140 CHARACTERS OF JUSTICE? THE PROMISE AND PERILS OF USING SOCIAL MEDIA TO REVEAL LAY PUNISHMENT PERSPECTIVES

Itay Ravid[*]
Rotem Dror[**]

For centuries, penal theorists have debated two key criminal justice questions: justifying state punishment power and determining proper punishment levels. Moral philosophers offered several theories to address these questions. Over time, calls emerged to move beyond theories and to consider community views on punishment rationales in criminal law and policy design, an approach that gained support alongside meaningful critique. Concurrently, social science advancements enabled empirically deepening understanding of public attitudes about punishment, largely through surveys and experiments.

One domain, however, remained untouched by those calling to assess lay intuitions of justice: social media. Such oversight is puzzling in light of social media's potential to reveal public perceptions without scientific intervention. This Article thus engages with two main questions. First, a methodological question: whether social media discourse can be used to reflect laypeople's attitudes about criminal culpability and punishment, and second, a normative question: should it be used for these purposes?

To answer these questions, the Article first synthesizes current scholarship about the promises and challenges of using social media data to

[*] Associate Professor of Law, Villanova University Charles Widger School of Law, JSD ('20), JSM ('13), Stanford Law School.

[**] Postdoctoral Researcher, Department of Computer and Information Science, University of Pennsylvania, & Incoming Assistant Professor, Department of Information Systems, University of Haifa, PhD ('20), Technion – Israel Institute of Technology. We would like to thank Elliott Ash, Gilat Bachar, Michelle Dempsey, Renana Keydar, Roi Reichart, Hadar Dancing-Rosenberg, and the rest of the participants at the "Law in Virtual Spaces" workshop, the Richmond Junior Faculty Forum 2023, and the panel participants "Normatively Interrogating Punishment 1: Theorizing About the Law" at the Law and Society Annual Meeting 2023, for helpful comments and suggestions. Mary McDermott, Liam Reeves, and Dave Goulden-Naitove provided excellent research assistance. We are also grateful for the editorial board of the Illinois Law Review for their diligent work on this piece. Professor Dror would also like to thank the US Defense Advanced Research Agency ("DARPA") for their financial support. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, the Department of Defense, or the U.S. Government.

study human behavior and applies it to the context of punishment justifications. The Article moves beyond theory, however, and utilizes recent technological developments in the field of Artificial Intelligence ("AI") and Law and Natural Language Processing ("NLP") to offer a novel empirical exploration of the potential promise of social media discourse in assessing community views on justice and punishment.

While our findings offer some support for the potentiality of using social media to assess laypeople's attitudes regarding punishment, we also expose the complex challenges of utilizing such data, particularly for penal law and policy design. First, due to a host of methodological challenges, and second, due to normative challenges, particularly social media's polarizing nature and the ambiguity around who's voice is amplified through these platforms. The Article thus urges caution when leveraging social media to evaluate the public's perceptions of justice.

Table of Contents

I. INTRODUCTION

For centuries now, penal theorists—mostly scholars of criminal law and philosophy—have been actively engaged in discussions and heated debates about two questions that stand at the heart of our (and, in fact, any) criminal legal system: given that punishment is an infliction of pain and suffering of the State on individuals, *how can we justify* the State's power to punish, and if we can, *how can we determine* what is the appropriate level of punishment?[1]

In attempts to answer these questions, a number of theories have emerged, some dating back to as early as the Hammurabi's Code. Today, we can think of several main "types" of justifications for punishment; some are thousands of years old, while others are a product of more modern debates: retributive, utilitarian, and expressive.[2] Over the years, however, these debates were predominantly held in philosophical silos in which moral philosophers engaged in sophisticated hypotheticals in attempts to advance one theory over the other.

Discussions that started around the mid-1970s questioned the efficacy and necessity of maintaining these debates isolated in a philosophical ivory tower, outside of the realm of public opinion, and calls to better understand the views of "the people" regarding the justification of punishment when adopting penal laws and policies got traction, culminating in Paul Robinson's theory of "empirical desert."[3] According to this theory, one can potentially overcome tensions between the goals of "doing justice" and "fighting crime" by adhering to the community's notions of justice, as opposed to "moral philosophy's deontological notion of justice."[4] Advocates of this approach argue that lay people's attitudes can not only be rigorously studied through social science methodology but also that adopting shared community's notions of justice as a distributive principle increases the moral credibility of the criminal legal system, which in turn affects its crime control capacity.[5] Furthermore, relying on lay intuition of justice[6] contributes to the democratization of the criminal legal system or at least increases community involvement in criminal punishment.[7] Given these arguments, Robinson and others hold that penal law and policies should align with ordinary

[1] *See infra* Section II.A.

[2] *See infra* Section II.A. Scholars also offer a mix of these justifications, often framed as "mix theories" of punishment. Furthermore, note that in Section II we also discuss restorative justice under the umbrella of punishment theory although many agree that it should be considered an alternative to state punishment more than a justification in and of itself. *See infra* Section II; *see generally* Hadar Dancing-Rosenberg & Tali Gal, *Restorative Criminal Justice*, 34 CARDOZO L. REV. 2313 (2013).

[3] *See* discussion *infra* Sections II.C, III.

[4] Paul H. Robinson, *Empirical Desert, in* CRIMINAL LAW CONVERSATIONS 29 (Paul H. Robinson, Stephen P. Garvey & Kimberly Kessler Ferzan eds., 2009).

[5] *Id*. at 29–30. *See infra* Sections II.C, III.

[6] As opposed to experts (criminologists, academics, etc.).

[7] Paul H. Robinson, *Democratizing Criminal Law: Feasibility, Utility, and the Challenge of Social Change*, 111 NW. U. L. REV. 1565 (2017); Paul H. Robinson, Joshua Samuel Barton, & Matthew Lister, *Empirical Desert, Individual Prevention, and Limiting Retributivism: A Response*, 17 NEW CRIM. L. REV. 312, 368–369 (2014).

people's innate sense of justice. It should be noted, however, that the theory of empirical desert attracted significant critique from scholars on different fronts, either related to its hypotheses or its perceived unjustness.[8] In the context of social media, certain crucial critiques hold significant meaning, which we will delve into further at a later stage.

Criticism aside, advancement in social science methodology also contributed to the development of the theory by providing research tools to explore views and perceptions in the general population. As such, surveys and experimental research designs were adopted by researchers to explore questions revolving around lay justifications for punishment.[9] The vast majority of these studies generally revealed that retributive theories of punishment are the most dominant justifications in the eyes of the public.[10]

From a methodological perspective, one domain, however, remained surprisingly untouched: social media. We say surprisingly because if the goal is to understand how laypeople think, talk, and justify punishments, debates and discussions occurring in virtual domains, particularly social media, are potentially meaningful source for such data. Furthermore, social media, unlike experiments and surveys, potentially offers access to community views with minimum scientific or other formal intervention.[11]

Can social media data fulfill its potential to reflect laypeople's attitudes or the "shared judgments of the community"[12] regarding punishment? This Article offers some answers. It asks two main—related—questions in this context: First, methodological—can social media platforms serve as a reflection of the community's view of criminal justice? And second, normative—if the answer to the first question is yes, should we rely on social media in assessing lay people's intuitions of justice?[13]

In answering the first question, we build on legal and social science scholarship from various disciplines such as psychology, political science, and communication and explore the promises and challenges of studying social views

[8] Robinson, *supra* note 4, at 31–38, among the critique one can find, *e.g.*, the challenges of using desert as a distributing principle due to its vagueness, the difficulty of reaching consensus of the appropriate desert, the existence of other (utilitarian) principles defining intuitions about criminal justice, concerns about the results of relying on empirical desert due to its draconian or brutish nature, or its potential immortality. *See infra* Section II.C.

[9] These were utilized either to support or refute the premise of empirical desert, *see* discussion *infra* Part III. This Article recognizes that there is indeed value in understanding the views of different communities regarding culpability and punishment but debates how and whether such information should be utilized in practice.

[10] *See* discussion *infra* Part III. Robinson, *supra* note 7, at 1565–66 (exploring the reasons to support and criticize lay deference and finding that lay intuitions of justice include retributive proportionality).

[11] As much as we can consider social media discourse as "natural" or "observational" data. For example, speech on social media might not be completely free as it can reflect some forms of informal social "regulation." Still, it is considered a domain generally free from official content-related regulation (but note current debates regarding content moderation on social media platforms).

[12] Robinson, *supra* note 7, at 1565–66 (exploring the reasons to support and criticize lay deference and discussing previous studies indicating that lay intuitions of justice include retributive proportionality).

[13] Particularly given "Empirical Desert[s]" argument, that criminal laws and policies should align with those intuitions.

through social media platforms, particularly Twitter and Facebook.[14] We further apply the current scholarship to the particular issue of criminal law and social views regarding justifications for punishments.

To answer the second question—and then reflect on the first—the Article moves beyond theory and utilizes recent technological developments in the field of Artificial Intelligence ("AI") and Law, Natural Language Processing ("NLP"), and Automated Content Analysis ("ACA") to offer a novel empirical exploration of social media discourse regarding culpability and punishment. Given the novelty of the questions we investigated, we adopted an exploratory approach and used two different methodologies: first, an unsupervised ACA methodology called topic modeling ("TM") to explore latent themes and semantic fields present in the large data set social media produces. We complemented the analysis with a qualitative investigation within each identified theme. Second, exploratory text classification methodology utilizing GPT 3.5.[15] Particularly, we analyze Twitter posts around the verdict or sentencing decisions in four criminal cases that received significant media attention: Casey Anthony (2011), Aaron Hernandez (2015), Kimberly Potter (2022), and Nikolas Cruz (2022).

The analysis of thousands of tweets pointed to the methodological and normative complexities of utilizing social media to assess community justice judgments. The findings first revealed that focal questions of interest of Twitter users did not necessarily revolve around questions of punishment but rather questions of guilt or innocence. As such, social media might have limited capacity to assess shared community views about punishment. If one still wishes to study social media to assess lay people's attitudes regarding punishment, the methodology adopted should strive to overcome this challenge. Second, when culpability and punishment were discussed, however, the leading narratives were most closely aligned with what criminal law scholars will consider a retributivist approach (or "just deserts") to punishment. Third, there were potential connections between satisfaction from the outcome of a criminal legal process and trust in the criminal legal system. The second and third findings reflected what social scientists have previously identified in other empirical studies, mostly experiments. As such, these latter findings suggested that, indeed, under the limitations of the first challenge we identified, social media data can potentially offer a better representation of the general population than one might predict.

