine whether the scale should include all 14 items, be limited to only the pro-retribution items, or be split into two subscales. The scale might also make a useful addition to the inventory of surveys hosted on yourmorals.org – its inclusion on this platform would create many opportunities to better understand what other social-psychological traits characterize people with high or low endorsement of retribution. Researchers who focus on legal issues could also make use of the scale in understanding the attitudes of judges, jurors, attorneys, plaintiffs, policymakers, law enforcement officers, and many others besides. Additionally, the Endorsement of Retribution scale could be utilized in comparative cross-cultural work, to better understand how different societal systems shape attitudes about retribution and punishment.

The scale also promises to be useful for researchers in neuroethics seeking to better understand the implications of the percolation of cognitive science into folk understanding, and for experimental philosophers to explore the use of concepts related to punishment

and desert. In particular, it would prove relatively simple and likely illuminating to use the Endorsement of Retribution scale in studies aimed at empirically assessing the much-discussed prediction that popular support for retribution is bound to wane (Greene & Cohen, 2004). Do people who subscribe to the notion that "you are your brain" tend to score lower on the Endorsement of Retribution scale? What about people who are more conversant in, or simply more frequently exposed to, neuroscience and its attendant cultural narratives?

The work detailed in Chapter 2 encompasses many opportunities for variations on its study design. Indeed, I hope that these experiments, especially in combination with the innovative and high-impact example set by the work of Shen, Jones, et al. in "Sorting Guilty Minds," will help catalyze further research applying the methods of experimental philosophy to issues in the domain of the law. The experiments from Chapter 2 in particular serve as proof of concept that legal researchers can usefully borrow the study-design and analytic techniques of experimental philosophy – most notably the contrastive vignette technique, which facilitates rigourous empiricism by relating attitudinal differences to fine-grained distinctions in situational details. Although the scenarios deployed in these studies spoke most directly to criminal law, there is no reason future studies could not employ similar methods to root out biases and double standards in people's intuitions regarding, e.g., torts or contracts.

Significance of Study Conclusions

Whereas the product of Chapter 1 – a validated scale – is immediately applicable as an instrument for use in further research, the results from Chapter 2 do not translate

straightforwardly into concrete recommendations for change or action. They do, however, pose a question worth addressing by scholars of criminal law. Contextualizing this question is the finding that, in certain situations, people exhibit a flaw in reasoning whereby they make inferences about an individual's *mens rea* inconsistently, with non-identical verdicts being rendered in pairs of cases that should elicit the same verdict. Although the evidence from the study does not conclusively pinpoint what drives this cognitive glitch, it likely involves implicit moral appraisals of the individuals involved in a given case. I presume that such erratic assignation of criminal guilt is legitimately problematic for the criminal justice system, and hence represents an issue that demands some effort toward corrective action. The difficult question, then, is this: what can be done?

Undoubtedly, this question will become less vexing as the nature of the problem is better understood through more research. Until then, it is challenging to envision solutions that are not radically revisionary; clearly the problem at hand would be rendered moot if society abandoned the practice of trial by jury, but taking aim at a right protected by the US Constitution and Canadian Criminal Code is perhaps not a strategy likely to garner much in the way of serious attention. What is needed is a patch for the functioning of the existing systems.

The best way to fashion a recommendation in this case is to heed the lessons of the data – in particular the difference between the main study and the first follow-up experiment. I noted, in comparing the results of these two iterations, that the inconsistency in *mens rea* attribution observed in the main paradigm is almost entirely swept away when participants are given the opportunity to make a comparative judgment. Perhaps, then, society's best hope of addressing this particular breed of juror bias lies in prompting actual

jurors to consider variations on the cases they are asked to decide in which key details about the identities of the *dramatis personae* have been switched around or altered.

