The Language of Mens Rea

Matthew R. Ginther[α]
Francis X. Shen[β,γ]
Richard J. Bonnie[δ]
Morris B. Hoffman[ε]
Owen D. Jones[ζ]
René Marois[η]
Kenneth W. Simons[θ,ι]

This Article empirically tests two key questions. First: How sensitive are jurors to variations in the language that delineates the criminal mental state

---

α. J.D. Candidate, Vanderbilt Law School; Ph. D. Candidate in Neuroscience, Vanderbilt University.

β. McKnight Land-Grant Professor & Associate Professor of Law, University of Minnesota; Executive Director of Education and Outreach, MacArthur Foundation Research Network on Law and Neuroscience.

γ. Coauthors listed beyond the second author are listed alphabetically by last name.

δ. Harrison Foundation Professor of Medicine and Law, Professor of Psychiatry and Neurobehavioral Sciences, Professor of Public Policy, and Director of the Institute of Law, Psychiatry and Public Policy at the University of Virginia.

ε. District Judge, Second Judicial District (Denver), State of Colorado; Adjunct Professor of Law, University of Colorado and University of Denver; Member, John D. and Catherine T. MacArthur Foundation Research Network on Law and Neuroscience; Research Fellow, Gruter Institute for Law and Behavioral Research.

ζ. Joe B. Wyatt Distinguished University Professor, New York Alumni Chancellor's Professor of Law, and Professor of Biological Sciences, Vanderbilt University; Director, MacArthur Foundation Research Network on Law and Neuroscience.

η. Professor, Department of Psychology and Center for Integrated and Cognitive Neuroscience, Vanderbilt University.

θ. Professor of Law and The Honorable Frank R. Kenison Distinguished Scholar in Law, Boston University School of Law.

ι. Acknowledgments: We received helpful comments from members of the MacArthur Foundation Research Network on Law and Neuroscience and from the organizers of and participants at the Adjudicating Guilty Minds Symposium at Duke Law School, the Guilty Minds: Neuroscience & Criminal Law Symposium of the Denver University Law Review, and conferences of the MacArthur Foundation Law and Neuroscience Project. Roland Nadler and Sarah Prentice-Mott provided valuable research assistance. Preparation of this Article was supported in part by a grant from the John D. and Catherine T. MacArthur Foundation. Its contents reflect the views of the authors and do not necessarily represent the official views of either the John D. and Catherine T. MacArthur Foundation or The MacArthur Foundation Research Network on Law and Neuroscience.

categories? Second: To what extent do jurors assign culpability in the manner assumed by the Model Penal Code (MPC)?

In prior work, we challenged numerous assumptions underlying the MPC mental state architecture, which divides guilty minds into four kinds: purposeful, knowing, reckless, and negligent. Our experiments showed that subjects had profound difficulty categorizing some of the mental states, particularly recklessness, in the context of scenarios in which hypothetical actors caused harmful results. And, when asked to punish hypothetical actors, subjects punished knowing behavior and reckless behavior indistinguishably.

Here, we extend our prior work in two main ways. First, we show that a person's ability to apply the MPC mental states is susceptible to subtle variations in the language that defines and communicates them. For instance, we demonstrate that using slightly different wording can significantly improve participants' ability to accurately identify the mental state of recklessness (notwithstanding that reckless and knowing mental states remain by far the hardest to classify). Second, we show that even when people can see the mental state distinctions that the MPC draws, they don't necessarily rank order the mental states—by culpability level—in the order the MPC assumes.

These findings raise questions about the normative basis for the knowing/reckless distinction in the MPC's mental state hierarchy in the context of result elements. Further, because even small changes in phrasing can produce significant differences in juror evaluation, the findings raise genuine concerns about the adequacy of MPC-based culpability instructions in criminal cases. Our results suggest the need for a critical reexamination of the substantial divide between the expectations and assumptions of drafters of criminal codes, on one hand, and empirical reality, on the other.

I. INTRODUCTION 1329
II. PREVIOUS EMPIRICAL RESEARCH 1331
A. *Research on Mental State Attributions* 1332
B. Sorting Guilty Minds*: Summary of Results* 1334
III. NEW EXPERIMENTS: DESCRIPTION & RESULTS 1339
A. *Three New Experiments* 1339
1. Revised MPC Definitions Experiment 1339
2. Variation in Signaling Phrases 1343
3. Revised Recklessness Experiment 1347
B. *Experimental Methods* 1349
C. *Results* 1351
1. Experiment 1: Revised MPC Definitions ... 1351
2. Experiment 2: Signal Variant Experiment 1353
3. Experiment 3: Revised Recklessness Experiment 1354

IV. DISCUSSION 1358
A. *Improving Sorting by Improving the Language of Recklessness* 1358
B. *Punishment Ratings Are Unaffected by Improved Sorting* 1360
C. *Study Limitations* 1361
V. CONCLUSION 1363
VI. APPENDIX A: TECHNICAL AND STATISTICAL DETAILS 1365
A. *The Participants* 1365
B. *The Experimental Paradigm* 1368
C. *Details of the Experimental Results* 1369
VII. APPENDIX B: FULL TEXT OF SCENARIOS 1372

I. INTRODUCTION

To be guilty of a crime, generally one must commit a bad act while in a culpable state of mind. But the language used to define, partition, and communicate the variety of culpable mental states (in Latin, mens rea) is crucially important. For depending on the mental state that juries attribute to him, a defendant can be convicted—for the very same act and the very same consequence—of different crimes, each with different sentences.

The influential Model Penal Code ("MPC") of 1962 divided culpable mental states into four now-familiar kinds: purposeful, knowing, reckless, and negligent.[1] Both before the MPC and since, scholars in criminal law and philosophy have actively debated how best to define and apply the mens rea categories.[2] Yet few empirical studies have explored the actual relationships between specific mens rea formulations and legally relevant outcomes.

A 2011 article coauthored by several of us, *Sorting Guilty Minds*, presented experiments that tested the MPC's twin assumptions that: (1) ordinary people naturally do—or at least can, when instructed—distinguish these four categories of mental states with reasonable

1. MODEL PENAL CODE § 2.02 (1962).

2. *See, e.g.*, LARRY ALEXANDER & KIMBERLY KESSLER FERZAN, CRIME AND CULPABILITY: A THEORY OF CRIMINAL LAW (2009); Darryl K. Brown, *Federal Mens Rea Interpretation and the Limits of Culpability's Relevance*, 75 LAW & CONTEMP. PROBS., no. 2, 2012, at 109; Claire Finkelstein, *The Inefficiency of Mens Rea*, 88 CALIF. L. REV. 895 (2000); Douglas N. Husak, *The Sequential Principle of Relative Culpability*, 1 LEGAL THEORY 493 (1995); Paul H. Robinson & Jane A. Grall, *Element Analysis in Defining Criminal Liability: The Model Penal Code and Beyond*, 35 STAN. L. REV. 681 (1983); Kenneth W. Simons, *Should The Model Penal Code's Mens Rea Provisions Be Amended?*, 1 OHIO ST. J. CRIM. L. 179 (2003).

reliability; and (2) holding the act and harm constant, the average person would punish acts reflecting these four mental states in the manner corresponding to the MPC hierarchy—that is, they would punish purposeful conduct the most severely and negligent conduct the least.[3]

Those experiments found that these assumptions held, for the most part. But an interesting and important exception emerged at the boundary between knowing and reckless conduct: in sorting the mental states and in assigning punishment, subjects were much less able to differentiate between knowing and reckless conduct.

On the basis of those findings, the article outlined several possible reforms—assuming the results were validated in future studies.[4] To validate and extend those results, we have conducted a series of additional experiments, reported here, with more than 1,600 additional subjects.

Two primary results emerge. First, we demonstrate that modifying the language used to communicate mens rea can significantly improve participants' ability to accurately identify the mental state of recklessness. However, subject accuracy in identifying the reckless and knowing mental states remains far below the classification accuracy for the other mental states.

Second, notwithstanding the gains in *sorting* accuracy, our subjects did not actually *punish* knowing and reckless behavior differently. Our observation that improved sorting of knowing and reckless mental states does not result in a corresponding distinction in the punishment ratings of knowing and reckless behavior suggests that subjects do not see a clear moral distinction between those two mental states, at least in relation to the "result" element of offenses. These findings raise, but do not fully answer, questions about the normative basis for including the knowing/reckless distinction in the MPC's mental state hierarchy in the context of result elements.

These findings also have implications for legal practice. Legislatures and courts use a variety of words to define and communicate mens rea. Typically overlooked is whether a particular formulation of a mental state will matter for juror understanding and decision-making. When jury instructions are reviewed on appeal, judges have only their experience and intuition to guide them as to the possible misunderstanding caused by particular wording. Our results here suggest that juror decision-making is sensitive to the precise

3. Francis X. Shen, Morris B. Hoffman, Owen D. Jones, Joshua D. Greene & René Marois, *Sorting Guilty Minds*, 86 N.Y.U. L. REV. 1306 (2011). We will subsequently refer to this article as the "original study."

4. *Id.* at 1348.

language of mens rea in ways that legal decision-makers may not anticipate.

The Article proceeds as follows. Part I describes our original study in the context of what little existing empirical literature there is on juror assessments of MPC mental states. Part II details the design and results of the new experiments. Part III discusses the implications. Two appendices provide additional details of the experiments, including the full text of the scenarios used.

II. PREVIOUS EMPIRICAL RESEARCH

The Model Penal Code, developed by the American Law Institute in the mid-twentieth century, has been highly influential in shaping the definition of mens rea terminology in state criminal codes and in judicial opinions.[5] The vast bulk of the states—thirty-four of them[6]—either have adopted or have been heavily influenced by the Model Penal Code, which since 1962 has divided the universe of potential culpable mental states into: purposeful, knowing, reckless, and negligent.[7] Even codes that continue to use common law terms have been interpreted in light of Model Penal Code concepts and definitions.[8] Due to the MPC's substantial and continued influence, scholars in criminal law and philosophy have actively debated how best to define and apply the mens rea categories.[9] However, surprisingly little research has examined how laypeople—who are most commonly charged with applying the Model Penal Code's mens rea provisions—actually interpret and apply the Code.

---

5. Peter W. Low, *The Model Penal Code, the Common Law, and Mistakes of Fact: Recklessness, Negligence, or Strict Liability?*, 19 RUTGERS L.J. 539, 540–41 (1988); Paul H. Robinson, *A Brief History of Distinctions in Criminal Culpability*, 31 HASTINGS L.J. 815, 815–16 (1980); Robinson & Grall, *supra* note 2, at 691–703 (discussing MPC approach to elements analysis); Simons, *supra* note 2, at 180–81.

6. Paul H. Robinson & Markus D. Dubber, *The American Model Penal Code: A Brief Overview,* 10 NEW CRIM. L. REV. 319, 326 (2007).

7. MODEL PENAL CODE § 2.02 (1962).

8. The U.S. Supreme Court and lower federal courts also cite to the Model Penal Code with some frequency. *See, e.g.*, *Global-Tech Appliances, Inc. v. SEB S.A.*, 131 S. Ct. 2060, 2069 (2011) (referencing the MPC to provide the appropriate definition for several mental state classifications in an attempt to distinguish these mental states from the concept of willful blindness as it has been articulated by the Courts of Appeals).

9. *See, e.g.*, ALEXANDER & FERZAN, *supra* note 2; Brown, *supra* note 2; Finkelstein, *supra* note 2; Husak, *supra* note 2; Robinson & Grall, *supra* note 2; Simons, *supra* note 2.