If the above findings indicated that there is a potential, if limited, to learn from social media about justice judgments in society at large, the fourth main finding, however, emphasized the normative challenges of following that path. The analysis revealed that the retributivism dominating social media discourse was often explained or discussed in conjunction with offensive, racist, and misogynist views. As such, the social media analysis offered some support for the

[14] *See* discussion *infra* Part IV. In this Article, we analyze Twitter data. While working on this project, meaningful changes have happened with regards to Twitter. Among these, its data access policies have changed, and its name was changed to "X". Whenever we mention Twitter we thus currently refer to "X".

[15] OpenAI, https://openai.com/ (last visited Aug. 21, 2023) [https://perma.cc/5ESP-NE6U].

empirical desert's immorality critique, emphasizing the risks of adhering to lay people's attitudes in designing penal law and policies. Our analysis further suggested that if one of the purposes behind empirical desert is to democratize criminal law, including amplifying voices of marginalized communities, then there might be a clash between the utilization of social media to assess lay people's attitudes and that purpose.

Overall, we argue that under certain conditions and methodological choices, social media can be a valuable tool for socio-legal scholars aiming to better understand community views about culpability and punishment. Nevertheless, we question the appropriateness of utilizing social media for the purposes of criminal law and policy design, especially due to the uncertainty regarding which voices are being magnified via social media platforms. That said, we consider this Article an invitation to further explore these questions normatively and empirically by utilizing current advancements in NLP methodology.

The Article proceeds in seven Parts. Part II discusses the evolution of moral theories of punishment and delves into the particularities of each approach as currently understood. It further discusses the justifications for theories such as empirical desert, calling to study and take into account laypeople's approaches to justification for punishment, and offers some critique of such theories. Part III surveys the current empirical landscape regarding studies aiming to assess laypeople's attitudes to punishment and offers a critique of existing designs. Part IV discusses the promises and challenges of using social media to learn about society at large. It also discusses potential solutions, methodological and others, adopted in other disciplines. Part V discusses the methodology adopted in this study and offers a quick overview of the cases selected for analysis. Part VI discusses the findings, and Part VII delves into the analysis and discussion, establishing our view about the promise, but mostly perils, of utilizing social media data to assess lay punishment perspectives. Part VIII offers our conclusion.

II. THEORIES OF PUNISHMENT: A BRIEF OVERVIEW

A. General

For centuries, penal theorists have been debating two core questions. First, how can we justify the State's power to punish individuals? Second, what amount of punishment should be inflicted? Historically, a number of theories rooted in moral philosophy were offered, all aiming to justify the use of imprisonment or any kind of suffering.[16] Punishment justifications are traditionally grouped into two main dominant groups: utilitarianism and retributivism.[17]

Utilitarianism, also known to some as consequentialism, suggests that punishment can be justified if its benefits outweigh its harms.[18] It is a forward-looking theory aspiring to prevent future criminal acts or any other benefits.[19] Traditionally, five main mechanisms that can potentially achieve that goal were identified: deterrence, incapacitation, rehabilitation, and more recently—denunciation (also known as the expressive function of punishment).[20] The emergence of deterrence theory can be attributed to Jeremy Bentham and his influential nineteenth-century writings.[21] Bentham believed punishment should be used as an end to a means: "[g]eneral prevention ought to be the chief end of punishment, as it is its real justification."[22] Bentham argued that punishment should be used to prevent crimes, finding that taking punishment out of a retribution context can then allow punishment to become a social good.[23] Deterrence theory comes in two forms: specific and general.[24] Specific deterrence aims to discourage the particular offender from "committing future crimes by instilling fear of receiving the same or a more severe penalty in the future,"[25] while general deterrence aims to discourage possible future offenders.[26] The aim of incapacitation is to prevent crime by physically restraining offenders, thus preventing them from future crimes.[27] Rehabilitation's purpose is to "reduce the offender's future criminality through education and treatment in prison or a nonprison program."[28] Denunciation will be discussed below in conjunction with the expressive goals of punishment.

Another mainstream philosophical foundation for punishment is retribution, which offers a deontological view of criminal punishments. Retributivists argue that punishment "is justified as an intrinsically appropriate, because deserved, response to wrongdoing."[29] Originally, retribution operated from the perspective of "an eye for an eye," essentially stating that "the state should punish

[16] *See generally* Zachary Hoskins & Anthony Duff, *Legal Punishment*, STAN. ENCYCLOPEDIA PHIL. ARCHIVE (Dec. 10, 2021), https://plato.stanford.edu/archives/sum2022/entries/legal-punishment/ [https://perma.cc/238B-YV62]; Joel Meyer, *Reflections on Some Theories of Punishment*, 59 J. CRIM. L., CRIMINOLOGY, & POLICE SCI. 595, 595 (1968) (exploring several theories of punishment and their foundations).

[17] *See id.*; *see also* Richard S. Frase, *Punishment Purposes*, 58 STAN. L. REV. 67, 70 (2005) (explaining the philosophical bases to the theories of punishment).

[18] *See* Frase, *supra* note 17, at 72.

[19] *See id.* at 72–73.

[20] *See id.* at 70.

[21] *Id.*

[22] JEREMY BENTHAM, THE RATIONALE OF PUNISHMENT 20 (1830).

[23] Morris J. Fish, *An Eye for an Eye: Proportionality as a Moral Principle of Punishment,* 28 OXFORD J. LEGAL STUD. 57, 63–64 (2008).

[24] *Id.*

[25] *Id.*

[26] *Id.* at 71.

[27] *Id.* at 70.

[28] *Id.*

[29] Hoskins & Duff, *supra* note 16.

those found guilty of criminal offences to the extent that they deserve, because they deserve it" (also known as "just desert").[30] Retribution, however, has evolved in contemporary times to focus on the punishment being *proportionate* to the crime rather than in total parity to the crime.[31] Immanuel Kant, the philosopher probably most known for advancing retributivist notions of punishment and for the idea that the use of proportionality to the crime derived from *lex talionis* to prescribe punishment.[32] Kant argued that "[j]uridical punishment can never be administered merely as a means for promoting another good either with regard to the criminal himself or to civil society, but must in all cases be imposed only because the individual on whom it is inflicted has committed a crime."[33] Furthermore, Kant argued that utilitarianist ideas could lead to situations in which those who wronged will not be punished, a decay of justice and righteousness and "if," according to Kant, "justice and righteousness perish, human life would no longer have any value in the world."[34]

In the middle of the twentieth century, and partially due to the challenges of reconciling retributivism and utilitarianism, some philosophers—with H. L. A. Hart and John Rawls as the most influential—offered what became known as "mixed theories" of punishments.[35] According to the mainstream version of these theories, both retributive and utilitarian principles are relevant in the consideration of punishment, but they simply answer different questions.[36] Utilitarianism justifies the "why" (*i.e.*, why the State can punish), and retributivism justifies the "how" (*i.e.*, how to punish wrongdoers).[37] These theories, however, were heavily criticized by others, preserving the tension between utilitarianism and retributivism.[38]

In an attempt to move away from either retributivism or utilitarianism, from the 1970s onward, we evidenced a new wave of philosophers calling to adopt another path to justify punishment through what became known as "expressive" theories of punishment. These theories are often traced back to Joel Feinberg's

[30] Hoskins & Duff, *supra* note 16. *See also* Kevin M. Carlsmith & John M. Darley, *Psychological Aspects of Retributive Justice*, 40 ADVANCES EXPERIMENTAL SOC. PSYCH. 193, 200 (2008) (offering a summary of retributivist principals).

[31] Hoskins & Duff, *supra* note 16.

[32] *See* Fish, *supra* note 23, at 62–63 (detailing Immanuel Kant's arguments for using retributivism as the main justification for punishment that shows the ancient law of *lex talionis* to have used proportionality rather than simple vengeance).

[33] IMMANUEL KANT, THE SCIENCE OF RIGHT 75 (W. Hastie trans., CreateSpace Independent Publishing Platform 2014) (1790).

[34] *Id.* at 76.

[35] Whitley Kaufman, *The Rise and Fall of Mixed Theories of Punishment,* 22 INT'L J. APPLIED PHIL. 37, 38 (2008).

[36] *Id.*

[37] *Id.* at 37–38. For the classic formulation of the mixed theory, see generally John Rawls, *Two Concepts of Rules*, 64 PHIL. REV. 3 (1955); H.L.A. Hart, *Prolegomenon to the Principles of Punishment, in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1, 1 (1968).

[38] *See, e.g.*, Alan H. Goldman, *The Paradox of Punishment*, 9 PHIL. & PUB. AFFS. 42, 42 (1979); Kaufman, *supra* note 35, at 51.

famous article, *The Expressive Function of Punishment*.[39] While there is "a family of views" under the expressive theories umbrella, scholars agree that the theory's core position is that "punishment is permissible"—at least to some extent—because it is society's best way to "express condemnation of the criminal offense."[40] Such condemnation is often aligned with retributivist ideas of punishment.[41] However, one can find support to the view that condemnation, under expressive theories of punishment, is not justified because it leads to certain outcomes (*e.g.*, deterrence) or because the offender deserves to be condemned (*i.e.*, retributivism).[42] It is merely the "last resort" of the State to reaffirm its values and to protect the dignity of victims when all other communicative forms are found unsuccessful.[43]

Finally, restorative justice should also be discussed in this context. Restorative justice is not a justification for punishment per se. Instead, it should be understood as an alternative to the use of state power in punishing offenders. The use of a restorative goal for punishment has been gaining traction in recent decades. Principles of restorative justice appear to have come about from the experience of practitioners working in the criminal legal system who were frustrated with perceived limitations from the traditional approaches.[44] The inspiration for restorative justice comes from non-Western community justice seen in Native American sentencing circles and New Zealand Maori Justice.[45] Restorative justice has been a more recent implementation following the "tough on crime" era seen in the late 1900s.[46] At its core, restorative justice diverts from traditional views of punishment while aspiring to advance accountability from the offender's side *vis-à-vis* the victims of the crime committed.[47] Restorative justice focuses on a process that offers a path for dialogue between offenders and victims while offering that the community play an important role in the restorative process.[48] In the American legal system, for example, "Neighborhood Justice

[39] Bernard E. Harcourt, *Joel Feinberg on Crime and Punishment: Exploring the Relationship Between the Moral Limits of the Criminal Law and the Expressive Function of Punishment*, 5 BUFF. CRIM. L. REV. 145, 145 (2001). *See generally* Joel Feinberg, *The Expressive Function of Punishment,* 49 MONIST 3 (1965).