Of course, people on juries are notoriously impressionable, with entire cases having been brought down on the basis of what the judge did or did not say to jurors (R. v. Nette, 2001). As such, most attempts to interfere with their deliberations will elicit objections. Hence, any interventions of this sort will need to be carefully and selectively piloted and fine-tuned; as is often the case in the law, this is a situation where a poorly conceived change would be far worse than no change at all. While I am not well positioned to offer more concrete recommendations for action, I hope to have illustrated the nature and seriousness of the problem in a way that will motivate those who do possess the requisite know-how to begin applying it, in the pursuit of ever-greater justice.

In summation, the studies I have presented here supply a new tool, dispel a potential worry, and issue a call to action. With the Endorsement of Retribution scale, future researchers will be able to include a valid measurement of an increasingly pertinent social attitude in empirical investigations of neurolaw-related issues. Thanks to the nullification of my main hypothesis of interest, legal scholars need not fear that jurors' views on the nature of punishment are wreaking systematic havoc on the attribution of *mens rea* and hence of criminal guilt. Nonetheless, in a more general sense, these studies further illustrate the widespread and severe nature of unduly biased reasoning about *mens rea*, and in doing so, they clearly establish both the need and the methodological tools for further research aimed at accurately fathoming and carefully mitigating these foibles of the human mind.

Bibliography

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126,* 556-574. doi:10.1037/0033-2909.126.4.556

Altemeyer, B. (1981). *Right-wing authoritarianism*. Manitoba: University of Manitoba Press.

Bagaric, M., & Amarasekara, K. (2000). The errors of retributivism. *Melbourne University Law Review, 24,* 124-189.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121,* 154-161. doi:10.1016/j.cognition.2011.05.010

Beebe, J., & Buckwalter, W. (2010). The epistemic side-effect effect." *Mind & Language, 25,* 474-498.

Bengson, J., Moffett, M. A., & Wright, J. C. (2009). The folk on knowing how. *Philosophical Studies, 142*, 387-401.

Bergman, R. (2002). Why be moral? A conceptual model from developmental psychology. *Human Development, 45,* 104-124.

Bornstein, R. F. (1996). Face validity in psychological assessment: implications for a unified model of validity. *American Psychologist, 51*, 983-984. doi:10.1037/0003-066X.51.9.983

de Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology, 23*, 43-57. doi:10.1080/09515080903532290

de Brigard, F., Mandelbaum, E., & Ripley, D. (2009). Responsibility and the brain sciences. *Ethical Theory and Moral Practice, 12,* 511-524. doi:10.1007/s10677-008-9143-5

Bronsteen, J. (2009). Retribution's role. *Indiana Law Journal, 84,* 6-37.

Buller, T. (2010). Rationality, responsibility, and brain function. *Cambridge Quarterly of Healthcare Ethics, 19*, 196-204. doi:10.1017/S0963180109990466

Carlsmith, K. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*, 437-451.

Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing / allowing judgments. *Cognition, 108*, 281-289. doi:10.1016/j.cognition.2008.02.005

Dalbert, C., Montada, L., & Schmitt, M. (1987). Glaube an eine gerechte Welt als Motiv: Validierungskorrelate zweier Skalen [Belief in a just world motive: validity correlation of the scale]. *Psychologische Beitrage, 29*, 596-615.

Dingwall, G. (2008). Deserting desert? Locating the present role of retributivism in the sentencing of adult offenders. *The Howard Journal of Criminal Justice 47*, 400-410. doi:10.1111/j.1468-2311.2008.00529.x

Dolinko, D. (2003). Restorative justice and the justification of punishment. *Utah Law Review, 2003*, 319-342.

Erickson, S. K. (2009). Blaming the brain. *Minnesota Journal of Law, Science & Technology, 11,* 27-77.

Frase, R. S. (2004). "Limiting retributivism: the consensus model of criminal punishment." In M. Tonry (Ed.), *The Future of Imprisonment* (pp. 83-119). Oxford, UK: Oxford University Press.