A. Research on Mental State Attributions

Empirical research on the ability of laypeople to distinguish specific mental states as required by law is only now emerging.[10] In 1992, researchers at the University of Washington investigated how students interpret and apply the legal definition of four culpable mental states: purpose ("P"),[11] knowledge ("K"), recklessness ("R"), and negligent ("N").[12] They found that subjects could only distinguish between the extremes of P and N. Subjects could not reliably distinguish in the middle of the hierarchy: P vs. K, P vs. R, K vs. R, K vs. N, and R vs. N.[13]

Jury instructions made no difference in subjects' ability to make these distinctions. When subjects were asked to assign punishment ratings, these ratings again did not differentiate between any mental states aside from the extremes of P and N. This result held true both for those subjects who did not have the legal definitions provided and for those who did.[14]

Also in the early 1990s, legal scholar Paul Robinson and psychologist John Darley ran a series of experiments that, in contrast

---

10. Kevin Jon Heller, *The Cognitive Psychology of Mens Rea*, 99 J. CRIM. L. & CRIMINOLOGY 317, 320–21 (2009):

> [C]ontemporary criminal law requires jurors to be latter-day Kreskins—to not only reliably distinguish nearly indistinguishable mental states, but also to accurately determine which of many possible mental states the defendant actually possessed at the time of the crime. Is such mindreading possible? . . . Given the centrality of mens rea to criminal responsibility, we would expect legal scholars to have provided a persuasive answer to this question. Unfortunately, nothing could be further from the truth.

Justin D. Levinson, *Mentally Misguided: How State of Mind Inquiries Ignore Psychological Reality and Overlook Cultural Differences*, 49 HOW. L.J. 1, 3 (2005) ("Scholars have not yet . . . empirically examined the psychological mechanisms involved in understanding others' minds in the legal setting."). For a more detailed discussion of these studies, see Shen et al., *supra* note 3, at 1320–26.

11. The authors of the 1992 study used the term "intent" to refer to the mental state category that we reference as purpose in the present work. Because they convey the same legal significance, we refer to the category as purpose ("P") throughout. We also use a blameless ("B") condition to signal the absence of a culpable mental state.

12. Laurence J. Severance, Jane Goodman & Elizabeth F. Loftus, *Inferring the Criminal Mind: Toward a Bridge Between Legal Doctrine and Psychological Understanding*, 20 J. CRIM. JUST. 107 (1992). Surprisingly, this study has to date been cited only once within the Westlaw JLR database.

13. The researchers found that, when rank-ordering mental states, "legally naïve subjects could not, on their own, reliably agree on differentiation between 'criminal knowledge' and 'criminal recklessness' nor reliably distinguish these from other legally relevant mental states." *Id.* at 115.

14. In addition, Severance et al. carried out a content analysis of subject-generated mental state definitions. They sought to determine, qualitatively, the extent to which subjects' definitions of the mens rea terms varied from the legal definitions. The researchers found that subjects often had their own set of preconceptions that deviated from the legal concepts of mens rea. *Id.* at 114.

to the 1992 study, found that individuals do typically assign liability and punishment in a manner generally consistent with the MPC.[15] As summarized by Robinson and Darley, "[T]he responses of our subjects . . . assign a higher degree of liability to the knowing versus the reckless commission of all offenses."[16]

A decade after the Robinson and Darley study, law professor Justin Levinson conducted an experiment that explored the mediating role of culture in the assessment of defendants' mental states.[17] While the primary objective of the study was to assess cultural differences in assessments of mental states and culpability, Levinson noted that even across cultures, in the majority of the scenarios the responses provided did not match the responses predicted by the MPC.[18]

Most recently, in 2012, psychologists Pam Mueller and John Darley and legal scholar Lawrence Solan published a study examining mental states and punishment in civil disputes.[19] The study used as its centerpiece a series of vignettes based on a case in which a workman is electrocuted while attempting an emergency repair job.[20] The researchers manipulated the level of knowledge that the employer had about the employee's risk of being electrocuted.

The primary finding of the study was that "not only do people regard the side effects of knowing acts as intentional when assigning liability, but they also regard the side effects of *reckless* acts as intentional when making liability judgments."[21] This is consistent with findings in moral psychology that subjects tend to characterize known negative side effects, but not positive side effects, as intended.[22]

The researchers then followed up by running an additional experiment manipulating the perceived level of risk (as communicated

---

15. PAUL H. ROBINSON & JOHN M. DARLEY, JUSTICE, LIABILITY AND BLAME (1995).

16. *Id.* at 87.

17. Levinson, *supra* note 10, at 2–3.

18. Only when averaging over all four fact patterns does Levinson find some evidence that "participants maintained a folk mental state hierarchy," placing "purpose above knowledge above recklessness" in their punishment ratings. *Id.* at 20. But these results were not robust, as they did not hold in each fact pattern when analyzed individually. *Id.* at 21.

19. Pam A. Mueller, Lawrence M. Solan & John M. Darley, *When Does Knowledge Become Intent?: Perceiving the Minds of Wrongdoers*, 9 J. EMPIRICAL LEGAL STUD. 859, 859 (2012). The researchers asked, "[W]hen judging behavior, do people distinguish between intentional and knowing acts, knowing and reckless acts, reckless and negligent acts, and so on?" *Id.*

20. *Id.* at 865 (noting that "[t]he scenarios are based loosely on *Parret v. Unicco Service Co.* (2005), a case decided by the Supreme Court of Oklahoma").

21. *Id.* at 875.

22. Joshua Knobe, *Intentional Action and Side Effects in Ordinary Language*, 63 ANALYSIS 190, 193 (2003); Thomas Nadelhoffer, *Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality*, 9 PHIL. EXPLORATIONS 203 (2006).

by percentage likelihood of the employee injury occurring).[23] Noteworthy from the results was the finding that—even with only a 3% perceived likelihood of the harm occurring—35% of subjects (when given the reckless treatment) concluded that the employer acted intentionally.[24]

Summarizing these empirical studies as a whole, it is apparent that there is much variation in jurors' abilities to accurately assess defendants' mental states. Methodological limitations in previous studies may explain this variation, including study designs in which experimental subjects were exposed to repeated variations of the same fact patterns as well as differences in subject pools across the studies.[25]

B. Sorting Guilty Minds*: Summary of Results*

In our original study,[26] we exposed subjects to a series of short, unique scenarios, each of which was designed to be straightforward and reasonably believable on its face, clearly communicative of a distinct MPC mental state, and concise enough so that subjects could read multiple scenarios within a reasonable time.[27] Moreover, because previous research has pointed to the interaction of harm level with mental state determinations and because the MPC and many states place differential importance on the mental state boundaries depending on the severity of the offense, we also varied the harm level across our scenarios.[28]

---

23. The risk levels used were 3, 20, 50, 80, 97, and 100%, respectively.

24. Mueller et al., *supra* note 19, at 888. The researchers conducted additional studies to determine whether the type of knowledge held by the employer affected intentionality judgments. They found that "knowing who is going to be injured, at least within a limited population, is irrelevant to judgments of intentionality; knowing when someone is going to be injured is relevant to intentionality judgments; and knowing *how* someone is going to be injured is essential for perceiving intentionality." *Id.*

25. For additional discussion of these issues, see Shen et al., *supra* note 3, at 1324–26.

26. In the present Article, when we refer to "we" in reference to this 2011 study, we are referring only to the subset of the current authorship who coauthored the original study.

27. These constraints raised a number of questions about how to effectively and efficiently communicate the protagonist's motivation and intent. John's action in each of our scenarios was explained to subjects with a simple, and typically neutral, motivation. For instance, in one scenario, subjects read that John acted because he was angry after an argument with a player on an opposing softball team. Scenario construction was also sensitive to the fact that, as found in research by philosopher Thomas Nadelhoffer, moral judgments about the actor involved may influence mental state assessment. Nadelhoffer, *supra* note 22, at 203.

28. For a discussion of the relationship between judgments of intentionality and harm caused, see Joshua Knobe, *The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology*, 130 PHIL. STUD. 203 (2006), and Edouard Machery, *The Folk Concept of Intentional Action: Philosophical and Experimental Issues*, 23 MIND & LANGUAGE 165 (2008).

Applying these principles, we drafted scenarios featuring a protagonist (always named John) whose actions resulted in various levels of harm, which we categorized into three levels.[29] We organized the individual scenarios within "themes." For purposes of this discussion, we use the term "theme," which is akin to previous researchers' "stem," to refer to the general fact pattern. Each theme had five variants, one for each mental state. The only difference between each variant was the manipulation of John's mental state as to the resulting harm. Because nothing else changed between the variants, this allowed us to attribute differences in behavior to variations in mental state.

Every mental state variant of a theme shared the same first and third sentence. The first sentence always served as an introductory sentence (e.g., "John is gardening in his backyard, where there are many plants and many small rocks."), and the third sentence always presented the resulting harm (e.g., "The rock hits the window, but since his neighbor's window is made of especially tough glass, the rock bounces off and causes no harm."). The second sentence was modified in each variant in order to introduce the scenario-specific mental state for a given theme (e.g., "Wanting to get rid of a small rock, he throws the rock over the fence, aware that there is a substantial risk that the rock will also hit his neighbor's nearby window, but choosing to ignore it.").[30] Thus, within each theme there were five scenarios: one each for purposeful, knowing, reckless, negligent, and blameless. We created thirty themes, ten in each of three harm-level categories, giving us a total of one hundred and fifty unique scenarios.[31] To ensure that subjects could not easily learn the nature of our manipulations, we systematically rotated the mental state signals across scenarios.[32] As we discuss in Part III, this approach did not allow us to determine whether particular signals were related to subject behavior—that is, whether a particular signal contributed to higher or lower accuracy in mental state identification.

---

29. While the nature of the harm was heterogeneous across scenarios, the resulting harm was classified into one of three categories: high harm (causing death or serious physical injury); medium harm (causing minor injury or great property damage); and low harm (causing no injury or minor property damage).

30. All scenarios were constructed so that they would have roughly the same total number of words. Scenario length was seventy-three words, plus or minus two words.

31. For the full set of scenarios, see *infra* Appendix B.

32. We also "counterbalanced" the presentation of mental state signals across the three harm levels. That is, subjects did not see all the scenarios and respective mental states in the same order. Rather, we also used randomization in presenting scenarios to mitigate against any foreseeable order effects.

Several experiments were run in the original study, but we focus here on the version of the experiment in which subjects had full access to the mental states definitions throughout the experiment. As noted above, one problem with some of the earlier research on this topic was that the design had subjects assess the same fact pattern multiple times, only changing the mental state of the fact pattern across trials. Doing so would have made it clear to the observant participant that the actor's mental state was being manipulated on each trial. To avoid this problem in our experiments, subjects saw a given fact pattern only once. Therefore, each subject read thirty of the total one hundred and fifty scenarios, six from each of the five mental states, and one, randomly assigned, from each theme.[33] In the rating experiments, after reading each scenario subjects were asked: "On a scale from 0–9, with 0 being no punishment and 9 being extreme punishment, how much should John be punished for his behavior?"[34]

The original study found that punishment ratings were highest for purposeful action. At the other end of the spectrum, blameless punishment averages were the lowest, and negligent averages were the second lowest.[35] In the middle, punishment for K and R was generally less than P and more than N. These results show not only that subjects

---

33. Subjects also were given five practice scenarios, one from each mental state and spanning the approximate range of harms, before the actual experiment, in order to familiarize them with the interface and the experimental task. These practice themes were developed in addition to the thirty themes used in the actual experiment.

34. In the original study and in this Article, when asked to rate punishment, subjects were given a 0-to-9 scale. We used text next to the numbers on the scale to communicate to subjects that 0 reflected no punishment and 9 reflected the most extreme punishment. In each study, we used a series of anchoring scenarios to introduce participants to the range of harms they would see during the experiment. The reported analysis standardized punishment ratings to control for the possibility that subjects likely had different understandings of the type of punishment associated with each number. Moreover, as an additional check, we ran a separate experiment in which we asked subjects (after they completed the punishment task) to provide us with a description of their personal scale, i.e., what type of punishment did they associate with the number one, five, and so forth. We found these subjective punishment scales to show a large amount of concordance across individuals.