[40] Joshua Glasgow, *The Expressivist Theory of Punishment Defended*, 34 LAW & PHIL. 601, 602 (2015). On expressive theories of law, see generally Cass R. Sunstein, *On the Expressive Function of Law,* 144 U. PA. L. REV. 2021, 2024 (1996); Elizabeth S. Anderson & Richard H. Pildes, *Expressive Theories of Law: A General Restatement*, 148 U. PA. L. REV. 1503, 1503 (2000); Richard H. McAdams, *A Focal Point Theory of Expressive Law*, 86 VA. L. REV. 1649, 1650 (2000).

[41] Hoskins & Duff, *supra* note 16.

[42] Glasgow, *supra* note 40, at 602–03.

[43] *Id*. Note, however, that some scholars consider expressive theories a form of utilitarian justifications. Frase, *supra* note 17, at 70 (discussing denunciation as a method to prevent future crimes).

[44] TONY F. MARSHALL, RESTORATIVE JUSTICE: AN OVERVIEW 5, 7 (1999).

[45] *Id.* at 15.

[46] *See id.* at 7 (explaining the rise of restorative justice uses as a result of frustrations with the limitations of traditional approaches).

[47] *Id.*

[48] *Id.*

Centers" have been used to divert certain offenders from the penal system to a mediation system.[49]

Regardless of which justification one advocates for, and as evident from the review above, for many years, the discussions pertaining to these justifications were dominated by criminal law or philosophy scholars. As such, they remained generally detached from the on-the-ground views of the many subjects of the criminal legal system: laypeople or "the public." In other words, discussions about which theory offers the most appropriate moral justification did not consider laypeople's attitudes with regard to these justifications. As the next Section will show, some changes have occurred on that front over the years, positing that laypeople's attitudes toward both the "why" and the "how" questions pertaining to punishment should also be taken into account.

B. Laypeople's Attitudes and Theories of Punishment

The influence of laypeople on theories justifying punishment became particularly clear with the politicization of penal policies,[50] mostly during the "tough on crime" era of the American Justice System.[51]

As discussed above, prior to the "tough on crime" era and for a relatively short period of time, rehabilitation aims were utilized in American penal reform.[52] This rehabilitation era was not based on popular views but rather on reports done by experts in the prison field.[53] This expertise-based era was short lived, however. In the 1970s, a shift occurred to focus on ensuring certain types of offenders were punished harshly, deterred from committing crimes and incapacitated by using longer sentences and longer paroles.[54]

The shift to more punitive ideals in sentencing was also due in part to both increases in crime rates and sensationalism by the media over crime problems, leading to public panic over perceptions of a flawed legal system.[55] Politicians found it popular—that is, in the eyes of the public—to appear "tough on crime"

[49] *Id.* at 15.

[50] Bruce Western and Christopher Muller, *Mass Incarceration, Macrosociology, and the Poor*, 647 ANNALS AM. ACAD. POL. & SOC. SCI. 166 (2013).

[51] *Id.* at 166–69.

[52] Jerome Hall, *Justice in the 20th Century*, 59 CALIF. L. REV. 752, 753 (1971).

[53] *See* HENRY KAMERLING, CAPITAL AND CONVICT: RACE, REGION, AND PUNISHMENT IN POST CIVIL WAR AMERICA, 111, 116 (2017) (explaining how penal reform centered around rehabilitation rose following the Civil War due to reports made by employees of different prisons in the United States).

[54] Western & Muller, *supra* note 50, at 166–69; *see generally* Judith Greene, *Getting Tough on Crime: The History and Political Context of Sentencing Reform Developments Leading to the Passage of the 1994 Crime Act*, SENT'G & SOC'Y: INT'L PERSPS., 1 (2002) (studying crimes rates of the 1970s–1990s, the legislation passed during this time period, the involvement of public movements (particularly the victims' rights movement), the effect both crime rates and legislation had on incarceration rates, and the social science research done at this time that influenced legislation).

[55] *See* Franklin E. Zimring, *Penal Policy and Penal Legislation in Recent American Experience*, 58 STAN. L. REV. 323, 330 (2005) (showing that the interactive relationship between the public, media, and lawmakers helped spur the "tough on crime"-era laws by examining prison trends from 1925 to 2000, penal legislation from the 1970s and on, and sentencing structures of the 1970s–2000s).

and thus imposed severe penalties for certain crimes that were plaguing America during the late 1970s–1980s.[56] Specifically, in 1974, the Federal Bureau of Investigation ("FBI") reported crime spikes, and the Attorney General, in turn, began critiquing "lenient judges" and policy makers who advanced rehabilitation as a punishment justification.[57] Further, sociological studies during this time came out finding that rehabilitative efforts had produced no change in levels of recidivism.[58] At the same time, liberal activists were rallying against the racial and class disparities that resulted from indeterminate sentencing schemes.[59] In response to criticisms from both sides, legislators began enacting policies utilizing both determinate schemes and retributive rationale (grounded in the theory of "just deserts").[60] State governments began adopting such policies in varying degrees across the late 1970s.[61] By 1979, almost half of the states had passed mandatory minimum laws, specifically for repeat offenders or crimes involving guns.[62]

From 1975 on, prison populations exploded as changes to penal policies were implemented.[63] Even as the FBI's reports on crimes showed a decrease, by 1982, the prison population had continued to explode.[64] The 1980s also showed politicians pushing for penal policies that did not concern themselves with the root causes of crimes but rather developing policies that would control the incapacitation and lead to harsh punishment of offenders.[65]

The political focus on crime at this time shifted to focus on "victim's rights," as illustrated by the 1988 presidential campaign of George Bush and his use of stories concerning violent offenders that wreaked havoc after being released from prison.[66] Again, leniency by judges and liberal penal policies in general were critiqued, while deterrence, incapacitation, and retributive justifications for punishment were pushed to the forefront.[67] The implementation of the "Three Strikes" laws of the early 1990s is probably the most striking illustration of this trend.[68] Meanwhile, in the 1992 presidential campaign, both Republican and Democratic candidates were now using the "tough on crime" narrative to solicit support.[69] In sum, throughout the 1970s and onwards, there was a clear profit by politicians in supporting policies that utilized deterrence,

[56] *Id.* at 333.

[57] *See* Greene, *supra* note 54, at 4.

[58] *Id.*

[59] *Id.*

[60] *Id.* at 6–7.

[61] *Id.* at 7–9.

[62] *Id.*

[63] *Id.* at 8.

[64] *Id.* at 12.

[65] *Id.* at 15.

[66] *Id.* at 18. For more on the victims rights' movement and its accomplishments, see generally Itay Ravid, *Inconspicuous Victims*, 25 LEWIS & CLARK L. REV. 529 (2021).

[67] *Id.*

[68] *See* Zimring, *supra* note 55, at 333.

[69] *Id.*

incapacitation, and retribution theories, thus leading to a focus on such theories in a bid for the approval of the public.

Considering that not only lawmakers but also prosecutors and judges occupy political positions, some deference to the laypeople they serve is to be expected. This is clear in Michael Nelson's research, which studied the influence of public opinion on prosecutors' and judges' behavior toward marijuana crimes.[70] Nelson studied a legalization initiative in Colorado to determine whether the prosecutor or judge would treat a drug offense involving marijuana in ways that align with the community's view of marijuana.[71] The research showed that district attorneys either behaved more leniently or harshly towards marijuana depending on whether the local opinion weighted in favor or against marijuana legalization.[72] The research further showed that judges responded to public opinion regarding marijuana legalization.[73] Further research has also shown that laypeople's opinion on even the use of the death penalty can sway judges.[74] Beyond controversial issues such as drug legalization or capital punishment, research suggests that general media reports on crime that reach the public can affect a judge's decision-making process.[75] One study showed, for example, that sentences were lengthened where there was increased media coverage of crime, and that these effects can vary based on the jurisdiction's method of selecting judges.[76] Thus, research shows that laypeople already influence–albeit indirectly–major parts of the criminal legal system.

Laypeople's influence on criminal laws can also be seen in the formation of the American Law Institute's ("ALI's") Model Penal Code ("MPC").[77] For example, Paul Robinson noted the MPC's deviation from traditional approaches to punishment appeared to show deference to "lay intuitions of justice."[78] Robinson specifically argue that the use of MPC's "excuse defenses," meaning the legal defenses of "insanity, involuntary intoxication, immaturity, and duress," were exemplary of how laypeople influence penal laws.[79] What supports this idea is the fact that classic deterrent strategies are undercut by the use of excuse

[70] *See* Michael J. Nelson, *Responsive Justice? Retention Elections, Prosecutors, and Public Opinion*, 2 J.L. & CTS. 117, 118 (2014).

[71] *Id.*

[72] *Id.* at 134.

[73] *Id.* at 118.

[74] *See* Paul Brace & Brent D. Boyea, *State Public Opinion, the Death Penalty, and the Practice of Electing Judges*, 52 AM. J. POL. SCI. 360, 362 (2008) (studying the influence that public opinion concerning capital punishment has on elected judges).

[75] *See* Itay Ravid, *Judging by the Cover: On the Relationship Between Media Coverage on Crime and Harshness in Sentencing*, 93 S. CAL. L. REV. 1121, 1127 (2021) (depicting empirical evidence to support the idea that media coverage affects judicial decision-making in criminal trials and arguing for ways to mitigate media's effect on judges).

[76] *Id.* at 1174.

[77] *See* Paul H. Robinson, *Why Does the Criminal Law Care What the Layperson Thinks Is Just? Coercive versus Normative Crime Control*, 86 VA. L. REV. 1839, 1840 (2000) (using the Model Penal Code to explore how punishment methods include lay intuitions of justice).

[78] *Id.* at 1841.