Gazzaniga, M. S. (2008). The law and neuroscience. *Neuron, 60,* 412-415. doi:10.1016/j.neuron.2008.10.022

Grau, E. (2011) Using factor analysis and Cronbach's alpha to ascertain relationships between questions of a dietary behaivor questionnaire. *Section on Survey Research Methods*, 3104-3110.

Greely, H. (2007) Neuroscience and criminal justice: not responsibility but treatment. *University of Kansas Law Review, 56,* 1103-1138.

Greene, J. D., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London: Biological Sciences 359*, 1775-1785. doi:10.1098/rstb.2004.1546

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: intentionality judgments and the moral valence of action. *Cognition, 117,* 139-150. doi:10.1016/j.cognition.2010.08.002

Haidt, J. & Graham, J. (2007). When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20,* 98-116. doi:10.1007/s11211-007-0034-z

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: implications for improving face validity of measures of unobservable constructs. *Journal of Business Research, 57,* 98-107. doi:10.1016/S0148-2963(01)00295-8

Huigens, K. (2005). On commonplace punishment theory. *University of Chicago Legal Forum, 2005,* 437-458.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. NYU Working Paper No. CEDER-10-01.

Johnson, D. D. P. (2005). God's punishment and public goods. *Human Nature, 16*, 410-446.

Johnson, J. (2008). Revisiting Kantian retributivism to construct a justification of punishment. *Criminal Law and Philosophy, 2,* 291-307. doi:10.1007/s11572-008-9052-7

Koleva, S., Graham, J., Haidt, J., Iyer, R., Ditto, P. H. (2009). The ties that bind: how five moral concerns organize and explain political attitudes. Unpublished manuscript, available at http://faculty.virginia.edu/haidtlab/articles/manuscripts/koleva.graham.submitted.ties-that-bind.pub604.doc (last retrieved 6 August 2012)

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190-194.

Knobe, J., & Fraser, B. (2008). "Causal judgment and moral judgment: two experiments." In W. Sinnott-Armstrong (Ed.), *Moral Psychology: Vol. 2. The Cognitive Science of Morality: Intuition and Diversity* (pp. 441-448). Cambridge, MA: MIT Press.

Leyens, J. P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M. P., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology, 31,* 395-411. doi:10.1002/ejsp.50

Lipkus, I. (1991). The construction and preliminary validation of a global belief in a just world scale and the exploratory analysis of the multidimensional belief in a just world scale. *Personality and Individual Differences*, *12,* 1171-1178.

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture, 6*, 87-112.

Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: the tension between folk and legal concepts of intentionality. *Behavioral Sciences and the Law, 21,* 563-580. doi:10.1002/bsl.554

Messick, S., & Jackson, D. N. (1961). Acquiescence and the factorial interpretation of the MMPI. *Psychological Bulletin, 58*, 299-304.

Mobbs, D., Lau, H. C., Jones, O. D., Frith, C. D. (2007). Law, responsibility, and the brain. *PLoS Biology*, *5,* 693-700. doi:10.1371/journal.pbio.0050103

Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology, 18*, 341-352. doi:10.1080/09515080500177309

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: some problems for juror impartiality. *Philosophical Explorations, 9,* 203-219. doi:10.1080/13869790600641905

Nahmias, E. (2006). Folk fears about freedom and responsibility: determinism vs. reductionism. *Journal of Cognition and Culture, 6,* 215-237.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22,* 287-293.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Nous*, *41,* 663-685.

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

Paulhus, D. L., & Carey, J. M. (2011). The FAD-Plus: Measuring Lay Beliefs Regarding Free Will and Related Constructs. *Journal of Personality Assessment*, *93*, 96-104. doi: 10.1080/00223891.2010.528483

Pizarro, D. A. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30,* 355-375.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: a personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*, 741-763. doi:10.1037/0022-3514.67.4.741

R. v. Nette, 78 Supreme Court of Canada. (2001) 3 SCR 488, retrieved from CanLII database: http://canlii.ca/en/ca/scc/doc/2001/2001scc78/2001scc78.html (last retrieved 6 August 2012)

Robinson, P. H. (2008). Competing conceptions of modern desert: vengeful, deontological, and empirical. *Cambridge Law Journal, 67*, 145-175. doi:10.1017/S000819730800010X

Robinson, P. H., & Darley, J. M. (1996). The utility of desert. *Northwestern University Law Review, 91,* 453-499.