It is important to recognize that research in moral psychology has found that individuals may assign blame differently than they assign punishment. *See, e.g.*, Jennifer K. Robbennolt, *Outcome Severity and Judgments of "Responsibility": A Meta-Analytic Review*, 30 J. APPLIED SOC. PSYCHOL. 2575, 2580 (2006) (discussing the variety of outcome variables that researchers have used to measure responsibility judgments). To account for this possibility, we ran a set of *blame*-rating experiments, identical to the punishment rating experiments, except for a change in the rating question asked. Thus, we reran experiments one, two, three, and five with a focus on blame rather than punishment. In these additional experiments, subjects were asked, after reading each scenario: "On a scale from 0–9, with 0 being not at all blameworthy and 9 being extremely blameworthy, how blameworthy is John for his behavior?" The results from the blame-rating experiments followed the same pattern as the punishment-rating experiments we discuss in the text.

35. Shen et al., *supra* note 3, at 1337–44.

punished in these categories in accord with the normative suppositions reflected in the MPC but also that subjects were very good at distinguishing these three categories of mental states from one another, even if they were not explicitly identifying the mental state of the actor in their evaluations of each scenario. However, subjects did not demonstrate differentiation of their punishment ratings at the junction between K and R. K was often punished no differently, or even less harshly, than R, a result not at all in keeping with the MPC's hierarchical assumptions.

Why didn't we see, even when we gave subjects the MPC definitions, higher punishment ratings for K scenarios than for R scenarios? There are at least two plausible explanations. First, it could be that subjects are capable of identifying a conceptual difference between knowing and reckless action but employ a moral calculus in which knowing and reckless actors are punished roughly the same. If this is the case, then the subjects' behavior focuses attention on the normative question of whether causing harm knowingly is indeed more culpable than causing harm recklessly.[36]

This deep normative question is beyond the more limited scope of the present empirical investigation, but a brief discussion of the MPC's normative presuppositions is warranted. Although the MPC often distinguishes between knowing and reckless action for the purpose of defining culpability, there are surprisingly few instances where the drafters of the MPC employed categorical differences between K and R for result elements of offenses.[37] For example, murder

---

36. In the original study, we ran versions of the experiment in which we asked subjects how much "blame" (as opposed to punishment) they would assign to the scenario protagonist. The results for the blame and punishment results were substantively similar, and to ease the presentation of results we described punishment ratings as reflecting subjects' assessments of moral culpability. As the original study recognized, however, blame and punishment are not synonymous. Two equally blameworthy individuals may be punished differently (by subjects and by the criminal justice system) on the basis of other utilitarian considerations such as incapacitation or deterrence. Thus, we cannot conclude with certainty that a difference in punishment ratings (or an absence of a difference) is due to a difference in subject assessment of moral culpability. This caveat aside, however, other empirical studies have demonstrated that punishment ratings are primarily driven by retributive notions of justice. *E.g.*, Kevin M. Carlsmith, *The Roles of Retribution and Utility in Determining Punishment*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 437 (2006).

37. *See* Robinson & Grall, *supra* note 2, at 723–24 (noting that the Model Penal Code refers to, but does not define, result and circumstance elements, and plausibly suggesting that result elements should be defined as circumstances changed by the actor, for example, causing a death in homicide or causing a fire in arson). Result elements are the consequence of one's actions (such as the death of the victim) that the state must prove the defendant brought about. For many offenses, the state must also, or instead, prove that one or more circumstance elements obtain (for example, that the package in the defendant's possession contains illegal drugs). There are not many pure result offenses in the MPC.

and aggravated assault include causing the result purposely, knowingly, or recklessly under circumstances manifesting extreme indifference to human life. A key feature of many extreme indifference cases is heightened probability of causing the harm, thereby blending into K. In other result offenses, the offense is typically punishable at the same level for P, K, and R if the forbidden result occurs.[38]

A second possible explanation for the subjects' failure to distinguish between K and R in their punishment ratings is that they are simply confused when trying to make the distinction. That is, subjects would punish K and R differently *if* they could identify the difference in mental states, but they cannot.

One experiment in our original study attempted to address this difference between subjects' ability to sort the mental states and their normative treatment of those states once sorted. Subjects were provided with the definitions of the mental states alongside each scenario and were instructed, "Please select from the question options below the definition that best matches John's mental state in this scenario." This allowed us to determine, for each mental state, subjects' ability to correctly classify the mental state of the actor in the scenario in terms of the MPC hierarchy.

The results suggested that subjects could identify purposeful and blameless scenarios with a high degree of accuracy. Subjects correctly identified purposeful scenarios 78% of the time and correctly identified blameless scenarios 88% of the time. Subjects were most likely to err in the middle categories of knowing (50% success rate), reckless (40% success rate), and negligent (48% success rate).

This low level of accuracy in identifying knowing and reckless mental states made it difficult to know which of our two proposed mechanisms explained the punishment ratings, but it was clear that we could not rule out the possibility that subjects simply couldn't distinguish the two mental states—and that this could explain the indistinguishable punishment ratings.[39]

---

38. Take, as an example, the crime of burglary in the second degree (inflicting harm during a burglary), MODEL PENAL CODE § 221.1(2)(a) (1962), and the crime of cruelty to animals, *id.* § 250.11. In both cases the MPC does not create gradations between P, K, and R. One counterexample is "causing catastrophe"—which is a second-degree felony if committed with P or K, but third degree if committed with R. *Id.* § 220.2(1). In this case, however, there may be a significant moral difference between knowing that a catastrophic result will occur and knowing that there is a real but small chance that it will. This offense probably should have included extreme recklessness ("eR") cases in the second-degree felony.

39. We also ran a final experiment combining sorting and rating. We designed this experiment to test whether exposing subjects to the sorting task first may result in punishment differences, perhaps due to better appreciation of the mental state gradations. This design also allowed us to test whether those who were better able to identify mental states showed greater differentiation in punishment ratings, particularly at the K/R boundary. To test this, we had

The original study thus suggested that jury-eligible subjects cannot distinguish between knowing, reckless, and negligent conduct with great accuracy. But the study also left open the possibility that this accuracy could be improved by refining the language used to define the MPC mental states and to signal them in the research scenarios. The research team therefore set out to investigate whether improving the language of mens rea could improve the ability of subjects to recognize, sort, and rate these mental states as expected.

III. NEW EXPERIMENTS: DESCRIPTION & RESULTS

A. Three New Experiments

The results of the original study were premised on the assumption that the scenario protagonist's mental state was clearly signaled to subjects and that the mental state categories that the scenarios were being sorted into were clearly defined. This assumption allowed us to interpret the failure to distinguish between knowing and reckless scenarios as the *subjects'* failure. But it is possible that the original results were sensitive to the way we defined or communicated the mental states.

Language might have mattered in three possible ways. First, the MPC definitions provided to subjects might not have been clear enough. Second, the specific language used in the scenarios to communicate mental states might have had a substantial effect on how individuals interpreted the scenarios. Third, the reckless signals in particular may not have properly conveyed the substance of the category as intended by the MPC drafters. We designed three new experiments to address these three possibilities.

1. Revised MPC Definitions Experiment

Essential to both the sorting and the rating tasks were the definitions provided to subjects for the five mental states (purposeful, knowing, reckless, negligent, and blameless). Would a different formulation of the MPC mental states definitions produce more

---

subjects first sort fifteen scenarios according to MPC definitions. These same subjects were then asked to rate fifteen different scenarios. The results from this experiment indicated that sorting the scenarios first did not materially change the punishment ratings these same subjects provided. Additionally, limiting the analysis of the punishment ratings task to only those who sorted with above 75% accuracy, there remains no significant difference in punishment ratings at the K/R boundary. Thus, even those who seem to be better able to utilize and understand the knowing/reckless distinction still fail to make a clear moral distinction between the two. The same is true for the "bad sorters" (i.e., subjects with overall accuracy below 50%).

accurate sorting and more differentiation in punishment, especially in the knowing and reckless scenarios? To find out, we ran a new experiment in which we modified the MPC definitions provided to subjects. Table 1 compares the original and new definitions.[40]

Because they were clearly the categories that subjects struggled with the most, we focused our revisions on the three middle categories: knowing (50% accurate in original study), reckless (40% accurate), and negligent (48% accurate).[41]

The Model Penal Code emphasizes that the main difference between knowing and reckless behavior is the actor's perceived probability of risk.[42] In the original experiment, we established this difference in perceived risk by telling subjects that a person acts "knowingly" when he is aware that his conduct is *practically certain* to cause the result, and that a person acts "recklessly" when he is conscious of a *substantial risk* that a result will occur. Thus, the difference in probability was that between practically certain and substantial risk. This tracks the language used in the MPC.[43]

Also tracking the MPC, in the original study, we told subjects that a person acts "knowingly" when he is aware that his conduct is practically certain to cause the result but that a person acts "recklessly" when he consciously disregards a substantial and unjustified risk that a result will occur or that a circumstance exists. The addition of the "consciously disregards" language was problematic because it may have erroneously suggested to subjects that such a conscious choice was not an element of knowing action. Knowing action also, of course, requires an actor to consciously disregard a risk—indeed, an even greater risk—that his actions will cause the harmful result.[44] By removing the

---

40. The study design—in which subjects read through thirty scenarios—requires parsimonious communication of the protagonist's mental state. In revising definitions, we aimed to further improve the clarity of the signals while maintaining this parsimony. This desire for parsimony, as well as the need to isolate the differences between scenarios, prevented us from developing more elaborate fact patterns in which each mental state is communicated primarily through circumstantial evidence and not through the signaling words alone. Future research can investigate the extent to which alternative research designs do a better (or worse) job at communicating mental states if they convey information in a way that more closely matches the types of evidence available at trial.

41. We also edited the blameless definition to improve clarity, as "lacked any of the culpable mental states" was ambiguous and could mean "lacked just one of the mental states," not "lacked all of them."

42. *See* MODEL PENAL CODE AND COMMENTARIES §2.02 cmt. at 236–37 (1985).

43. MODEL PENAL CODE § 2.02.

44. An alternative would have been to insert a choice clause into the knowing scenarios, but this would have been more cumbersome, and, at least anecdotally, the "consciously disregard" language could have generated even more confusion. In addition, this change had the added advantage of removing what seems a priori a negatively valenced word ("disregard"), which, when compared to the more positive word "knowing," might have contributed to the subjects' confusion.

"consciously disregards" language from the reckless definition, we more clearly conveyed to subjects that the primary difference between K and R was the perceived probability of risk.

Another difference between the original K and R definitions was the inclusion of the word "unjustified" in the original reckless definition. Exactly how the dual requirements of substantial and unjustified risk are meant to operate is ambiguous and has been debated by MPC commentators.[45] For our purposes here, the goal was not to settle these deep normative debates about the proper contours of the reckless category but rather to operationalize the definition in a simple way that more clearly differentiated it from the other mental states.[46]

The same logic led us to remove the word "unjustified" from the revised definition of negligent. In the original study, we told subjects that a person acts negligently when, through a gross deviation from the standard of care that a reasonable person would exercise, he fails to perceive a substantial and unjustified risk that a result will occur or that a circumstance exists.

In addition to striking the word "unjustified," we made one further modification to both the reckless and negligent definitions. The original definitions both referred to a "risk that a result will occur or that a circumstance exists." This tracks the MPC language, which is designed to apply both to result and circumstance elements. But because we focus in this set of studies only on result elements, we refined the definition and removed the language that referenced circumstance elements.

Using these new definitions, but without otherwise changing the experimental design, we ran a Revised MPC Definitions Experiment.[47]

---

45. *See* Larry Alexander, *Insufficient Concern: A Unified Conception of Criminal Culpability*, 88 CALIF. L. REV. 931, 933–35 (2000); Joshua Dressler, *Does One Mens Rea Fit All?: Thoughts on Alexander's Unified Conception of Criminal Culpability*, 88 CALIF. L. REV. 955, 956–59 (2000); Simons, *supra* note 2, at 189–92; David M. Treiman, *Recklessness and the Model Penal Code*, 9 AM. J. CRIM. L. 281, 362–67 (1981).