[79] *Id.* at 1842.

defenses because excuse defenses focus on the "blameworthiness" of the offender.[80] Robinson further explored the use of inchoate offenses as examples of lay-intuition deference.[81] The fact is that, from a deterrent perspective, it should not matter whether an offender was successful in their crime or not, but the MPC takes into consideration the proportionality of the harm done in choosing the punishment for an inchoate offender, which demonstrates lay-intuition deference.[82]

This use of proportionality—which derives from retributive principles—shows a consideration of laypeople's intuition. In fact, as will be elaborated in the next Section, many studies that have been conducted to determine what laypeople think about punishment have found that laypeople instinctively rely on retributive principles when considering punishment for an offender.[83] Specifically, laypeople's views are rooted in the retributive notion of proportionality.[84] What is interesting in this notion of proportionality is that what laypeople think of as proportional is often a far lighter punishment than what is actually written into law.[85] This then calls into question the legitimacy of the use of "tough on crime" policies if they do not reflect what laypeople actually think of punishment.[86]

The impact public opinion has had on the enactment of penal policies that have contributed to our problem of mass incarceration led scholars to advocate for leaving laypeople out of the conversation in creating penal policies.[87] This view is understandable, considering that the current inclusion of laypeople does not truly align with what laypeople think of crime when confronted with

[80] *Id.* at 1844.

[81] *Id.* at 1850.

[82] *Id.*

[83] *See* Carlsmith & Darley, *supra* note 30, at 199–200 (researching laypeople's psychological use of retributive justice through empirical studies).

[84] *See* Robinson, *supra* note 7, at 1580 (exploring the reasons to support and criticize lay deference and finding that lay intuitions of justice include retributive proportionality); *see also* discussion *infra* Part III (describing different studies aiming to assess how laypeople justify punishment).

[85] Robinson, *supra* note 7, at 1580 ("[A]lthough it may seem that community views on punishment are draconian or brutish, in reality, those views are rooted soundly in principles of proportionality and in fact seriously conflict with the harsh and disproportionate penalties found in many modern crime-control doctrines."); *see also* Paul H. Robinson, Geoffrey P. Goodwin & Michael D. Reisig, *The Disutility of Injustice*, 85 N.Y.U. L. REV. 1940, 1947 (2010).

[86] *See generally* Robinson, Goodwin, & Reisig, *supra* note 85 (showing that lay intuitions of justice conflict with actual laws); *see also* Leif P. Olaussen, *Concordance Between Actual Level of Punishment and Punishments Suggested by Lay People—But with Less Use of Imprisonment*, 2 BERGEN J. CRIM. L. & CRIM. JUST. 69, 71 (2014) (studying lay understanding of the Norway justice system through lay judges' proscription of sentences compared to professional judges); *see also* Liz Turner, *Penal Populism, Deliberative Methods, and the Production of "Public Opinion" on Crime and Punishment*, 23 GOOD SOC'Y 87, 89 (2014) (arguing that certain methods for gaging public opinion fail to truly show the public's view of punishment).

[87] *See generally* Anthony Bottoms, *The Philosophy and Politics of Punishment and Sentencing*, *in* THE POLITICS OF SENTENCING REFORM 40 (Clarkson and Morgan eds., 1995). For more on "Penal Populism" see generally JOHN PRATT, PENAL POPULISM (2007).

punishing a potential offender.[88] Critics point out that laypeople are uneducated about both punishment approaches and the actual laws governing punishment.[89] It would thus make sense to instead rely on scholars educated in the necessary fields to help reform the system and eradicate mass incarceration.[90] But other scholars (the leading among them is Paul Robinson) advocate for a shared criminal legal system informed by the "empirical desert" theory, which calls to adopt laypeople's (or "community") justice judgement on liability and punishment.[91] The next Section further explores arguments for and against incorporating lay intuitions on punishment theory into penal policy design.

C. Do Laypeople's Attitudes Matter?

According to a number of scholars, the inclusion of laypeople in reforming penal institutions would give a variety of benefits, and further, the exclusion of laypeople in consideration of punishment goals will not fix the current problems.[92] Scholars have identified three (related) main benefits from using laypeople's understanding of punishment in designing penal policies: (1) legitimizing the penal institutions, (2) supporting deterrence aims, and (3) increasing participation of laypeople in penal institutions (as a form of education and responsibility enhancement mechanism).[93]

First, including laypeople will help legitimize penal institutions because it would reflect the contemporary social beliefs of society and thus carry more "moral credibility" with the public.[94] Moral credibility is important because it promotes better trust in the government.[95] Albert Dzur's argument for including laypeople notes that hiding the criminal process from the public would be akin to the relationship the public had with the government following the response to

[88] *See* Robinson, *supra* note 77, at 1841 (showing that lay intuitions of justice conflict with actual laws); *see generally* Olaussen, *supra* note 86, at 91 (studying lay understanding of the Norway justice system through lay judges' proscription of sentences compared to professional judges); *see also* Turner, *supra* note 86, at 90 (arguing that certain methods for gaging public opinion fail to truly show the public's view of punishment).

[89] *See* FRANK E. ZIMRING, GORDON HAWKINS, & SAM KAMIN, PUNISHMENT & DEMOCRACY: THREE STRIKES AND YOU'RE OUT IN CALIFORNIA, 181–91 (2001).

[90] *Id*.

[91] Robinson, *supra* note 7, at 1580; Albert W. Dzur, *The Myth of Penal Populism*, 24 J. SPECULATIVE PHIL. 354, 360 (2010) (discussing the critique of laypeople's influence on the contemporary criminal legal system, also known as "penal populism," while supporting the adherence to lay participation).

[92] *See* Robinson, *supra* note 7, at 1580 (arguing for the inclusion of laypeople on the basis that it will not solve the criticisms against laypeople inclusion and that laypeople are a fundamental part of the criminal legal system); *see also* Robinson, *supra* note 77, at 1839 (reasoning the benefits including lay intuitions of justice when determining penal policies); Robinson, *supra* note 7, at 1566; Dzur, *supra* note 91, at 360.

[93] *See* Dzur, *supra* note 91, at 360–62 (arguing that inclusion of lay deference in penal policy-making is fundamental to the criminal legal system and American democracy more broadly); *see also* Robinson, *supra* note 7, at 1580–88 (depicting the benefits that come from including laypeople consideration in criminal law); Robinson, *supra* note 129, at 1861 (exploring the value of including lay intuitions of justice in the Model Penal Code).

[94] Robinson, *supra* note 7, at 1581.

[95] Dzur, *supra* note 91, at 360–61 (arguing that the distrust seen in the public towards the government following the 2008 recession will be seen in the criminal justice system if laypeople are not included).

the 2008 financial crisis in that it would only create skepticism towards government actions.[96]

Legitimizing penal institutions is also connected to the second benefit identified in the literature: supporting deterrence goals. The increased legitimacy stemming from the adherence to laypeople's views on theories of punishment could actually deter potential offenders better than the current systems' use of harsh sentences.[97] Furthermore, the use of laypeople's intuitions of justice would independently aid deterrence efforts in that inclusion of community views of punishment lend the criminal law moral credibility, which "can harness the power of stigmatization."[98] The power of stigmatization is that it not only costs less than imprisonment, but it can "endanger . . . personal and social relationships."[99] As Robinson notes, punishment for an action that society is less likely to condemn does not garner much respect for the government or lead to social condemnation for that action.[100] By using communal views to increase trust in the government and display what is currently condemned by society, respect for the law that properly reflects societal views of criminal law may increase.[101] These benefits reflect deterrence efforts in that the aim of the punishment is to prevent future offenders from choosing the criminal path, and at least according to Robinson, it purports the punishment to be a societal good as it reflects societal morality.[102]

Third, including laypeople's views in determining punishment supports the public's inclusion in the criminal legal system.[103] The public needs to be included in the justice system because the public has conceptually always been a part of the justice system to begin with.[104] When a criminal trial begins, the State is only considered a party to the trial because the public—not just the individual victim—is considered to be harmed by the criminal offense, which allows the State to act on behalf of the people when it prosecutes an offender.[105] Beyond the traditional approach to the criminal legal system, restorative justice, which has been gaining traction as an alternative to traditional punishment, inherently considers the public a participant in the healing process.[106] As previously noted, restorative justice involves the offenders, victims, community members, and some government entity working together to meet the "*victims' needs*," "to

[96] *Id.*

[97] *See id.* (arguing that laypeople's stigmatization of offenses will help prevent future offenders and enforce social punishment outside of legal punishment).

[98] Robinson, *supra* note 7, at 1581.

[99] *Id.*

[100] *Id.*

[101] *Id.* at 1581–82.

[102] Frase, *supra* note 17, at 70 (explaining that deterrence aims of punishment are couched in utilitarian philosophy because it focuses on the prevention of future crimes).

[103] *See* Dzur, *supra* note 91, at 365 (arguing that inclusion of lay intuition is necessary because the public are fundamental parts of the justice system).

[104] *Id.*

[105] *Id.*

[106] MARSHALL, *supra* note 44, at 6.

prevent re-offending by *reintegrating offenders* into the community," and "to provide a means of *avoiding escalation* of legal justice and the associated costs and delays."[107] As such, it is important to include laypeople in such a process as the actual community members that reintegrate the offenders.[108] Therefore, in both traditional and newer approaches to justice, laypeople are fundamental parts of the system and thus necessary to include. Further, as argued by Robinson, laypeople generally impose more lenient punishments than the government currently uses.[109] Having a transparent view of the criminal process could thus—at least in theory—help reform current laws to better reflect the public's view, which in turn could help lighten mass incarceration problems.[110]

As mentioned, some disagree with the above claims and call to exclude laypeople's justice judgments from the set of considerations used by the government in designing penal policies. Those in support of this view largely blame the adherence to laypeople's attitudes from around the mid-1970s and through the 1990s as one of the main causes of mass incarceration due to "penal populist" policies that came about during the "tough on crime" era.[111] As such, those criticizing empirical desert find this theory to be potentially unjust as it lends support to harsh populist views on punishment and expansion of the prison system.[112] Relatedly, others argue that relying on individuals' intuitive morals opens the door for considerations that should be left outside of the criminal system including racism, misogyny, xenophobia, and more.[113] But when exploring the effects of both including and excluding laypeople, it remains unclear whether excluding laypeople will solve penal populism.[114] Even if policy-making was transferred out of the hands of legislators, prosecutors, and judges and into an expert board, the appointment of board members would likely become politicized as well.[115] Further, some argue that the fact that penal populism has been shown in some instances not to align with laypeople's intuitions shows that the "tough on crime" laws were not made using actual lay intuition but rather were made by and for

[107] *Id.*

[108] Dzur, *supra* note 91, at 372.

[109] *See generally* Robinson, *supra* note 7 (showing that lay intuitions of justice conflict with actual laws); Olaussen, *supra* note 86; Turner, *supra* note 86 (arguing that certain methods for gaging public opinion fail to truly show the public's view of punishment).

[110] *See* Robinson, *supra* note 7, at 1580–88 (arguing that inclusion of lay intuitions will better support the implementation of criminal laws).