Roskies, A. L., & Nichols, S. (2008). Bringing moral responsibility down to earth. *The Journal of Philosophy, 105*, 371-388.

Rubin, E. (2003) Just say no to retribution. *Buffalo Criminal Law Review, 7*, 17-83.

Rubin, Z., & Peplau, L. A. (1975). Who believes in a just world? *Journal of Social Issues, 31*, 65-89.

Schweitzer, N. J., Saks, M. J., Murphy, E. R., Roskies, A.L., Sinnott-Armstrong, W., & Gaudet, L. M. (2011). Neuroimages as evidence in a mens rea defense: No impact. *Psychology, Public Policy, and Law, 17*, 357-393. doi:10.1037/a0023581

Severance, L. J., Goodman, J., & Loftus, E. F. (1992). Inferring the criminal mind: toward a bridge between legal doctrine and psychological understanding. *Journal of Criminal Justice*, *20,* 107-120.

Shafer-Landau, R. (2000). Retributivism and desert. *Pacific Philosophical Quarterly, 81,* 189-214.

Shen, F. X., Jones, O. D., Hoffman, M. B., Greene, J. D., & Marois, R. (2011) Sorting guilty minds. *New York University Law Review, 86,* 1306-1360.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466-469. doi:10.1038/nature04271

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*, 300-308.

Snead, O. C. (2011). "Cognitive neuroscience and the future of punishment." In J. Rosen & B. Wittes (Eds.), *Constitution 3.0: Freedom and Technological Change* (pp. 130-154). Washington, DC: Brookings Institution Press.

Spanos, A. (2006.) "Econometrics in retrospect and prospect." In T. C. Mills & K. Patterson (Eds.), *New Palgrave Handbook of Econometrics, vol. 1* (pp. 3-60). London, UK: Palgrave Macmillan.

Strauss, A. R. (2001). Losing sight of the utilitarian forest for the retributivist trees: an analysis of the role of public opinion in a utilitarian model of punishment. *Cardozo Law Review, 23*, 1549-1595.

Whitman, J. Q. (2003). A plea against retributivism. *Buffalo Criminal Law Review, 7,* 85-107.

Young, L. & Saxe, R. (2011) When ignorance is no excuse: different roles for intent across moral domains. *Cognition, 120,* 202-214. doi:10.1016/j.cognition.2011.04.005

Appendices

N.B.: in all appendix pages, *italic font* denotes survey logic markup not visible to participants.

A: Demographic Items, All Questionnaires.

Please enter your age (in years).

[ ]

Please indicate your gender identity.

- ○ Female
- ○ Male
- ○ Other / prefer not to say

Please select the category that best describes your highest level of education.

- ○ Some high school education
- ○ High school diploma or equivalent
- ○ Non-university certificate / diploma / degree
- ○ Some university / college education
- ○ Associate degree or equivalent
- ○ Bachelor's degree or equivalent
- ○ Some graduate / professional education
- ○ Master's degree or equivalent
- ○ Doctoral degree or equivalent
- ○ Professional degree or equivalent

Do you consider yourself religious?

	1	2	3	4	5	6	7	8	9	
Not at all	○	○	○	○	○	○	○	○	○	Very much so

Which, if any, religion / spiritual tradition / belief system do you identify with?

- ○ Buddhism
- ○ Christianity
- ○ Hinduism
- ○ Islam
- ○ Judaism
- ○ Humanism
- ○ Sikhism
- ○ Jainism
- ○ Baha'i Faith
- ○ another organized religion not listed here
- ○ no organized religion

We recognize that political views are multifaceted, but sometimes people encounter forced choices between left-leaning and right-leaning policies or politicians. When faced with such a choice in a North American political setting, where on this scale do you tend to fall?