46. We did not implicate the unjustifiable requirement in the fact patterns used in these experiments. Rather, in all of the fact patterns (except the blameless ones), the scenario was deliberately constructed so that the actor's conduct was unjustifiable. Thus, including "unjustifiable" in some but not all of the culpable mental state definitions would have been an additional and unnecessary distraction. While the concept of justification is sometimes an important issue, and the MPC embraces the view that causing a result with a mental state of K is more difficult to justify than causing a result with a mental state of R, the issues surrounding justification are not a principal aim of this investigation.

47. Because our question of interest was only whether the above changes, en masse, had an effect on our experimental results, and not the differential effect of each change, we did not run multiple experiments for the various manipulations. This limits our ability to causally link any one definitional change to differences in the punishment ratings or sorting accuracy. Causal inferences about the effect of a particular phrase should be further investigated if deciding upon a

We ran both a sorting and a punishment task, with independent samples, as a part of this experiment. For both tasks, all components of the experimental design, other than the definitional changes, stayed the same as in the original study, including all of the text of the scenarios.[48] Thus, any difference in sorting or punishment rating can be attributed to these modifications in the mental states definitions.

single phrase with which to instruct jurors, but our initial research goal here was to see if improving the set of definitions employed in the experiment could improve sorting accuracy.

48. One additional change concerned the recruitment of subjects. In the original study, we paid Qualtrics to recruit subjects for us. In the new series of studies, we used Amazon's Mechanical Turk service to recruit subjects directly. Multiple studies have validated results using Amazon's Mechanical Turk on a variety of assessments, especially when compared to samples of convenience. *See, e.g.*, Tara S. Behrend et al., *The Viability of Crowdsourcing for Survey Research*, 43 BEHAV. RES. METHODS 800 (2011); Adam J. Berinsky et al., *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk*, 20 POL. ANALYSIS 351 (2012); Michael Buhrmester, Tracy Kwang & Samuel D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?*, 6 PERSP. ON PSYCHOL. SCI. 3 (2011); Joseph K. Goodman et al., *Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples*, 26 J. BEHAV. DECISION MAKING 213 (2012); Jon Sprouse, *A Validation of Amazon Mechanical Turk for the Collection of Acceptability Judgments in Linguistic Theory*, 43 BEHAV. RES. 155 (2011).

Table 1: Definitions Used in Original Study and in New Experiment 1[49]

Original Definitions	***New Definitions***
1. Purposefully: A person acts "purposefully" when his conscious objective is to cause the specific result.	1. Purposefully: A person acts "purposefully" with respect to a result when his conscious objective is to cause the specific result.
2. Knowingly: A person acts "knowingly" when he is aware that his conduct is practically certain to cause the result.	2. Knowingly: A person acts "knowingly" with respect to a result when he is aware that his conduct is practically certain to cause the result.
3. Recklessly: A person acts "recklessly" when he consciously disregards a substantial and unjustified risk that a result will occur or that a circumstance exists.	3. Recklessly: A person acts "recklessly" with respect to a result when he is aware of a substantial risk that his conduct will cause the result.
4. Negligently: A person acts "negligently" when, through a gross deviation from the standard of care that a reasonable person would exercise, he fails to perceive a substantial and unjustified risk that a result will occur or that a circumstance exists.	4. Negligently: A person acts "negligently" with respect to a result when, through a gross deviation from the standard of care that a reasonable person would exercise, he fails to perceive a substantial risk that his conduct will cause the result.
5. Blamelessly: A person is "blameless" even though he may have caused harm, if he lacked any of the culpable mental states defined above.	5. Blamelessly: A person acts "blamelessly" when he does not have any of the culpable mental states defined above.

2. Variation in Signaling Phrases

Signaling mental states requires developing a fact pattern (i.e., a "theme"), and using words to describe the mental state within that fact pattern (i.e., a "scenario"). For the original study, we communicated John's mental state with regard to the harm being caused in the following way.

---

49. In addition to receiving the definitions, subjects were told:

> A crime is committed when the defendant has committed a voluntary act prohibited by law accompanied by a culpable mental state. Voluntary act means an act performed consciously as a result of effort or determination. Culpable mental state means either purposefully, knowingly, recklessly or negligently, as explained in this instruction. Proof of the commission of the act alone is not sufficient to prove that the defendant had the required culpable mental state. The culpable mental state is as much an element of the crime as the act itself.

If we label harm as the *y* variable and John's action in the scenario as the *x* variable, then within each theme *x* varies, *y* remains constant, and the general relationship between *x* and *y* is as follows:

- Purposefully: John decides to cause [or bring about] *y* by doing *x*.
- Knowingly: John does *x*, practically certain that it will result in *y*.
- Recklessly: John does *x*, aware there is a substantial risk that *y* will occur.
- Negligently: John does *x*, failing to perceive a substantial risk that x may cause *y*.
- Blamelessly: John does *x*, and despite being as careful as he could be, *y* happens.

The words used above (for P, K, R, and N) to describe the relationship between mental state (actus reus) and the resulting harm are words taken fairly directly from the language provided by the MPC to describe each mental state—for instance, "practically certain" describing knowing.

One experimental design could strictly adhere to MPC language by using only these formulations. But using just these MPC signaling terms creates two problems. The first concerns habituation. If we exposed subjects to identical signaling language for each mental state, it's likely that over the course of the experiment (as they spotted that same word multiple times) they would recognize the phrase as a sort of code word for the mental state. The second problem is that using only the MPC formulation leaves us unable to say anything about whether the MPC language could be improved. For instance, would substituting "almost positive" for "practically certain" be easier for subjects to understand? In our new Signal Variant Experiment we modified our experimental design to allow us to answer such questions.[50]

In the Signal Variant Experiment, we tested the effect of signaling phrases as follows. First, from the thirty themes used in the original study, we selected nine themes on the basis of behavioral data from the original study indicating that subjects were more capable of accurately parsing the K/R distinction for those themes.[51] The nine

50. In the original study, five alternatives for each mental state signaling phrase were developed. But each of the 150 scenarios was assigned one and only one signaling phrase. Because we did not rotate all five alternative phrases for each scenario, the design did not allow us to determine whether a particular phrase was more (or less) helpful in allowing subjects to accurately sort the mental states.

51. Of our original thirty scenarios, we selected those scenarios where the mean punishment ratings were higher for the K scenarios as compared to the R scenarios and where the classification accuracy for both K and R scenarios were above 35% in the classification experiment. We took this

themes we used included three themes each from our three harm levels (low, medium, and high). For each theme, we had already developed (from the original study) five scenarios, one for each mental state. We dropped the blameless scenario[52] and then created twenty scenarios for each theme: five signal variants for each of the four mental states. Table 2 illustrates the twenty variations that were generated for a single theme.

We ran both a sorting task and punishment rating task as part of the Signal Variant Experiment. In each, subjects were randomly assigned to read one of these twenty variants for the nine different scenarios included in the experiment. This allowed us to determine the extent to which particular signals differentially contributed to the punishment effects seen in the original study. In addition, by comparing the results of the sorting and punishment rating tasks, we can tell whether the differences in sorting accuracy reliably correlate with the punishment ratings. In other words, do those R signals that are more likely to be confused with K result in higher punishment, and do those K signals that are more likely to be confused with R result in lower punishment?

---

approach recognizing the possibility that there may have been scenarios in our original experiment that made the K/R boundary particularly difficult for subjects to grasp. Because these scenario-specific errors are not of greatest interest, we selected those scenarios where the scenario itself was least likely to drive an effect. Further, due to the sheer number of scenarios that needed to be written (twenty variants for each scenario), performing the study using all thirty scenarios (six hundred unique variants in total) was impractical. These two filters isolated eleven candidate scenarios and we selected three scenarios from each harm level, leaving us with the nine scenarios used in the present study. This selection process did have the effect of creating a K/R punishment difference across the group where one did not exist in the full set of thirty scenarios. The difference was minor but significant. For this reason we compare the punishment difference between K and R in Experiment 3 to the results from Experiment 2 (where a marginal punishment difference was present) as opposed to our original results (where no punishment difference was present).

52. Unlike the other mental states, the pertinent facts in the blameless scenarios were much more dependent on the context provided beyond the signaling language than other scenarios—for example, a gust of wind or an unforeseen natural event. This made it nearly impossible to create multiple variants of a blameless scenario just by changing the signaling language. Even though we no longer included blameless scenarios, we kept a blameless scenario in the anchoring "practice" questions provided at the start of the survey, and we also kept blameless as an answer choice in the sorting tasks. This allows for meaningful comparison between the various studies.

Table 2: Constructing the Signal Variant Experiment: Varying Mental State *and Signal Variant* Within A Single Theme

Sentence 1 of 3: Setting up the fact pattern, e.g., "John is attending an outdoor concert and is sitting behind a row of other concertgoers."

Note: Every subject saw the same Sentence 1 for each theme.

Sentence 2 of 3: Communicating a single mental state with one of five signals, e.g., (for Purposeful) "During the concert John gets angry that fans in the row in front of him keep standing and blocking his view, so John decides to hurt one of them by throwing his soda can at the row of fans standing in front of him."	
Purposeful	Decides to Intends [to/that/of] Desires [to/that] Wants to Chooses to
Knowing	Practically certain that [the harm will occur] Aware that [the harm] will almost certainly occur Almost positive that [the harm will occur] Virtually certain that [the harm will occur] Understands that [the harm] is almost guaranteed to occur
Reckless[53]	Aware there is a substantial risk that [the harm will occur], but chooses to ignore [it/the risk] Realizes it is very likely that [the harm will occur], but decides to [act] anyway Conscious of the likelihood that [the harm will occur], but simply doesn't care Understands that [the harm could easily happen], but decides to risk it Knows there is a good chance that [the harm will occur], but chooses to [act] anyway
Negligent	Carelessly Wasn't paying attention Hurriedly [and] not seeing Without even noticing Overlooks

Sentence 3 of 3: Describing the harm, e.g., "The soda can hits one of the fellow concertgoers in the face, breaking his nose."

Note: Every subject saw the same Sentence 3 for each theme.

Note: For each of nine themes (three involving high harm, three involving medium harm, and three involving low harm), we constructed four mental state scenarios and further created five variations of each mental state (one each for each of the signal variants listed in the table.) This table uses as an example one of the medium harm themes. Text of all scenarios is presented in Appendix B. Each scenario uses the same first and third sentence, varying only the middle sentence.

53. The Signal Variant Experiment included revised definitions of the mental states (with the phrase "consciously disregards" removed), but we did not additionally revise the reckless definitions (to remove the choice clause) until the Revised Recklessness Experiment. *See* discussion in Section II.A.3 *infra*.

3. Revised Recklessness Experiment

A third modification of the original study concerns the construction of the recklessness signaling phrases (Table 3). The most significant concern was the inclusion of the choice language attached to the end of the original signals (e.g., "but chooses to ignore" the risk). As discussed previously in the context of the MPC definitions, this language, which was not included in the knowing signals, may have contributed to the K/R confusion. To make our signaling language consistent with the revised definitions, we removed the "but chooses to ignore" language from our R signals.

In addition, two of the original recklessness signaling phrases—"realizes it is very likely" and "conscious of the likelihood"—may have communicated too high a probability level, thus conflating them with knowing.[54] Specifically, both signals may have conveyed to the reader that the result was not only possible but *probable*. It can be argued, however, that this level of risk was not consistent with the intent of the MPC, which only requires the probability of the risk to be "substantial" in order to reach the threshold for recklessness.