[111] *See generally* Pratt, *supra* note 87; NICOLA LACEY, THE PRISONERS' DILEMMA: POLITICAL ECONOMY AND PUNISHMENT IN CONTEMPORARY DEMOCRACIES 18 (2007); Zimring, Hawkins & Kamin, s*upra* note 89.

[112] Robinson, *supra* note 4, at 35–36; for the effects of penal populism, see also LACEY, *supra* note 111; ZIMRING, HAWKINS & KAMIN, *supra* note 89, at 201–203.

[113] Deborah W. Denno, *The Perils of Public Opinion*, 28 HOFSTRA L. REV. 741, 752–753 (2000). *See also* Ilya Rudyak, *Promoting Equality Through Empirical Desert*, 7 TEXAS A&M L. REV. 187 (2019) (discussing the critique of empirical desert based on "immorality objections" but offering reconceptualization of the theory which, according to the author, immunizes it from the immorality critique alongside additional objections offered in the literature).

[114] *See* Dzur, *supra* note 91, at 364–66 (laying out reasons why the complaints against inclusion of laypeople in the criminal legal system will not solve the current problems and will further lead to more problems).

[115] *Id.* at 365.

politicians to appear "tough on crime" without truly considering the aims of punishment.[116] In fact, as mentioned earlier, Robinson and others suggest that contrary to the prescribed legal punishments, the penalties administered by lay people were comparatively less severe. [117]

Excluding laypeople, however, will likely serve to isolate the government from the public's trust.[118] This lack of trust will only make criticism of the government more severe when penal policies fail to produce meaningful results.[119] So too, if criminal laws define offenses without consideration of public views of what offenses should be prosecuted, adherence to criminal laws and the possibility of social punishment will likely be negatively affected.[120] Empirical studies indeed have connected the functionality of the legal system and the level of trust laypeople had in the institution, and that trust in the system was "a vital factor in legal compliance."[121] Further studies have shown how public opinions affected prosecutors and judges implementation of actual laws.[122] Where marijuana legalization is supported, the prosecutors will choose not to prosecute marijuana laws because the public does not view those laws as worthy of being followed.[123] Excluding laypeople is thus not only unfavorable to supporting a criminal legal system but is practically difficult when so many aspects of the system are politicized and dependent on approval from laypeople.[124]

Finally, the exclusion of laypeople implies that laypeople are incapable of contributing to the criminal legal system.[125] It suggests that laypeople cannot handle the responsibility of being informed and active in the legal system and thus, the system as a whole.[126] As Dzur noted, "[i]t is to say that the public, like

[116] *See generally* Robinson, *supra* note 7 (showing that lay intuitions of justice conflict with actual laws); Olaussen, *supra* note 86; Turner, *supra* note 86.

[117] Joshua Kleinfeld & Hadar Dancig-Rosenberg, *Social Trust in Criminal Justice: A Metric*, 98 NOTRE DAME L. REV. 815, 863–64 (2022).

[118] *See* Dzur, *supra* note 91, at 364 (explaining the lack of lay deference in fiscal policies has led to major distrust between the public and government concerning the 2008 recession).

[119] *See id.* (arguing that the exclusion of laypeople will only lead to even more severe criticism from the public for the government's faults in the criminal system than what is occurring now).

[120] Robinson, *supra* note 7, at 1565.

[121] *See* Kleinfeld & Dancig-Rosenberg, *supra* note 117, at 845 (advocating for a metric based on social trust in assessing criminal systems).

[122] *See* Nelson, *supra* note 70, at 117–18 (researching how support or opposition of marijuana legalization affects the implementation of drug laws that involve marijuana in Colorado); *see also* Brace & Boyea, *supra* note 74, at 360 (researching how local opinions of the death penalty affect judicial decision-making for judges that retain elected positions); Ravid, *supra* note 75, at 1166–69 (showcasing empirical data to prove that media coverage over criminal cases can cause judges to lengthen sentences and generally affect judicial decision-making).

[123] Nelson, *supra* note 70, at 118.

[124] *See id.* (showing that local opinions of marijuana affect whether prosecutors and judges will prosecute marijuana offenses); *see also* Brace & Boyea, *supra* note 74, at 360 (showing that local opinions will sway how judges affirm cases involving the death penalty); Dzur, *supra* note 91, at 364–66 (supporting the inclusion of laypeople by weakening the criticism that exclusion would be better for reform of penal policies).

[125] Dzur, *supra* note 91, at 365.

[126] *See id.* at 364 (arguing that "failure to engage the public is risky because sealing off the criminal justice process does nothing to educate, 'responsibilize,' or build trust, which is what experts and professionals require

a criminal offender, is careless regarding the lives of others and needs restraints, expert guidance to dampen down normally poor impulse control."[127] Such an implication will only add to the distrust that an exclusionary, opaque system would create.[128]

As the next Section will further discuss, studies show that laypeople's intuitions focus on retributivism and proportionality but that, when well-informed, laypeople can utilize more utilitarian principles as well.[129] This, in turn, shows that laypeople can properly understand the complex reality of justifications for punishment and do not only rely on what some consider the more barbaric aspects of retributivism (*e.g.*, "eye for an eye" as a guiding principle).[130] Considering that including laypeople in reforming a criminal legal system will potentially democratize the system, help increase respect for the government and its laws, and give respect to the laypeople who this system has been created for, institutional deference to laypeople may indeed be beneficial.

At the same time, this potential outcome seems too ideal, as merely suggesting taking into account lay people's attitudes does not resolve questions related to whose voices are de-facto amplified when exploring "lay" attitudes. To some, empirical desert by itself does not guarantee increased public participation in its deepest sense, that is, full participation of different groups in society, including those who were traditionally excluded from the decision-making table. These concerns similarly challenge the justness of the empirical desert theory and offer more support to its immorality critique. The theory was further criticized on several additional grounds. For example, some scholars questioned the use of desert as a distributive principle,[131] while others raised doubts regarding the theory's core assumptions.[132] Among these is the assumption that one can ever point at "community" views of justice, given the deep disagreement on issues of crime and punishment. [133]Robinson responded to this critique by providing empirical support that at least for the "core of wrongdoing," there exists a substantial consensus among all demographics.[134] Additionally, one may raise a normative distinction between the value of *learning about* the views of

to do their work," thus implying that engaging the public would as well gain responsibility in being an active participant in the criminal legal system).

[127] *Id.* at 365–66.

[128] *See id.* at 363–66 (determining that a nontransparent system for the public will not be beneficial).

[129] Carlsmith & Darley, *supra* note 30, at 200 (researching laypeople's psychological use of retributive justice through empirical studies).

[130] *See* Fish, *supra* note 23, at 58 (arguing that the use of retributivism is not barbaric as some critics argue—but focused on proportionality in the implementation of punishment); *see also* Carlsmith & Darley, *supra* note 30, at 194 (concluding that laypeople utilize proportionality factors when proscribing punishment).

[131] *See* Alice Ristoph, *The New Desert, in* CRIMINAL LAW CONVERSATIONS 48 (Paul H. Robinson, Stephen P. Garvey & Kimberly Kessler Ferzan, eds., 2009).

[132] *See, e.g.*, Christopher Slobogin & Lauren Brinkley-Rubinstein, *Putting Desert in Its Place*, 65 STAN. L. REV. 77, 79 (2013); Michael D. Cahill, *A Fertile Desert?*, *in* CRIMINAL LAW CONVERSATIONS 43 (Paul H. Robinson, Stephen P. Garvey & Kimberly Kessler Ferzan eds., 2009); Christopher Slobogin, *Some Hypotheses About Empirical Desert*, 42 ARIZONA ST. L.J. 1189 (2011).

[133] Robinson, *supra* note 7, at 1567.

[134] *Id.*

communities regarding culpability and punishment, and the suggestion to implement such views in the design of criminal law and policy, as advanced by the empirical desert theory.[135] Some of these concerns, and others, will be further discussed in section V.

III. ASSESSING LAYPEOPLE'S ATTITUDES: THE CURRENT EMPIRICAL LANDSCAPE

Since the mid-1980s, and with the realization that lay attitudes regarding punishment could be beneficial from the perspective of different stakeholders in the criminal legal system, we have experienced a significant increase in studies that adopt social science methodologies to empirically assess questions related to lay people's attitudes toward the criminal legal system. The research in this domain seems to have taken similar paths to other social science studies aiming to assess public perceptions, that is, heavier use of direct surveys early on with a transition into methodologies that better assess causation, like lab and online survey experiments.

One of the earliest attempts to move beyond general views regarding punishment towards examining the motivations behind why individuals support a particular form of punishment was conducted by Ellsworth and Ross.[136] The study—which centered on perspectives regarding the death penalty—revealed that participants' justifications were most likely reflective of their existing positions on the issue and not a product of "reasoned and knowledgeable investigation of the factual issues involved."[137] While some inconsistencies in reasoning were elucidated by this method,[138] Ellsworth and Gross concluded in later studies[139] that surveys that ask direct questions of participants were relatively rudimentary methodologies.

As such, further studies which relied on participant self-reports should be interpreted with these limitations in mind. For example, Weiner, Graham, and Reyna[140] attempted to relate the objectives of imposing punishments to attributions related to the cause of a particular crime based on attribution theory,[141]

[135] *See* Ristoph, *supra* note 131.

[136] *See generally* Phoebe C. Ellsworth & Lee Ross, *Public Opinion and Capital Punishment: A Close Examination of the Views of Abolitionists and Retentionists*, 29 CRIME & DELINQ. 116, 117 (1983) ("Although there are numerous empirical studies that provide us with a record of the changing levels of overall support for capital punishment, there are very few that attempt to probe deeper, to understand what people mean when they say that they favor or oppose the death penalty.").

[137] *Id*. at 162.

[138] For example, while Ellsworth & Ross' study found some support for deterrence as a key factor in opinions for and against the death penalty, these views were not changed even after statements indicating the inefficiency of the death penalty were presented. *Id.*

[139] *See generally* Phoebe C. Ellsworth & Samuel R. Gross, *Hardening of the Attitudes: Americans' Views on the Death Penalty*, 50 J. SOC. ISSUES 19 (1994).

[140] Bernard Weiner, Sandra Graham & Christine Reyna, *An Attributional Examination of Retributive Versus Utilitarian Philosophies of Punishment*, 10 SOC. JUST. RSCH. 431, 431 (1997).