	1	2	3	4	5	6	7	8	9	
Very liberal	○	○	○	○	○	○	○	○	○	Very conservative

Do you have any formal legal education or legal work experience (e.g., attended law school, hold a law-related degree, worked at a law firm or as a paralegal)?

- ○ Yes
- ○ No

To the best of your knowledge, are you eligible or ineligible to serve on a jury?

- ○ Eligible
- ○ I have no idea
- ○ Ineligible

B: Comprehension Checks, All Questionnaires:

Comprehension Check
The following list contains three topics that this survey asked you about, and one topic that did NOT appear in this survey. Please select the topic that this survey did NOT cover.

- ○ Dominance among groups in society
- ○ How and why to punish offenders
- ○ Principles about what is morally right or wrong
- ○ Strategies for managing personal finances

Comprehension Check
Earlier in the survey, you read a story about a criminal case in which a person was holding on to a moving vehicle and eventually lost his grip. What did the driver of the vehicle do that caused this to happen?

- ○ The driver came to a sudden stop.
- ○ The driver steered in a zig-zag fashion.
- ○ The driver opened and shut his door forcefully.
- ○ The driver fired a gun at the individual.

C: Vignettes and Questions, Chapter 2 – Main Study

Scenario *(shoplifter version)*

Please read the following paragraphs carefully. Their details will be important later, and you will not be able to browse back to check the text again. Suppose that you are sitting on a jury for a criminal case. The facts of the case are as follows: A notorious shoplifter is driving his car, and stops at a red light. A local doctor approaches the vehicle; the doctor recognizes the shoplifter, recalling a recent alert on the news. Although the doctor is clearly unarmed, the shoplifter thinks the doctor may try to apprehend him, and so he panics and speeds off through the intersection. Amazingly, the doctor manages to hold on to the side of the car as it speeds off. The shoplifter swerves in a zigzag fashion in the hopes of escaping — understanding that doing so may place the doctor in grave danger. But the shoplifter doesn't care; he just wants to get away. Unfortunately for the doctor, the shoplifter's attempt to shake him off is successful. As a result, the doctor rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

Juror Task

The driver is now on trial for homicide. As a juror on the case, you must determine whether he is guilty or not guilty. Before you begin to deliberate, the judge says the following to you and the rest of the jury: "Some of the commentators on this case have been talking about whether the driver's actions were justifiable, considering the circumstances. But I must remind you that, according to the laws of our region, the question of whether a homicide is justifiable may only be taken up during sentencing, and so it is never the job of the jury to consider this aspect. Your task is strictly to determine whether the driver's actions count as homicide according to the region's legal definition, whether justified or not." As you go to deliberate, you glance over some reference material provided to you for the case. You note that in the region this court case is being decided, the law is as follows: "to be considered guilty of homicide, the defendant must have intentionally and knowingly brought about the death of the victim." Thinking this through, you remind yourself that if the facts of this case meet the standard in that definition, then the driver is guilty. If the standard is not met, then the driver is not guilty. Taking all of this into consideration, what is your vote?

- ○ Guilty
- ○ Not guilty

(if:guilt=0) Which part of the standard for homicide did the facts of the case fail to meet?

- ○ The defendant did not knowingly bring about the death of the victim.
- ○ The defendant did not intentionally bring about the death of the victim.
- ○ The defendant did not cause the death of the victim in the first place.
- ○ Other, please specify: _____