The decision about what words to use for recklessness raises fundamental questions about what the MPC drafters intended and, more generally, how judges and commentators understand and apply the distinction between K and R. We recognize that the meaning of "substantial" risk is meant to be contextualized in relation to the nature of the harm; however, at the same time, we think that a relatively low (though real) probability may be sufficient to establish recklessness, especially in high-harm cases.[55] With this in mind, we adjusted some of

---

54. In addition, one of our original signals for recklessness inadvertently included the word "knows" ("knows there is a good chance"). In Experiment 3 we replaced the word "knows" with "recognizes."

55. This is the position taken in the Colorado Supreme Court case *People v. Hall*:

> Some risks may be substantial even if they carry a low degree of probability because the magnitude of the harm is potentially great. For example, if a person holds a revolver with a single bullet in one of the chambers, points the gun at another's head and pulls the trigger, then the risk of death is substantial even though the odds that death will result are no better than one in six. . . . Conversely, a relatively high probability that a very minor harm will occur probably does not involve a "substantial" risk. Thus, in order to determine whether a risk is substantial, the court must consider both the likelihood that harm will occur and the magnitude of potential harm, mindful that a risk may be "substantial" even if the odds of the harm occurring are lower than fifty percent.

999 P.2d 207, 217–18 (Colo. 2000). For further discussion, see Simons, *supra* note 2, at 189–92.

The Commentary to the Model Penal Code states as follows:

> Even substantial risks . . . may be created without recklessness when the actor is seeking to serve a proper purpose, as when a surgeon performs an operation that he knows is very likely to be fatal but reasonably thinks to be necessary because the patient has no other, safer chance. [Footnote 14] . . . Some standard is needed for determining *how* substantial and *how* unjustifiable the risk must be in order to warrant

the signals meant to convey likelihood in order to communicate the relatively low probability that we believe is sufficient for a finding of recklessness. These changes are noted in Table 3. Recognizing that our conceptualization of what is and what is not recklessness is debatable, our goal here was to push the lower bounds of the requirements for recklessness in order to determine whether doing so produces a separation between subject responses to recklessness and knowledge scenarios. Because we do not observe such a distinction in the data, we are not concerned that we might have departed from the canonical meaning of recklessness in our signaling language.

Whether to eliminate the "choice" language is more complicated because it opens up the question of the actor's motivation for taking the risk and whether he had good (possibly justifiable) or understandable reasons. However, we designed the scenarios so that the actor's reasons for acting were palpably not good enough to justify taking the risk.

As with the Revised MPC Definitions Experiment and the Signal Variant Experiment, we ran a sorting task and a punishment rating task as part of the Revised Recklessness Experiment. As before, we used a different set of subjects to perform the two tasks.

> a finding of culpability. There is no way to state this value judgment that does not beg the question in the last analysis; the point is that the jury must evaluate the actor's conduct and determine whether it should be condemned. The Code proposes, therefore, that this difficulty be accepted frankly, and that the jury be asked to measure the substantiality and unjustifiability of the risk by asking whether its disregard, given the actor's perceptions, involved a gross deviation from the standard of conduct that a law-abiding person in the actor's situation would observe.

MODEL PENAL CODE & COMMENTARIES § 2.02 cmt. at 237 & n.14 (1985). Footnote 14 states, in part: "On the other hand, less substantial risks might suffice for liability if there is no pretense of any justification for running the risk." *Id.* at 237 n.14.

Table 3: Comparison of Phrasing in Original Study and in the Revised Recklessness Experiment

Signal Phrasing for Reckless in Original Study	***Signal Phrasing for Reckless in Experiment 3***
Aware there is a substantial risk that [the harm will occur], but chooses to ignore [it/the risk].	Aware there is a substantial risk that [the harm will occur].
Realizes it is very likely that [the harm will occur], but decides to [act] anyway.	Realizes there is some risk that [the harm will occur].
Conscious of the likelihood that [the harm will occur], but simply doesn't care.	Conscious of the real risk that [the harm will occur].
Understands that [the harm could easily happen], but decides to risk it.	Understands that [the harm could easily happen].
Knows there is a good chance that [the harm will occur], but chooses to [act] anyway.	Recognizing there is a good chance that [the harm will occur].

B. Experimental Methods

The three experiments discussed in the previous section were conducted in a similar way to our original study. Each subject who participated in an experiment was asked to read a series of short scenarios and answer a single question about the scenario's protagonist after each one. As noted above, for Experiment 2 and Experiment 3 we utilized a selected subset of nine themes, and for Experiment 1 we utilized the entire original set of thirty themes.

The experiments were conducted between December 2012 and May 2013.[56] We used a web-based experimental platform called Qualtrics. Research using Qualtrics-based experiments has been published and presented in a number of academic fields, suggesting that it meets scholarly expectations for quality online web-based experiments.[57]

---

56. The University of Minnesota Institutional Review Board Human Subjects Committee determined that the study was exempt from review under federal guidelines. *See* 45 C.F.R. § 46.101(b)(2) (2014). The study, Study No. 1211E24181, is on file with the authors.

57. Studies relying on Qualtrics experiments include Jonathan S. Abramowitz et al., *Obsessive–Compulsive Symptoms: The Contribution of Obsessional Beliefs and Experiential Avoidance*, 23 J. ANXIETY DISORDERS 160, 162 (2009); Yany Grégoire et al., *When Customer Love Turns into Lasting Hate: The Effects of Relationship Strength and Time on Customer Revenge and Avoidance*, 73 J. MARKETING 18, 21 (November 2009); and Paul H. Robinson et al., *The Disutility of Injustice*, 85 N.Y.U. L. REV. 1940, 2004 (2010).

All subjects were recruited via modest, market-rate payments made available through Amazon Mechanical Turk's payment service.[58] Separate samples were recruited for each experiment. No personally identifying information was collected. Studies assessing the quality of Turk subjects have found them to be engaged by the online experimental stimuli and to be more representative than the convenience samples that would otherwise be used.[59] As discussed in more detail in Appendix B, filtering questions were used to ensure that subjects were actively participating throughout the course of the experiment.

All subjects recruited were self-reported United States citizens age eighteen to sixty-five. The number of subjects for each experiment, reported in Table 4, allowed sufficient statistical power to robustly test our hypotheses. At the end of the experiment, we collected demographic information from subjects. Table A1, located in Appendix A, shows these results. While not a truly nationally representative sample, the 1,613 subjects who participated in the experiments came from all 50 states, Washington, D.C., American Samoa, Puerto Rico, and the U.S. Virgin Islands. Our sample was roughly equal in terms of gender, with 52% of subjects being female and 48% male. Our subjects were younger on average than the comparable U.S. population. Our sample was 77% white, about equivalent to the national average. In terms of education, our subjects are slightly skewed toward having more education. Income distributions of our subjects and the U.S. population as a whole are similar, though not identical.

Table 4: Number of Subjects in Study by Experiment

Experiment	***Number of Subjects***
1a. Revised MPC Definitions Experiment - Sorting	96
1b. Revised MPC Definitions Experiment - Punishment	94
2a. Signal Variant Experiment - Sorting	531
2b. Signal Variant Experiment - Punishment	509
3a. Revised Recklessness Experiment - Sorting	186
3b. Revised Recklessness Experiment - Punishment	197
TOTAL SUBJECTS, ACROSS ALL EXPERIMENTS	*1,613*

58. No personally identifying information was collected aside from a thirteen-character ID number provided by the worker for the purposes of tracking survey completion, obtaining payment, and preventing the same individual from completing the same or related surveys.

59. *See supra* note 48.

C. Results

In this Section we report on the results from each of our three new experiments. Each experiment consists of two subparts: a sorting task and a punishment rating task. In Appendix A, we provide additional discussion of the details of the statistical analysis.[60]

1. Experiment 1: Revised MPC Definitions

The results from the sorting portion of the Revised MPC Definitions Experiment reveal the same basic pattern found in the original study: subjects are best at sorting purposeful and blameless mental states, worst at sorting recklessness, and in the middle with knowing and negligent mental states. A summary of sorting results is presented in Table 5, along with a comparison to the results of the original study.

Although the overall pattern remained the same, accuracy of N and R did increase in the new study. R improved from 40% to 47%, and N improved from 48% to 63%. This at least suggests that the new language may better communicate the distinction between these two mental states. But subjects in the new study were also less likely to correctly identify blameless scenarios, dropping from 88% to 78%.

Turning to the punishment task, we find that, even with the revised MPC definitions, subjects do not significantly differentiate between K and R in their punishment ratings across the thirty themes. Purposeful action was punished at 5.7, knowing at 4.9, reckless at 4.8, negligent at 3.4, and blameless at 1.4. Graphically, Figure 1 (which bears remarkable resemblance to the similar figure in the original study) clearly shows that, even with the revised definitions, there is no significant punishment differentiation between the knowing and reckless scenarios.

---

60. In this Section, we use several types of statistical analysis to assist us in drawing inferences from the data we collected. Using widely accepted methods, we estimate the likelihood that a difference in the sample means reflects a true difference in the population means. When we make an inferential statement in the body of the text (e.g., classification accuracy increased), this indicates that the statistical analysis (details available in Appendix A) indicated that there is a less than 5% chance that the effect we observed was due to chance. Another way of stating this is that our presented results are all statistically significant at a p-value of 0.05. This is the conventional standard for statistical significance in the natural and social sciences. Smaller p-values indicate a greater certainty that the observed effect is real. Statistical significance is not the same as legal or policy significance because a statistically significant difference is not necessarily great in magnitude. Whether an observed effect has legal significance involves policy and normative questions beyond the statistical test.

Table 5: Comparing Accuracy of Mental State Sorting: (A) Revised MPC Definitions Results and (B) Original Study Results

A. Revised MPC Definitions

	Correct Mental State				
	Purposeful	Knowing	Reckless	Negligent	Blameless
Subject chose: **Purposeful**	81%	7%	6%	2%	2%
Subject chose: **Knowing**	10%	53%	36%	7%	2%
Subject chose: **Reckless**	5%	27%	47%	18%	5%
Subject chose: **Negligent**	4%	9%	10%	63%	14%
Subject chose: **Blameless**	1%	3%	1%	10%	78%

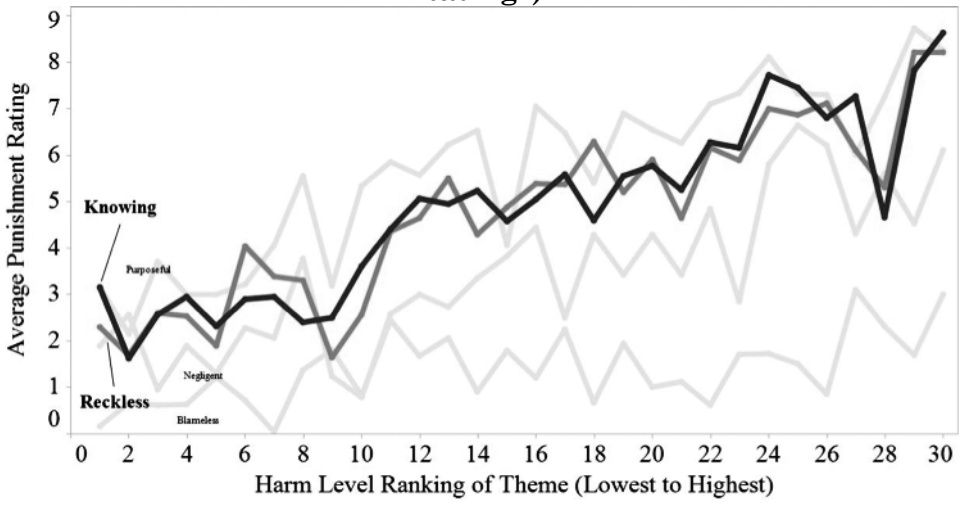
B. Original Study

	Correct Mental State				
	Purposeful	Knowing	Reckless	Negligent	Blameless
Subject chose: **Purposeful**	78%	9%	5%	2%	0%
Subject chose: **Knowing**	14%	50%	42%	5%	1%
Subject chose: **Reckless**	5%	30%	40%	31%	3%
Subject chose: **Negligent**	2%	10%	12%	48%	8%
Subject chose: **Blameless**	1%	2%	1%	15%	88%

What to Notice in Table 5: Even with the Revised MPC Definitions, subjects struggle to differentiate knowing from reckless and reckless from negligent. Subjects continue to do well at identifying purposeful and blameless actions.