[141] *Id.* at 432.

which concerned the causality (or rather perceived causality) behind a particular outcome.[142] The study drew upon several methodologies.[143] The study, again attained primarily by participants' self-reporting, found that people asserted justifications for punishment based on their beliefs about what caused a crime. As such, the findings of the study remained in question, as the data could have merely reflected "post-hoc justifications for the participants' attitudes rather than the casual antecedents."[144]

Other social psychology research that has focused on decision-making has employed subjective expected-utility models, which present individuals with preselected information sets based on different behavioral alternatives and their expected utility.[145] The data gathered, however, is not only a function of the presented options but also subject to further self-reporting when participants are asked to discuss their choices.[146] Other studies that apply this model simply ask subjects to list information they believe they considered when making a decision and thus may be affected by *post hoc* biases, rationalization, and memory lapses.[147]

On top of these challenges of self-reporting discussed above, social-psychology research is plagued by the fact that most mental processing occurs subconsciously, or too rapidly, to fully describe with the conscious mind. Given the challenges mentioned above, Policy Capturing ("PC"), sometimes also referred to as Behavioral Process modeling ("BP"), has become a viable methodology that addresses the methodological deficiencies of studies based on self-reporting.[148] The methodology has been used as a statistical method in social psychology to quantify the relationship between a judgment and the information used to make that judgment without relying on direct introspection by the participants.[149]

One of the earlier studies using a BP model was done by Jacoby, Jaccard, Kuss, Troutman, & Mazursky in 1987.[150] Their goal was to describe and use procedures that control the information inputs involved in participants decisions, allowing them to be traced and identified in a more rigorous manner.[151] While this process method was a step forward from simple self-reporting studies, some limitations that were nevertheless encountered by Jacoby *et al.* were that results were recorded in verbal format (thus subjecting the study to some of the limitations of self-reporting studies) and the controlled nature of the information

[142] *Id*.

[143] *Id*. at 446–47.

[144] Kevin M. Carlsmith, *The Roles of Retribution and Utility in Determining Punishment*, 42 J. EXPERIMENTAL SOC. PSYCH. 437, 439 (2006).

[145] *See* Jacob Jacoby, James Jaccard, Alfred Kuss, Tracy Troutman & David Mazursky, *New Directions in Behavioral Process Research: Implications for Social Psychology*, 23 J. EXPERIMENTAL SOC. PSYCH. 146, 148 (1987).

[146] *See id.*

[147] *Id.*

[148] *See id.* at 149–59.

[149] *See id.* at 149.

[150] *See generally id.*

[151] *Id.* at 146–49.

provided to the participants.[152] It was concluded, however, that BP methods could be tailored to isolate some of these problems as well as provide a large degree of customizability to the experiment being conducted.[153]

BP methods have been later used by subsequent researchers in the context of laypeople's attitudes regarding the justification of punishment. In this context, studies have found that people are usually most sensitive to factors associated with retributivism rather than utilitarianism. For example, Darley, Carlsmith, and Robinson[154] sought to determine what motivates a person's desire to punish intentional wrongdoers. In that study, participants were presented with specific vignettes that described individuals committing crimes with varying levels of severity and varying criminal histories associated with the perpetrators themselves.[155] Participants were then asked to recommend a sentence for each of the vignettes.[156] The results of the study indicated that participants were most sensitive to information relating to the severity of the offense when compared to any other information about the perpetrators.[157] With the assumption that severity of the offense was a retributive-based piece of information, the study showed that individuals based their decisions more on retributive motives than motives designed to incapacitate.[158]

In a follow-up study, the findings of Darley, Carlsmith, and Robinson were reaffirmed when researchers manipulated not only the severity of the crime but also the underlying justification for the perpetrator's actions (*i.e.*, stealing funds to pay off illegal gambling debts vs. to benefit underpaid factory workers overseas).[159] Once again, Carlsmith, Darley, and Robinson found that participants typically ignored deterrence factors and instead focused on retributive factors, suggesting that punishment decisions are in some way driven by moral outrage.[160] This method of varying motivations and magnitudes of harm was a key development in the body of literature devoted to studying the motivations behind punishment decisions.

Further studies by Carlsmith incorporated confidence ratings into the determination of sentences assigned to perpetrators.[161] In keeping with the same systematic manipulations of information related to specific aspects of an infraction, participants were asked to rate the relevance of certain information related to retribution, deterrence, or incapacitation.[162] Acquisition of information in this

[152] *Id.* at 154.

[153] *Id.*

[154] *See generally* John M. Darley, Kevin M. Carlsmith & Paul H. Robinson, *Incapacitation and Just Deserts as Motives for Punishment*, 24 LAW & HUM. BEHAV. 659 (2000).

[155] *Id.* at 659.

[156] *Id.* at 662.

[157] *Id.* at 667–68.

[158] *Id.*

[159] Kevin M. Carlsmith, John M. Darley & Paul H. Robinson, *Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment*, 83 J. PERSONALITY & SOC. PSYCH. 284, 289 (2002).

[160] *Id.* at 290.

[161] Carlsmith, *supra* note 144, at 444.

[162] *Id.*

study, however, had no cost associated with it, which is yet another variable that can be manipulated in BP studies. Further, the participant's confidence in their sentencing decisions was recorded via a self-reporting scheme, wherein they were asked to rank their confidence in their decisions on a numerical scale both before and shortly after receiving each type of information.[163] This study ultimately confirmed that information related to retribution justification is most relevant to people tasked with sentencing perpetrators.[164]

As part of these BP studies, specific information types have been linked to particular theoretical justifications for punishment. Keller, Oswald, Stucki & Gollwitzer outlined many of these relationships for the purposes of tracking informational items to punishment decisions in present and future studies.[165] For example, retributive motivations were tied with information items relating to the magnitude of harm, motivation, and intent behind the harm.[166] Preventative motivations were tied to information items relating to the frequency with which the crime was committed, the publicity associated with the crime, and the future detection rate of similar crimes.[167] Incapacitation motivations were tied to specific attributes about the perpetrator themselves, such as the likelihood of violence during the crime, whether the offender was a repeat offender, or whether the offender could be shown to responsibly control his/her impulses.[168] Additionally, by showing the participants all the information they requested in making their sentencing decision at once, rather than sequentially, the study intentionally targeted people's punishment motivations rather than how they formed their punishment decisions.[169] Eventually, this study confirmed results from prior researchers, finding that retributive-related information played the biggest role in an individual's punishment motivations.[170] While a rich body of scholarship indeed supports such a conclusion, one can also find studies that challenge this premise. For example, Slobogin and Brinkley-Rubinstein conducted a set of studies that revealed, among other things, willingness to move from desert-based justifications under certain circumstances (such as crime seriousness).[171]

While these sets of experimental studies offer powerful insights into understanding laypeople's attitudes towards punishment, their main limitations remain: they are, at the end of the day, a set of lab experiments that mimic responses to potentially real-life events. Further, many of these studies capture post-hoc justifications as a response to scientific interventions. As such,

[163] *Id.*

[164] *Id.* at 444–45.

[165] *See generally* Livia B. Keller, Margit E. Oswald, Ingrid Stucki & Mario Gollwitzer, *A Closer Look at an Eye for an Eye: Laypersons' Punishment Decisions Are Primarily Driven by Retributive Motives*, 23 SOC. JUST. RSCH. 99 (2010).

[166] *Id.* at 103 tbl.1.

[167] *Id.*

[168] *Id.*

[169] *Id.* at 102.

[170] *Id.* at 110.

[171] Slobogin & Brinkley-Rubinstein, *supra* note 132, at 118.

participants in these experiments do not respond under natural, real-life settings to these issues. This study aims to address some of these concerns by investigating social discourse with respect to theories of punishment in a more realistic setting: social media. Some recent studies have attempted to elucidate certain aspects of the relationship between punishment motivations and social media, albeit in more indirect ways than our study. Blackwell, Chen, Schoenebeck & Lampe ventured to examine when individuals perceive online harassment as justified and when it is not.[172] That study recruited users through Twitter profiles and provided different control groups with different situations wherein users were harassed for either no conduct whatsoever or increasingly more severe conduct against an elderly couple.[173] Participants were asked to then rank how appropriate they believed the online harassment of the perpetrator was.[174] The study found a strong correlation between individuals who believe in retributivist forms of punishment and those who believe that harassment of an online user who violated community standards was justified.[175] There was, however, some variability in the justification responses depending on the context of the harassment.[176]

A separate study aimed to examine the primary emotions involved when individuals participate in third-party punishment ("TPP").[177] There, Ginther, Hartsough, & Marois determined that most individuals are primarily motivated by a sense of moral outrage rather than any other emotion when making decisions involved in TPP.[178] The participants in that study were assigned to view a particular test scenario.[179] Their emotional responses to that scenario were subsequently recorded, and the participants were asked to make a punishment determination based on a numerical scale.[180] While the results of the study were still somewhat subject to self-reporting and thus may suffer from the limitations discussed above, the experiment did reveal a strong correlation between moral outrage and punishment response to the scenario.[181] This finding confirmed the retributive basis for punishment decisions, as seen in previous experiments.

With regard to the results of the studies, as mentioned, most researchers report participants making punishment decisions based on retributivism rather than any other theoretical justification (this could be based on the information

[172] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck & Cliff Lampe, *When Online Harassment Is Perceived as Justified*, 12 PROC. INT'L. AAAI CONF. ON WEB & SOC. MEDIA 22, 22 (2018).

[173] *Id.* at 25.

[174] *Id.*

[175] *Id.* at 27.

[176] *Id.* at 26.

[177] Matthew R. Ginther, Lauren E. S. Hartsough & René Marois, *Moral Outrage Drives the Interaction of Harm and Culpable Intent in Third-party Punishment Decisions*, 22 EMOTION 795, 795 (2022).

[178] *Id.*

[179] *Id.* at 797.

[180] *Id.*

[181] *Id.* at 798–99. For additional limitations, see *id.* at 799–802.

chosen, the stated goals, and the justifications provided).[182] Indeed, a small number of studies found that individuals could have utilitarian motives and goals, but even some of these findings either relied on self-reporting or were heavily scrutinized by their own authors. Further, on an emotional level, it appears that an individual's feelings towards harassment or condemnation as punishment for prior wrongdoings are driven primarily by moral outrage of the harms caused, recognized as an intuitive response that is often "driven by just deserts" concerns.[183]

Despite the vast usage of experimental design in this setting, an additional domain of potential relevance to the understanding of laypeople's attitudes regarding punishment has largely remained untouched: social media discourse. As such, a gap in the current body of research exists for studies to examine motivations and reactions to punishment decisions through the lens of a social media platform, like Facebook or Twitter. Furthermore, current developments in ML methodologies offer new potential research designs to investigate such questions. These exciting, albeit underutilized opportunities, raise important methodological and normative questions about the potential promise of social media in the empirical desert context. *Could* conversations about crime and punishment serve as a basis to assess communal views about punishment? *Should* they? Moreover, what exactly are the potential methodological and substantial limitations of assessing the community's view of criminal justice through social media platforms?