Scenario *(doctor version)*

Please read the following paragraphs carefully. Their details will be important later, and you will not be able to browse back to check the text again. Suppose that you are sitting on a jury for a criminal case. The facts of the case are as follows: A local doctor is driving his car, and stops at a red light. A notorious shoplifter approaches the vehicle; the doctor recognizes the shoplifter, recalling a recent alert on the news. Although the shoplifter is clearly unarmed, the doctor thinks the shoplifter may try to carjack him, and so he panics and speeds off through the intersection. Amazingly, the shoplifter manages to hold on to the side of the car as it speeds off. The doctor swerves in a zigzag fashion in the hopes of escaping — understanding that doing so may place the shoplifter in grave danger. But the doctor doesn't care; he just wants to get away. Unfortunately for the shoplifter, the doctor's attempt to shake him off is successful. As a result, the shoplifter rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

Juror Task

The driver is now on trial for homicide. As a juror on the case, you must determine whether he is guilty or not guilty. Before you begin to deliberate, the judge says the following to you and the rest of the jury: "Some of the commentators on this case have been talking about whether the driver's actions were justifiable, considering the circumstances. But I must remind you that, according to the laws of our region, the question of whether a homicide is justifiable may only be taken up during sentencing, and so it is never the job of the jury to consider this aspect. Your task is strictly to determine whether the driver's actions count as homicide according to the region's legal definition, whether justified or not." As you go to deliberate, you glance over some reference material provided to you for the case. You note that in the region this court case is being decided, the law is as follows: "to be considered guilty of homicide, the defendant must have intentionally and knowingly brought about the death of the victim." Thinking this through, you remind yourself that if the facts of this case meet the standard in that definition, then the driver is guilty. If the standard is not met, then the driver is not guilty. Taking all of this into consideration, what is your vote?

- ○ Guilty
- ○ Not guilty

(if:guilt=0) Which part of the standard for homicide did the facts of the case fail to meet?

- ○ The defendant did not knowingly bring about the death of the victim.
- ○ The defendant did not intentionally bring about the death of the victim.
- ○ The defendant did not cause the death of the victim in the first place.
- ○ Other, please specify: _____

D: Vignettes, Follow-Up Study 2

(telemarketer version)

Please read the following paragraphs carefully. Their details will be important later, and you will not be able to browse back to check the text again. Suppose that you are sitting on a jury for a criminal case. The facts of the case are as follows: Smith, a telemarketer by profession, is driving his car, and stops at a red light. Jones, a local doctor, stands on the sidewalk nearby. As it happens, the two look very much alike. In a remarkable coincidence of mistaken identity, each man mistakes the other for a notorious car thief whose photograph has frequently been broadcast on the local news, and who happens to resemble them both. The doctor begins to approach the vehicle, hoping to apprehend its driver. Although the doctor is clearly unarmed, the telemarketer thinks the doctor may try to carjack him, and so he panics and speeds off through the intersection. Amazingly, the doctor manages to hold on to the side of the car as it speeds off. The telemarketer swerves in a zigzag fashion in the hopes of escaping — understanding that doing so may place the doctor in grave danger. But the telemarketer doesn't care; he just wants to get away. Unfortunately for the doctor, the telemarketer's attempt to shake him off is successful. As a result, the doctor rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

(doctor version)

Please read the following paragraphs carefully. Their details will be important later, and you will not be able to browse back to check the text again. Suppose that you are sitting on a jury for a criminal case. The facts of the case are as follows: Jones, a doctor by profession, is driving his car, and stops at a red light. Smith, a local telemarketer, stands on the sidewalk nearby. As it happens, the two look very much alike. In a remarkable coincidence of mistaken identity, each man mistakes the other for a notorious car thief whose photograph has frequently been broadcast on the local news, and who happens to resemble them both. The telemarketer begins to approach the vehicle, hoping to apprehend its driver. Although the telemarketer is clearly unarmed, the doctor thinks the telemarketer may try to carjack him, and so he panics and speeds off through the intersection. Amazingly, the telemarketer manages to hold on to the side of the car as it speeds off. The doctor swerves in a zigzag fashion in the hopes of escaping — understanding that doing so may place the telemarketer in grave danger. But the doctor doesn't care; he just wants to get away. Unfortunately for the telemarketer, the doctor's attempt to shake him off is successful. As a result, the telemarketer rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.