Note: The gray cells in Table 5 display the sorting success rate for each mental state. The clear cells display the percentage of responses across the other four (incorrect) options. For instance, looking at the column labeled "Purposeful" in section A, subjects correctly identified these scenarios 81% of the time; 10% of the time mistook them for knowing; 5% of the time mistook them for reckless; 4% of the time mistook them for negligent; and 1% of the time mistook them for blameless.

Figure 1: Average Punishment Ratings for Scenarios from Experiment 1B ("Revised MPC Definitions, Punishment Rating")



What to Notice in Figure 1: Even using the Revised MPC Definitions, the average punishment ratings for knowing and reckless scenarios cross each other repeatedly, visually presenting what is confirmed by the statistical analysis discussed in Appendix A: there is no significant difference between punishment ratings of knowing and reckless scenarios.

Notes: Data for this figure are from the Revised MPC Definitions Experiment Punishment Rating task. The y-axis plots average harm rating for each purposeful, negligent, and blameless scenario in each of thirty themes (averaged across all subjects who rated the particular scenario). Shading indicates the mental state of the scenario.

2. Experiment 2: Signal Variant Experiment

In the next two experiments we tested the effects of changing the language used in our scenarios on sorting accuracy and punishment ratings. Specifically, in the Signal Variant Experiment we identified higher-accuracy and lower-accuracy variants of our signal for recklessness. Then, in the Revised Recklessness Experiment, we replaced the lower-accuracy language with revised phrases aimed at improving sorting and punishment differentiation.

Starting with the Signal Variant Experiment, our analysis revealed a robust effect of signaling language on subjects' ability to accurately identify reckless scenarios (Table 6). The results show that the R signals cleanly divided into two higher-accuracy R signals and three lower-accuracy R signals. The two higher accuracy signals were "understands that [the harm could easily happen], but decides to risk

it" (54% correct) and "aware there is a substantial risk that [the harm will occur], but chooses to ignore [it/the risk]" (52% correct). The three lower accuracy signals were "realizes it is very likely that [the harm will occur], but decides to [act] anyway" (42% correct), "conscious of the likelihood that [the harm will occur], but simply doesn't care" (39% correct), and "knows there is a good chance that [the harm will occur], but chooses to [act] anyway" (39% correct).

The results suggest that some words can better communicate recklessness than others. But even with the improved sorting results, our analysis revealed no statistically significant difference in punishment ratings across the various signals that we used to communicate the reckless mental state.

To further examine the nature of the relationship, we asked whether there was a significant correlation between an R signal's likelihood of being misinterpreted as a K signal and the mean punishment rating assigned to that signal. We found no correlation between these two factors.

Table 6: Sorting Accuracy of Recklessness in Experiment 2A

More Accurate R Signals:	***% Accurate***
1. Understands that [the harm could easily happen], but decides to risk it.	54%
2. Aware there is a substantial risk that [the harm will occur], but chooses to ignore [it/the risk].	52%
Less Accurate R Signals:	***% Accurate***
3. Realizes it is very likely that [the harm will occur], but decides to [act] anyway.	42%
4. Conscious of the likelihood that [the harm will occur], but simply doesn't care.	39%
5. Knows there is a good chance that [the harm will occur], but chooses to [act] anyway.	39%

3. Experiment 3: Revised Recklessness Experiment

The Revised Recklessness Experiment utilized the modified mental states definitions (as used in the two experiments just described) and also modified the signals for recklessness, as presented in Table 3. In brief, we modified the recklessness language by removing the choice language, which had been present in R signals but not K signals, and by reducing the probability of risk communicated in certain signals.

As presented in Table 7, we found that the adjustments to our R signals produced a marked improvement in sorting accuracy. The improvements in sorting accuracy were not limited to the R scenarios. The changes also seemed to improve our participants' ability to understand the K/R and R/N boundaries, as indicated by robustly improved classification of the K and N mental states as well (see Figure 2). This is despite no changes being made to how we communicated either of these mental states between Experiments 2 and 3.

We also investigated whether the changes we made to the signaling language—which resulted in higher sorting accuracy—were accompanied by other changes in the way subjects classified the various scenarios. For instance, it is possible that improvement in sorting accuracy for R scenarios is associated with an increase in the instances where subjects misidentify R scenarios as N scenarios while reducing instances where subjects misidentify R scenarios as K scenarios.

Comparing the results from the Revised Recklessness Experiment with the results using the Original Reckless Definitions Experiment, we found that there was no statistically significant difference in the breakdown of incorrect responses for the P and R mental states. There was, however, a significant difference in the distribution for the K and N mental states. This difference is driven by an increase in subjects misclassifying K scenarios as R in the Revised Recklessness Experiment, an increase in subjects misclassifying N scenarios as B, and a decrease in subjects misclassifying N scenarios as K or R. In sum, when we do see a statistically significant difference, we see subjects more frequently classifying scenarios into less culpable mental states as a result of our changes to the language of recklessness.

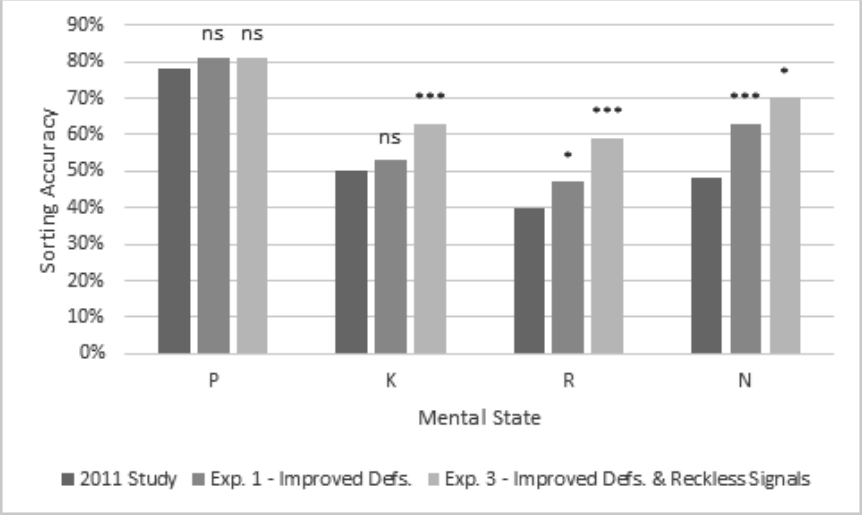
Table 7: Change in Sorting Accuracy of Recklessness: Experiment 2A vs. Experiment 3A

Experiment 2A Original Reckless Definitions	***Accuracy***	***Experiment 3A Revised Reckless Definitions***	***Accuracy (Change)***
Aware there is a substantial risk that [the harm will occur], but chooses to ignore [it/the risk].	52%	Aware there is a substantial risk that [the harm will occur].	65% (+13)
Realizes it is very likely that [the harm will occur], but decides to [act] anyway.	42%	Realizes there is **some risk** that [the harm will occur].	70% (+28***)
Conscious of the likelihood that [the harm will occur], but simply doesn't care.	39%	Conscious of the **real risk** that [the harm will occur].	53% (+14*)
Understands that [the harm could easily happen], but decides to risk it.	54%	Understands that [the harm could easily happen].	52% (-2)
Knows there is a good chance that [the harm will occur], but chooses to [act] anyway.	39%	**Recognizes** there is a good chance that [the harm will occur].	56% (+17*)

Notes: Table 7 compares the ability of participants to accurately classify scenarios when using the original reckless signals and when using the modified reckless signals (see Table 3). Aside from changes to the signaling language, no other changes were made to the scenarios or study design between Experiment 2 and Experiment 3. The number in parentheses in the fourth column notes the change in sorting accuracy in terms of absolute percentage points. * = Stat. Sig. at $p < 0.05$; *** = Stat. Sig at $p < 0.0005$.

What to Notice in Table 7: Reducing the communicated probability (e.g., from "very likely" to "some risk" and from "likelihood" to "real risk") improved the ability of participants to accurately identify the mental state. Further, removing the term "know," which might have further confused participants, from one of the reckless indicators also resulted in an improvement in sorting accuracy for that signal. While it is true that we also removed "choice language," the improvements we see are not even across the board, as would be expected if the improvements were due to the removal of such language. In fact, we only see statistically significant improvements on those signals that implement the modified probability language. This is strong evidence that removing the choice language was not a *primary* causal factor in the improvement we see in sorting accuracy.

Figure 2: Sorting Accuracy in Original Study, Revised MPC Definitions Experiment, and Revised Recklessness Experiment



Notes: The above figure presents the sorting accuracy for the culpable mental states across three experiments: the original study ("2011 Study") and the sorting components of the Revised MPC Definitions Experiment ("Exp. 1 – Improved Defs.") and the Revised Recklessness Experiment ("Exp 3 – Improved Defs. & Reckless Signals"). We denote the significance of the change between the original study and Revised MPC Definitions Experiment over the middle column bars in each cluster. We denote the significance of the change between the Revised Recklessness Experiment and the Revised MPC Definitions Experiment over the third column bars in each cluster. ns = $p > 0.05$; * = Stat. Sig. at $p < 0.05$; *** = Stat. Sig. at $p < 0.0005$.

What to Notice in Figure 2: This figure visualizes the net improvement that the various changes implemented to our paradigm revealed. Alterations to the mental state definitions result in modest improvements to R sorting accuracy and robust improvements to N sorting accuracy. We did not see substantial improvement to the various mental state categories until adjusting the language used to communicate recklessness to participants. Notably, this change not only improved sorting accuracy for R but for K and N as well. This is despite no alterations being made to these signals between the Revised MPC Definitions Experiment and the Revised Recklessness Experiment.

In the punishment rating task for the Revised Recklessness Experiment, we found no significant difference in the punishment ratings that subjects assess when employing the revised signals.

We then tested whether the improvement in sorting accuracy was followed by a greater separation in punishment ratings for K and

R scenarios. In other words, now that participants can better differentiate K and R scenarios, do they change the relative punishments assessed to K and R scenarios? We find that there is no change in the relative punishments given to R and K variants of the nine themes.

IV. DISCUSSION

There are both practical and theoretical implications from our results. At a practical level, the results offer concrete, if preliminary, guidance to courts and legislatures when constructing mental states definitions. While the mental states will likely always be difficult for a lay juror to grasp, we show that this confusion can be mitigated through careful word choice.

The results also contribute to an ongoing scholarly debate about the utility of the MPC's present construction of its mental state categories. We argue in this Part that our new results, while not yet dispositive, give strong support to the conclusions that lay subjects can, under the right conditions, see a distinction between knowing and reckless conduct causing harmful results, but that they typically fail to see a corresponding distinction in moral culpability.

A. Improving Sorting by Improving the Language of Recklessness

In our original study, we wrote that, if the distinction between knowing and reckless behavior is to have import in the criminal law, "legislatures and courts will have to do a better job of articulating it in their codes and jury instructions."[61] We also noted that, because our original study was "but one set of experiments in a young—indeed almost nonexistent—empirical literature," caution and additional study was needed.[62] One concern was that the language we used to define the several mental states and then to communicate these mental states in the scenarios could have been improved.

Our new results demonstrate that, indeed, the original language could be improved. By modifying the language of recklessness in the new experiments, we were able to bring about significant improvements to participants' sorting accuracy.