IV. SOCIAL MEDIA AS A TOOL TO MEASURE SOCIETY

As discussed earlier, thus far, the empirical scholarship assessing lay attitudes about the criminal legal system has largely focused on experiments.[184] Social media has not been broadly utilized by legal scholars to address similar questions. Other disciplines, however, particularly social scientists, offered a more robust utilization of social media texts to answer different questions about political and social phenomena.[185] Scholars have addressed the methodological and

[182] For example, Carlsmith showed that people seek information related to retribution sooner and more frequently than they do utilitarian information. Carlsmith, *supra* note 144, at 437, 444–45. For a summary of findings, see also Carlsmith & Darley, *supra* note 30, at 233–34.

[183] Carlsmith & Darley, *supra* note 30, at 233.

[184] *See* discussion *supra* Part III.

[185] David Lazer et al., *Meaningful Measures of Human Society in the Twenty-first Century*, 595 NATURE 189, 191 (2021) (claiming that "thousands of papers based on Twitter data" were written in recent years). *See, e.g.*, Erik Tjong Kim Sang & Johan Bos, *Predicting the 2011 Dutch Senate Election Results with Twitter*, 13 PROC. CONF. EUR. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS 53, 53 (2012) (using Twitter and Facebook data to forecast elections); Juliet E. Carlisle and Robert C. Patton, *Is Social Media Changing How We Understand Political Engagement? An Analysis of Facebook and the 2008 Presidential Election*, 66 POL. RSCH. Q. 883, 883 (2013) (using it to study political mobilization); David Garcia & Bernard Rimé, *Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack*, 30 PSYCH. SCI. 617, 617 (2019) (using it to learn about collective emotions after terrorist attacks); Johan Bollen, Huina Mao & Xiao-Jun Zeng, *Twitter Mood Predicts the Stock Market*, 2 J. COMPUTATIONAL SCI. 1, 1 (2010) (using it to predict stock market values).

substantive advantages of studying social media. Substantively, it was recognized that social media has become an arena in which social and political identities and positions are debated and discussed.[186] As such, social media can contribute to our understanding of collective positions on different issues, both in online and offline communities.[187] Importantly, social media can also offer access to individual voices traditionally marginalized from mainstream discourse, including traditional media, the academy, and more.[188] Thus, if the goal is to deepen our understanding of lay attitudes in the broadest sense, with an eye toward democratization and communal participation, studying social media can be an important tool.

There are also numerous methodological advantages of studying social media, including the accessibility, scope, and costs of the data.[189] Furthermore, at least in principle, analyzing social media data can overcome some of the concerns raised by social scientists studying public perceptions, specifically the concerns that surveys, and even experiments, are able to capture mostly post-hoc rationalization and not real, intuitive views of the public. More broadly, while well-designed experiments inherently involve scientific interventions, social media data might be closer to what one may define as observational data and, as such, can offer a better, more accurate reflection of real societal views about certain issues, including regarding punishment.

Scholars have also argued, however, that inherent characteristics of social media data make it challenging, not to say impossible, to offer generalizable conclusions about society and human behavior.[190] In the context of Twitter, for example, some argue that "the large majority of Twitter research is making inferences about accounts or tweets" and that "very little of Twitter research can reasonably claim to be making statements about the behaviors of humans."[191] Scholarship has identified a number of core generalizability challenges to the ability to learn about society from social media data.

First, issues of demographic representations are often raised in this context. That is, while it is tempting to use the data available through different social media platforms to draw inferences about the general population, it is often a problematic leap. Thinking about this issue through the concept of sampling, one

[186] Sarah Jackson, Bailey Moya & Brooke Foucault Welles, *#GirlsLikeUs: Trans Advocacy and Community Building Online*, 20 NEW MEDIA & SOC'Y 1868 (2018).

[187] Jeffrey L. Blevins, James Jaehoon Lee, Erin E. McCabe, & Ezra Edgerton, *Tweeting for Social Justice in# Ferguson: Affective Discourse in Twitter Hashtags*, 21 NEW MEDIA & SOC'Y 1636 (2019); Rob Eschmann, Julian Thompson & Noor Toraif, *Tweeting Toward Transformation: Prison Abolition and Criminal Justice Reform in 140 Characters*, 93 SOCIO. INQUIRY 496 (2023).

[188] ANDRÉ BROCK, JR., DISTRIBUTED BLACKNESS: AFRICAN AMERICAN CYBERCULTURES (2020).

[189] Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Poppov & David Stillwell, *Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines*, 70 AM. PSYCH. 543, 543 (2015). It should be noted, however, that these advantages are contingent on the platforms' varying approaches to research, and these can rapidly change as we recently witnessed in the case of Twitter. Recent changes to their policies has also affected this study.

[190] Lazer et al., *supra* note 185, at 192–93.

[191] *Id.* at 191.

can hardly argue that using social media data is a form of random sampling drawn from a representative sample of the general population.[192] Indeed, social media users are often not representative of the population; thus, as is often the case with nonprobability samples, drawing broader social meanings beyond the data itself is likely problematic.[193] For example, studies have already identified that Facebook and Twitter users are often younger and more educated than the general population,[194] Twitter users are found more in wealthier and younger urban areas,[195] and specific topics of conversation are not equally distributed among the population (*e.g.*, politically active Twitter users tend to be more male, urban, and extreme).[196]

These issues of representation are not only challenging statistically but also have an equally concerning qualitative angle. The idea of social media—and the Internet more broadly—as the new, "modern public square"[197] was reflective of similar ideas of representativeness. Those supportive of this pluralistic and democratic ideal have characterized social media as an inclusive digital space that allows everyone, regardless of gender, race, class, and ability, to participate in the Habermasian marketplace of ideas.[198] As such, ideas of representation on social media seem compelling and advance support for the potential contribution of social media data to the understanding of society as a whole. Scholars like Mary Anne Franks, however, believe the idea that social media is an inclusive platform that reflects and promotes democratic participation is a fallacy.[199] Instead, she argues, social media—as any other traditional American public

[192] Jonathan Mellon & Christopher Prosser, *Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users*, 2017 RSCH. & POL. 1, 1 (2017). Recall that when one is sampling data from social media, the sampling happens "at the level of who is a user of the system from which the data are collected as well as who is most active on said system." Lazer et al., *supra* note 185, at 192. As such, this form of sampling is not a sample of the whole population but "at best" 'convenience census' of the platform under investigation. *Id.*

[193] Mellon & Prosser, *supra* note 192, at 1. This raises selection bias concerns, that is, running a "risk of error if there are non-ignorable confounding relationships between the probability of self-selection into samples and outcome variables of interest." *Id.*

[194] Maeve Duggan, *Mobile Messaging and Social Media*, PEW RSCH. CTR. (Aug. 19, 2015), https://www.pewresearch.org/internet/2015/08/19/mobile-messaging-and-social-media-2015/ [https://perma.cc/N5GN-EHUL].

[195] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela & J. Niels Rosenquist, *Understanding the Demographics of Twitter Users*, 5 PROC. INT'L AAAI CONF. ON WEBLOGS & SOC. MEDIA 554, 555 (2011); Momin Malik, Hemank Lamba, Constantine Nakos & Jürgen Pfeffer, *Population Bias in Geotagged Tweets*, 9 PROC. INT'L AAAI CONF. ON WEBLOGS & SOC. MEDIA 18, 18 (2015).

[196] Pablo Barberá & Gonzalo Rivero, *Understanding the Political Representativeness of Twitter Users*, 33 SOC. SCI. COMPUT. REV. 712, 712, 720 (2014). Generally, Twitter is used by only about 20% of the U.S. population and is even less popular in most other countries. *See id.* at 720.

[197] Packingham v. North Carolina, 582 U.S. 98, 107 (2017).

[198] Eschmann, Thompson & Toraif, *supra* note 187.

[199] Mary Anne Franks, *Beyond the Public Square: Imagining Digital Democracy*, 131 YALE L.J.F. 427, 428 (2021) ("But if the goal is to promote a space for democratic deliberation and to realize the values underlying the First Amendment, the public-square analogy is both misleading and misguided.").

squares—serves as a site "for the assertion of violent white male supremacy."[200] Therefore, the ideas expressed on social media platforms are not only not representative of society as a whole, but they are also biased in their inability to truly offer sufficient opportunities for minority communities' views.[201] In the context of empirical desert, when these same communities are often those most affected by the criminal legal system, this gap in representation seems challenging for those hoping to design criminal law and policy based on the narratives extracted from social media.[202]

Moreover, differences in the platforms themselves might affect the published content as people might behave differently depending on the platform (*e.g.*, the same person might respond differently on Facebook versus Twitter).[203] This reflects unique generalizability concerns that focus on technology itself since the ability to draw inferences from the observed behavior depends not only on the demographics of the population but also on the particular "observational context."[204] Furthermore, given the nature of these platforms, the "processes that underlie our online social actions, relationships and structures" are extremely dynamic and change rapidly.[205] This requires adopting a dynamic system to follow such structural changes to embed them within our generalizable framework.

A *second* generalizability concern relates to the potential to influence outcomes on social media platforms either through algorithmic design (*i.e.*, "algorithmic confounding") or other manipulations such as bots. As a result, it might be difficult to distinguish between information that is a product of "typical human behavior" and information that is a product of a nonhuman intervention or the platform's rules implemented through its algorithmic design.[206] As for the latter, Lazer, Hargittai, Freelon, Gonzalez-Bailon et al. summarize the problem succinctly: "[w]ithout knowing how a system is designed, we could easily attribute social motives to behavior driven by algorithmic decisions."[207] In fact, manipulation of human behavior is at the core of what platform creators hope to achieve and for different purposes, from increasing engagement on the platform to advancing the sale of different products. This poses a meaningful challenge in

[200] *Id. See also* Danielle Citron & Mary Anne Franks, *Cyber Civil Rights in the Time of COVID-19*, HARV. L. REV. BLOG (May 14, 2020), https://blog.harvardlawreview.org/cyber-civil-rights-in-thetime-of-covid-19 [https://perma.cc/38UG-YBY4]; Azmina Dhrodia, *Unsocial Media: A Toxic Place for Women*, 24 IPPR PROGRESSIVE REV. 381, 381 (2018); Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 105 (2009).