These results suggest that legal actors should not be cavalier in describing mental states. Small differences in wording—as evidenced

61. Shen et al., *supra* note 3, at 1347.

62. *Id.* at 1344.

in all of our experiments—can produce significant changes in perception of the mental state.[63]

While we can offer a general recommendation that the language be scrutinized, it is not yet clear exactly which words should be used in a particular context and what level of accuracy should be expected. The underlying question remains a normative one: What is the nature of the risk that makes it appropriate for one to be deemed criminally reckless?

Although the drafters of the MPC clearly meant for "substantial" to be interpreted contextually, we believe that they intended that a perceived probability of harm much less than 50% would suffice, at least in high harm cases.[64] Such a threshold of culpable risk creation is widely accepted and quite defensible in the types of factual scenarios used in our experiments—where the actor is aware of a genuine risk and creates or takes the risk without any morally plausible justification. Based on these observations, it became clear that certain words and phrases used in the original study communicated recklessness in a manner that materially contributed to lower classification accuracy by the participants.

Classification accuracy may be the goal of the sorting experiment, but courts must weigh competing concerns. For instance, experimenters might develop language that generates clearer distinctions between knowing and reckless conduct but in doing so may create a separation that is not consistent with the intent of the Code. Whether the MPC should be revised to improve subjects' ability to make distinctions or whether the confusion should instead be accepted as a necessary consequence of deliberately flexible definitions is a question beyond the scope of this Article.

Our results suggest that changing language can improve sorting, but they also suggest that those improvements are limited. Even in our best case, only 59% of subjects are accurately identifying R scenarios. More than one out of every three times they read a reckless scenario, subjects fail to identify it as such. About 70% of these misidentifications are subjects believing that an R scenario demonstrates knowing conduct on the part of the protagonist. We are still left with the basic conclusion we reached in the original study: laypeople have great difficulty identifying and distinguishing reckless and knowing behavior. If jurors cannot reliably distinguish between the two, on what basis are they deciding whether a defendant charged with murder acted knowingly or recklessly?

---

63. Additional research can investigate whether K/R accuracy can be further improved through more drastic changes to the experimental design. What if, for instance, subjects engaged in a short training exercise before reading and rating the scenarios?

64. *See supra* note 55 and accompanying text.

B. Punishment Ratings Are Unaffected by Improved Sorting

Despite the improvements in classification accuracy for the R mental state scenarios, we did not observe a statistically significant change in punishment ratings for the R mental state as compared with the K mental state. That our subjects improved their sorting, but did not similarly adjust their punishment ratings, lends support to the argument that we see the conflation of K and R punishment because subjects do not see a clear moral distinction between the K and R mental states, at least as it concerns the result element of offenses.

Two results particularly support this conclusion. First, as presented in Figure 2, we find that the propensity for an R signal to be misinterpreted as a K signal makes no impact on how subjects punish the scenarios containing that signal. This indicates that subjects do not reliably distinguish their punishment ratings even when perceiving a difference between the reckless and knowing mental states.

The results from Experiment 3 provide further support for this proposition. In Experiment 3, we saw significant and robust improvements in R classification accuracy, but subjects continued to demonstrate almost no difference in how they punish (1) cases in which they are told that the actor consciously took a significant risk that a victim would be harmed, with no semblance of a justificatory motivation for doing so, and the harm actually occurs; and (2) otherwise identical cases in which they are told that the actor knew that the harmful outcome was "practically certain" to occur.

What implications does this finding have for the law? The answer is not straightforward. Though the drafters of the MPC expressly adopted the K/R distinction and used it extensively in defining the mens rea for circumstance elements of criminal offenses, they rarely graded "result" crimes (i.e., offenses defined as engaging in conduct that causes specific harms) so as to punish knowingly causing the result more than doing so recklessly.[65] There is, however, one important exception to this pattern: the grading of homicide[66] (and a parallel distinction in assault[67]). Under the MPC and in virtually every state, a knowing murder (often called second-degree murder) is classified as significantly more serious than a reckless murder (often called manslaughter). If jurors see little or no moral distinction between the two, why does the law?

---

65. In other result offenses (of which there are not very many in the MPC), the offense is typically punishable at the same level for P, K, and R. *See, e.g.*, MODEL PENAL CODE § 221.1(2)(a) (1962) (inflicting harm during a burglary); *id.* § 250.11 (cruelty to animals).

66. *Id.* § 210.2.

67. *Id.* § 211.1(2).

A partial explanation for this puzzle may lie in the fact that the MPC and the law of most states also recognize a mental state in addition to the standard PKRN hierarchy, typically called "extreme recklessness ("eR")," that is generally treated as demonstrating a level of culpability commensurate with acts committed purposely or knowingly.[68] The effect of this additional mental state is to add a new moral gradient to differentiate R behavior from these more culpable mental states. While our studies have indicated that subjects do not see a moral distinction between K and R, it remains to be seen whether they are able to distinguish between R and eR. Of course, even if subjects were able to differentiate between R and eR, whether the eR mental state is a workable concept in practice depends on the extent to which juries are actually instructed on extreme recklessness in homicide cases and on how clearly the moral line between ordinary recklessness and extreme recklessness is defined.[69] We intend to explore these moral judgments in subsequent studies, including an effort to create scenarios that draw a reasonably defensible and reliable distinction between cases of "extreme" recklessness and "ordinary" recklessness.

C. Study Limitations

As in our previous work, we recognize that the implications of the experiments are limited in important ways by our experimental design.[70]

First, as with our original study, we cannot generalize beyond the online experimental context, which we continue to employ. It may be that jurors, when collectively deliberating, will understand and behave differently than when they are individually asked to render a punishment decision. Moreover, we cannot predict in any given case how the much richer set of facts (and their presentation through testimony and argument) will affect juror decision-making. Mock jury studies focusing on the attribution of mental states seem a promising

---

68. *Id.* § 210.2(1)(b) ("[Criminal homicide constitutes murder when] it is committed recklessly under circumstances manifesting extreme indifference to the value of human life."). In addition to the MPC, extreme indifference murder, sometimes called depraved-heart murder, has been adopted by most states.

69. In one of our authors' (Morris B. Hoffman) judicial experience, jurors in Colorado murder cases are rarely instructed on extreme recklessness, while jurors are routinely instructed on ordinary recklessness as a lesser included offense. This means that the difference between K and R is having huge effects on the outcomes of these cases because second-degree murder (requiring K) is punishable by mandatory prison up to forty-eight years while reckless manslaughter may be punished only by a term of probation.

70. Shen et al., *supra* note 3, at 1345–46.

avenue to address these concerns. Postverdict interviews with real jurors might also be fruitful.

Second, we readily acknowledged in the original study and again note here that, when mens rea is at issue in an actual case, jurors are not told what the defendant's mental state is, as they are in our scenarios. Real jurors must rely on descriptions of the defendant's conduct and the circumstances under which it occurred to infer mental states. And rather than receiving a single signal about the mental state, they will hear conflicting stories from the prosecution and the defense about what was going on inside the defendant's head during the alleged commission of the crime. How jurors synthesize this circumstantial evidence in order to arrive at a conclusion about the defendant's mental state remains a mystery ripe for further empirical investigation.

Third, we have again limited the focus of our experiments to *result* elements of crimes. Our experiments do not address *circumstance* elements of a crime—that is, elements having to do with the existence of a particular existing or historical fact (e.g., whether the property that defendant possesses is stolen, or whether the person with whom the defendant has intercourse is younger than sixteen). The results do not speak to whether people can distinguish between when a wrongdoer "knew" that a circumstance existed, was aware of a risk that it existed, or merely "should have known" that it existed. In separate work, we are now testing the MPC assumptions as they operate for circumstance elements.

Fourth, our modifications to mental state language do not address questions about the significance of the fact patterns themselves and whether laypeople are better at identifying knowing and reckless action when it arises from a particular type of behavior. In separate work we are investigating, for instance, whether crimes involving property damage result in blurring of the K/R boundary to a greater extent than crimes that result in bodily injury.

Fifth, our definitions and signals have aimed for simplicity—to isolate a single operative fact that distinguishes each mental state from the others. Obviously, criminal codes differentiate between mental states in numerous ways and often depart from the basic MPC five-level hierarchy. Extreme recklessness is one example that is particularly pertinent to result offenses.[71]

Finally, it is worth remembering that, outside the death penalty context and a few outlier states, judges (not jurors) typically perform the punishment function. Thus, future research should use a sample of judges for the punishment rating tasks.

---

71. *See* MODEL PENAL CODE § 210.2(1)(b); *id.* § 211.1(2)(a).

These limitations serve as an important reminder that the findings are incomplete and thus inadequate for deriving clear policy prescriptions. It has been said that "replication is the best statistic," and only with replication and further extensions of this work can it serve policymakers and courts in their specific formulation of mens rea.[72]

V. CONCLUSION

The fairness, utility, and effectiveness of the criminal justice system hinges on how well jurors can understand and apply the mens rea categories. Yet the mens rea categories are notoriously difficult to conceptualize and define, even for experts. Every day the subtleties of those categories used in most jurisdictions must be explained to jurors. And every day the effectiveness of those explanations remains uncertain.

It is vitally important that the language of mens rea conveys to actual jurors what the legal system has long assumed it will. Here, we have demonstrated that specific variations in the phrases used to define and to communicate criminal mental states can significantly increase an individual's ability to accurately classify mental states. Yet we also find that there are limits to the added value of new language. Despite the substantial changes we made to the language used to communicate recklessness, subjects continued to be categorically worse at accurately classifying reckless behavior compared to other mental states.

There are two practical lessons for the legal system. First, when it comes to communicating mental states, phrasing matters. Courts should therefore exercise care when considering the appropriate instruction; subtle variations may have substantial effects. And empirical legal scholars should provide courts with more data on what the effects of those variations are likely to be.

Second, our results raise deeper questions about the normative foundations of the MPC's mental states hierarchy. Improving subject accuracy in distinguishing between knowing and reckless behavior that causes harmful results did not translate into corresponding changes in the relative punishment that subjects would impose on knowing, as opposed to reckless, acts. Although real-life jurors are not typically called upon to decide punishment, it is nonetheless troubling that citizens apparently do not see the clear moral distinction that the MPC

72. STEVEN J. LUCK, AN INTRODUCTION TO THE EVENT-RELATED POTENTIAL TECHNIQUE 251 (2005) ("Replication does not depend on assumptions about normality, sphericity, or independence. Replication is not distorted by outliers. Replication is a cornerstone of science. Replication is the best statistic.").

presupposes between unjustifiably causing a criminal harm knowingly or instead recklessly.

VI. APPENDIX A: TECHNICAL AND STATISTICAL DETAILS

This Appendix provides additional detail on the research design employed in our study, the statistical procedures used to analyze the data, and the results of the statistical analyses.

A. The Participants

Participants were recruited through Amazon's "Mechanical Turk." Mechanical Turk is an online service provided by Amazon that allows individuals and institutions to offer online tasks (called "human intelligence tasks," or "HITs") to people across the country for pay. This service provided us with a sample that, while not truly nationally representative, was substantially more representative than convenience samples that would otherwise be used. In Table A1, we include the self-reported demographic information of the subjects included in the analysis.

Table A1: Demographics of Experimental Subjects (N = 1613)[73]

Education	***Subjects***	***U.S. Census***
Less than HS	1%	18%
High school / GED	11%	30%
Some college	31%	20%
Assoc. degree	10%	7%
Bachelor's	35%	17%
Graduate Degree	12%	10%

Income	***Subjects***	***U.S. Census***
< $20k	32%	$1 to $25k: 22%
$20k - $40k	29%	$25k to $35k: 19%
$40k - $60k	22%	$35k to $50k: 21%
$60k - $80k	10%	$50k to $65k: 14%
$80k - $100k	5%	$65k to $75k: 6%
> $100k	3%	$75k to $100k: 8%

Gender	***Subjects***	***U.S. Census***
Male	48% (42–56%)	49%
Female	52% (44–58%)	51%

Age Groups	***Subjects***	***U.S. Census***
18-24	22% (16–28%)	13%
25-34	42% (35–48%)	18%
35-44	20% (14–25%)	19%
45-59	13% (9–18%)	27%
60 +	4% (1–7%)	23%

Race	***Subjects***	***U.S. Census***
White	77%	74%
Non-White	23%	26%

Jury Member in Criminal Case?	***Subjects***
Yes	9% (5–14%)
No	91% (86–95%)

73. Some demographic information was only collected on a subset of the surveys. In those instances, we provide a bootstrap estimate of the 95% confidence interval for the population estimate to the right of the breakdown for the reported cases. If no range is provided, then that demographic data was collected from all 1,613 participants.

In the Revised MPC Definitions Experiment, we used approximately the same number of subjects as we used in comparable experiments in our original study. In the Signal Variant Experiment, we increased the number of subjects to account for the reduced number of observations per subject (since we were using nine themes instead of thirty), as well as the decrease in the expected effect size. We then reduced our sample size for the Revised Recklessness Experiment because, unlike the Signal Variant Experiment, we were not testing the differential effect of signaling language across four different mental states. Our statistical power requirements were thus substantially reduced for the Revised Recklessness Experiment.

Concerns about subjects' compliance with task instructions are of special concern with online experiments because subjects cannot be monitored while engaged in the experimental tasks.[74] To address this issue, experimental psychologists have developed "attention filters" designed to ascertain whether subjects are in fact following instructions and paying attention to the material being presented to them online. In each of our experiments, we employed a modified version of the filter developed by psychologist Daniel Oppenheimer and his colleagues.[75]

The design of the primary attention filter question was such that users who did not read carefully would see, in large font, a headline reading "Background Questions on Sources for News" as well as another large, bold question: "From which of these sources have you received information in the past month?" A series of check-box options were provided (e.g., local newspaper, local TV news). Subjects reading carefully, however, were instructed *not* to check any of the boxes, but instead to type "123" into the text box provided.[76] In several of our experiments we deployed an additional attention filter. This filter presented each subject with a scenario that appeared similar to other scenarios, except that it directed participants to select a specific response. The results presented in this Article are based only on those subjects who were paying attention as assessed by these attention filters.

---

74. A filter employed after data collection allowed for the experiment to exclude from the dataset subjects with duplicate IP addresses.

75. *See* Daniel M. Oppenheimer, Tom Meyvis & Nicolas Davidenko, *Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power*, 45 J. EXPERIMENTAL SOC. PSYCHOL. 867, 867–68 (2009) (describing a filter in which subjects must carefully read instructions which, counter to the boldface headline above the instructions, tell subjects not to actually click on an answer to the question).

76. Across the five experiments, 87% of subjects successfully answered the attention filter question.

B. The Experimental Paradigm

For the Revised MPC Definitions Experiment, the experimental paradigm remained largely unchanged from the original paradigm used in the original study and described in Section II.B above. There were, however, two changes to note. First, we transitioned to using participants recruited through Amazon Mechanical Turk instead of Qualtrics-recruited panels. Second, as discussed in the main text, we modified the MPC definitions (see Table 1). In brief, we crafted thirty distinct fact patterns and five variants of each fact pattern. The five variants served to manipulate the mental state of the actor in each scenario but nothing else. Participants saw only one of the five variants for each fact pattern and saw all fact patterns once. Therefore, participants saw thirty different fact patterns in total. The fact patterns were presented to participants in a pseudorandomized order.

For the Signal Variant Experiment and the Revised Recklessness Experiment, we used a modified version of this experimental paradigm. The primary modification was that for each fact pattern we crafted twenty variants instead of five. These twenty variants were comprised of five different signals used to communicate the four different mental states in the experiment (we did not test the blameless mental state). Instead of using thirty different fact patterns, we selected nine fact patterns from the original thirty. We selected those nine fact patterns where subjects demonstrated the greatest ability to distinguish the K and R scenarios in the sorting tasks and, on average, punished K greater than R in the punishment ratings task. Again, participants were presented with a pseudorandomized order of the nine stems, seeing no stem twice. Likewise, participants did not see the same signaling language twice, and the presentation of the mental states was evenly distributed for each subject.

Upon accepting the HIT on Amazon Mechanical Turk, all subjects were directed to the survey, which was hosted by Qualtrics. Before starting the survey, all participants were informed about the nature of the survey and their rights as participants in the study. At the start of the survey, we provided all subjects with the instructions necessary to complete the survey. Aside from changes to the mental state definitions, these instructions did not change between the three sorting tasks or the three punishment tasks. All subjects were exposed to five anchoring scenarios prior to starting the trials; these scenarios spanned the range of harm and intent that subjects were exposed to in the trial scenarios. Following the anchoring questions, participants began rating or sorting the trials of interest. Each trial was presented on a separate screen. On each screen, subjects were asked to read the

scenario and then either select a punishment level (in the punishment rating version) or identify the protagonist's mental state (in the sorting version). There was no time constraint placed on subjects' responses.

Punishment responses were provided on a 0-to-9 scale, with 0 being no punishment and 9 being the most extreme punishment the participant personally endorsed. Sorting responses were provided by participants clicking a radio button next to one of the five mental states (and accompanying definition). Both punishment and sorting responses were made at the bottom of the same screen that presented the scenario. Subjects then had to click another button to advance to the next trial. At the end of the trials, we presented the instructional manipulation check, followed by the demographic questions. Finally, subjects were debriefed and provided with a code to enter into Amazon Mechanical Turk in order to receive their compensation. We kept a record of each subject's Mechanical Turk ID in order to ensure that no subject completed a survey more than once.

C. Details of the Experimental Results

In this section we detail the statistical analyses that support the inferential conclusions we discuss in the body of the Article.

Starting with the Revised MPC Definitions Experiment, which tested the sorting accuracy of participants using the revised MPC definitions, we compared the sorting accuracy with our 2011 results using pairwise chi-squared tests. For reckless, we found that subjects using the revised definition were 1.3 times more likely to be correct than subjects using the original definitions ($\chi^2(1) = 4.26$, $p < .05$). For negligent, we found that subjects using the revised definitions were 1.8 times more likely to be correct ($\chi^2(1) = 18.93$, $p < .001$). We found no difference in the knowing condition, and for blameless, subjects using the revised definitions were 2.1 times more likely to be *incorrect* ($\chi^2(1) = 13.30$, $p < .001$).

We then examined the effect of the revised MPC definitions on punishment ratings. We first examined whether the changes were able to create a noticeable difference in how knowing and reckless scenarios were punished. A t-test comparing punishment ratings for knowing and reckless scenarios across the thirty themes revealed no significant difference $t(29) = .71$, $p = .23$.

In the Signal Variant Experiment we analyzed the effect of specific signaling language on subjects' ability to properly categorize the reckless mental state. A logistic regression analysis revealed a robust effect of signaling language on subjects' ability to accurately identify the reckless mental state scenarios ($\chi^2(4) = 21.04$, $p < 0.001$).

We next analyzed the effect of specific signaling language on how subjects punish scenarios describing reckless conduct. We ran a two-way ANOVA, with harm level of the scenario as one factor and signal as the other. This analysis did not reveal a main effect of signal ($F(4,505) = 1.781$, $p = 0.131$). As expected, there was a significant main effect of harm level ($F(2,507) = 684.5$, $p < .0001$). There was no significant interaction between variant and harm level ($F(8,501) = 1.414$, $p = .188$). We also examined whether there was a correlation between the likelihood that a reckless signal was misinterpreted as a knowing signal and the punishment rating that subjects assigned to scenarios using that signal. We found no reliable correlation ($r = 0.24$, $p = 0.70$).

In the Revised Recklessness Experiment, we assessed the effect of changes to the reckless mental state signals on subjects' sorting accuracy. We compared the sorting accuracy with the sorting accuracy in the Signal Variant Experiment using a 2-proportion z-test. The results are presented in Table 7. We also compared the sorting accuracy to the sorting accuracy of subjects in our original study as well as in the Revised MPC Definitions Experiment. All comparisons were made using a two-proportion z-test. We also examined whether the improvement in reckless sorting accuracy was accompanied by other changes to the sorting behavior. To test this, we observed the frequency of subject responses, broken down by the actual mental state presented in a scenario, and compared the results between the Signal Variant and Revised Recklessness Experiments. There was no statistically significant difference in the breakdown of incorrect responses for the P ($\chi^2(3) = 2.933$, $p = 0.402$) and R ($\chi^2(3) = 5.2$, $p = 0.158$) mental states. There was, however, a significant difference in the distribution for the K ($\chi^2(3) = 9.02$, $p < 0.05$) and N ($\chi^2(3) = 9.20$, $p < 0.05$) mental states. This difference is driven by an increase in subjects misclassifying K scenarios as R in the Revised Recklessness Experiment, an increase in subjects misclassifying N scenarios as B, and a decrease in subjects misclassifying N scenarios as K or R.

We next examined the effect of changes to our reckless signaling language on punishment ratings. We entered our data into a two-way ANOVA, with harm level of the scenario as one factor and signal as the other. This analysis revealed a marginal effect of signal ($F(4,193) = 2.321$, $p = 0.06$). As expected, there was a significant main effect of harm level ($F(2,195) = 173.3$, $p < .0001$). There was no interaction between variant and harm level ($F(8,189) = 0.207$, $p = .989$). As indicated, we do observe a trend towards significance with the main effect of signal variant. Post hoc analysis indicates that this trend is the result of nonsignificantly higher punishment ratings for the "aware

there is a substantial risk that [the harm will occur]" signal. We also examined whether our improved reckless signals had an effect on how subjects punished reckless as compared to knowing behavior. To accomplish this, we compared the difference in punishment ratings for the reckless and knowing variants of the nine themes. We found no reliable difference ($t(8) = -1.253$, $p = .246$).

We also combined the data from the Signal Variant and Revised Recklessness Experiments to test the marginal effects of our other mental state signals on sorting accuracy and punishment ratings, respectively. We found no effect of signal on sorting accuracy for P, K, or N. We performed a separate logistic regression analysis on each mental state, using signal variant as a categorical independent variable. This revealed no significant variation in sorting accuracy across signals for P ($\chi^2(4) = 1.731$, $p = .785$), K ($\chi^2(4) = 4.013$, $p = .404$), or N ($\chi^2(4) = 8.137$, $p = .087$). We did reveal some differences in punishment ratings that reached statistical significance. We ran three distinct two-way ANOVAs, with harm level of the scenario as one factor and signal as the other, for each of the three mental states. The models revealed no main effect of signal variant on punishment rating for P ($F(4,1521) = 1.075$, $p = .367$), a marginal effect—that does not reach the Bonferroni adjusted threshold of $p = 0.0167$—for K ($F(4,1514) = 2.899$, $p = 0.021$), and a robust effect for N ($F(4,1521) = 7.662$, $p < .0001$).

Post hoc analysis of the K mental state variants revealed that the main effect was driven by a single variant ("Understands that [the harm] is almost guaranteed to occur"), which subjects punished marginally less than the others. Post hoc analysis of the N mental state variants revealed that the effect was driven by higher than average assessed punishment on scenarios where we describe the act as being done "carelessly" and lower than average punishments on scenarios where we describe the act as being done "hurriedly and without noticing [a risk of harm]."

VII. APPENDIX B: FULL TEXT OF SCENARIOS

The full text of the scenarios used in the experiments discussed in this Article are available for download. The text of the scenarios used in the Revised MPC Definitions Experiment is available at: http://www.vanderbiltlawreview.org/content/articles/2014/08/Ginther-Revised-MPC.pdf. The text of the scenarios used in the Signal Variant and Revised Recklessness Experiments is available at: http://www.vanderbiltlawreview.org/content/articles/2014/08/Ginther-Signal-Variant-Experiment.pdf.