[201] According to Franks, online spaces "merely replicate existing hierarchies and reinforce radically unequal distributions of social, economic, cultural, and political power." *Id.* at 429.

[202] In this context, see Robinson's definition of the relevant communities: "the relevant community is that which will be bound by the rule being enacted." Robinson, *supra* note 7, at 1573.

[203] Lazer et al., *supra* note 185, at 193.

[204] *Id.*

[205] *Id.* at 192.

[206] *Id.*; Claudia Wagner et al., *Measuring Algorithmically Infused Societies*, 595 NATURE 197, 197 (2021).

[207] Lazer et al., *supra* note 185, at 192; *see also, e.g.*, David Lazer, Ryan Kennedy, Gary King & Alessandro Vespignani, *The Parable of Google Flu: Traps in Big Data Analysis*, 343 SCI. 1203, 1203 (2014) (illustrating that changes to Google's search algorithms was the main drive behind increasing overprediction of flu prevalence of Google Flu Trends).

assessing the connections between the behavior on the platform and the general behavior, even of the observed users.

Beyond these two core generalizability issues, there are additional issues raised by scholars that should be taken into account when working with social media data, particularly with Twitter. For example, when analyzing Twitter data, there are challenges related to linguistic meaning-making due to computers having challenges decoding more nuanced expressions, irony, or sarcasm.[208] Furthermore, even comparisons to previous studies of social media should be carefully assessed. First, due to the length of the unit of analysis, tweets are much shorter "and contain much less content than, for instance, news articles and traditional blogs,"[209] which raises some questions as to their informational value.[210] Second, much of the information delivered through Twitter is not merely in the written text, as 19% of all messages include links to other websites.[211] To paraphrase Tumasjan, Sprenger, Sandner & Welpe, it thus remains contested whether 140-character messages provide information sufficient to transform knowledge about human behavior.[212]

But despite the above-mentioned concerns, we also find solid and consistent evidence that with sufficient methodological remedies,[213] these concerns can be mitigated. That is, Twitter analysis can, in fact, reflect more about the general population than critics suggest. For example, studies find that data gathered from posts on Twitter are associated with public reaction,[214] effectively identify topics of public importance,[215] can predict election results,[216] reflect historic urban-landscape values,[217] and can be correlated with emotional

[208] Lazer et al., *supra* note 185, at 192. According to Lazer et al., the severity of this challenge depends "on the structure of the noise and, again, on what matters—that is, the research question." *Id.*

[209] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner & Isabell M. Welpe, *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*, 4 PROC. INT'L AAAI CONF. ON WEBLOGS & SOC. MEDIA 178, 179 (2010).

[210] *Twitter Study: Usage- 40% is Pointless Babble*, PEAR ANALYTICS, https://pearanalytics.com/twitter-study-reveals-interesting-results-40-percent-pointless-babble/ (last visited Aug. 21, 2023) [https://perma.cc/VAQ9-YCPP].

[211] Dan Zarrella, *Announcing the June 2009 State of the Twittersphere Report*, HUBSPOT (Oct. 20, 2016), https://blog.hubspot.com/blog/tabid/6307/bid/4829/announcing-the-june-2009-state-of-the-twittersphere-report.aspx [https://perma.cc/3B8W-A228].

[212] Tumasjan et al., *supra* note 209, at 179.

[213] *See, e.g.*, Mellon & Prosser, *supra* note 192, at 8; Nicholas Beauchamp, *Predicting and Interpolating State-Level Polls Using Twitter Textual Data*, 61 AM. J. POL. SCI. 490, 502 (2017).

[214] *See generally* Nicholas A. Diakopoulos & David A. Shamma, *Characterizing Debate Performance Via Aggregated Twitter Sentiment*, 28 PROC. SIG CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1198 (2010).

[215] *Id.*

[216] *See* Beauchamp, *supra* note 213, at 502; *see generally* Tumasjan et al., *supra* note 209; Adam Bermingham & Alan F. Smeaton, *On Using Twitter to Monitor Political Sentiment and Predict Election Results*, *in* PROC. WORKSHOP ON SENTIMENT ANALYSIS WHERE AI MEETS PSYCH. 2 (2011). Bermingham and Smeaton even argue that the predictive power of Twitter even comes close to traditional election polls. *See id.* at 6–9.

[217] Manar Ginzarly, Ana Pereira Roders & Jacques Teller, *Mapping Historic Urban Landscape Values Through Social Media*, 36 J. CULTURAL HERITAGE 1, 9 (2019).

experiences in society at large.[218] All these studies, and others, highlight the potential promise of Twitter in identifying general views and perspectives among the general population. There is thus similar potential in using social media to examine lay perspectives regarding the distribution of criminal liability and punishment. Moreover, as discussed earlier, views appearing on social media might affect stakeholders, including policy-makers, judges, and prosecutors. As a result, there is value in studying these views, even if they do not necessarily represent the views among the general population in a strict statistical form.

With this in mind, the Article next describes its exploratory empirical component: using NLP methodologies to explore first, what narratives about criminal liability and punishment dominate social media discourse and, second, whether these narratives can and should serve as the basis to assess "criminal liability and punishment rules derived from the governed community's principles of justice,"[219] also known as principles of "empirical desert."

V. RESEARCH DESIGN

The study adopts a text-as-data approach to social media discourse regarding criminal punishment. We investigate, through two different methodologies, the narratives that laypeople advance when communicating about crime and punishment in the virtual realm. To do so, we first leverage recent advances in NLP that make it possible to discover hidden thematic structure in large collections of documents. Second, we utilize current advancements in NLP for performing text classification through text generation using an autoregressive language model named GPT-3.5, which is a sub class of GPT-3 Models created by OpenAI in 2022.

The first methodology we rely on is a machine learning ("ML") methodology known as Topic Modeling ("TM"). TM is an exploratory technique, useful for imposing order upon large bodies of textual data and for discovering information that helps analysts see beyond their priors.[220] TM algorithms are a suite of ML methods for discovering hidden thematic structure in large collections of documents. As such, TM is a method of large-scale text analysis that represents each document in a collection as a member of one and only one of several more general "topics" or "themes" appearing in a collection.[221] Employing TM allows expansion beyond what humans often consider "close reading" methodology (which is limited, by its nature, to a small number of texts) and focuses instead on "distant reading" that has the ability to analyze "large corpora of text."[222]

[218] David Garcia, Max Pellert, Jana Lasser & Hannah Metzler, *Social Media Emotion Macroscopes Reflect Emotional Experiences in Society at Large*, ARXIV (July 28, 2021), https://arxiv.org/abs/2107.13236 [https://perma.cc/LVH7-AD8X].

[219] Robinson, Barton & Lister, *supra* note 7, at 313.

[220] David M. Blei, *Probabilistic Topic Models*, 55 COMMC'NS ACM 77, 84 (2012).

[221] *Id.* at 82.

[222] *See* Renana Keydar, *Listening from Afar: An Algorithmic Analysis of Testimonies from the International Criminal Courts*, 2020 ILL. J. L., TECH. & POL'Y 55, 60 (2020).

This approach builds on the distributional hypothesis of linguistic theory, which suggests that the meaning of a given word can be derived from words that occur around it.[223] With a collection of documents as input, a topic model can produce a set of interpretable ‘‘topics’’ (*i.e.*, groups of words that are associated under a single theme) and assess the strength with which each document exhibits those topics.[224]

Using NLP and ML algorithms that detect semantic structure patterns, large bodies of text units can be classified into semantically similar clusters without human direction. The researcher will later assign a label to the clusters based on the key words and the documents identified as the core of a semantic cluster.[225] As such, the approach is unsupervised. That is, the researcher typically selects the number of topics to be estimated, and the algorithm identifies the topics inductively from word co-occurrence patterns in the documents under analysis.[226] Unlike qualitative analysis based on information retrieval, where researchers know what they are looking for, topic modeling is attractive because it offers a formalism for exposing a corpus's themes by discovering groups of words that often appear together in documents ("topics"). To offer a deeper dive into the data and to further understand the topic distribution and the content related to each topic, we contextualized the TM analysis by qualitatively analyzing tweets that were chosen by the algorithm to be representative of each of the topics.[227]

The second ML methodology we rely on is known as text classification. Text classification is a methodology that assigns a set of *predefined* categories to an open-ended text.[228] For this study, we use GPT 3.5 as our classifier. GPT 3.5 is a language model that was trained to perform the task of next word prediction, *i.e.*, given a sentence, the model will output the most probable word to follow with respect to the word distribution it has learned from billions of texts it was trained on. To utilize GPT 3.5 as a text classifier, we provide the model with a prompt that specifies the classification task and includes a description of the labels we want to classify the text into. In this Article, we provided the model with four recognized justifications or alternatives for punishment based on the existing scholarship discussed in Section II of this Article (retributive, utilitarian, expressive, and restorative) and requested the model to classify each tweet based on the definitions provided in the prompt. This approach, known as unsupervised

[223] *See generally* Tomas Mikolov, Kai Chen, Greg Corrado, & Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, ARXIV (2013), https://doi.org/10.48550/arXiv.1301.3781 [https://perma.cc/987J-G5LH]; Emma Rodman, *A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors*, 28 POL. ANALYSIS 87 (2020).

[224] Blei, *supra* note 220, at 78.

[225] J.B. Ruhl, John Nay & Jonathan Gilligan, *Topic Modeling the President: Conventional and Computational Methods*, 86 GEO. WASH. L. REV. 1243, 1248–49 (2018).

[226] *See* Keydar, *supra* note 222, at 69–70.

[227] For a methodologically related approach, see Renana Keydar, Yael Litmanovitz, Badi Hasisi, & Yoav Kan-Tor, *Modeling Repressive Policing: Computational Analysis of Protocols from the Israeli State Commission of Inquiry into the October 2000 Events*, 47 L. & SOC. INQ. 1075 (2022).

[228] Kwangil Park, June Seok Hong, & Wooju Kim, *A Methodology Combining Cosine Similarity with Classifier for Text Classification*, 34 APPLIED ARTIFICIAL INTELLIGENCE 396 (2020).

zero-shot learning, allows us to utilize a model that was not explicitly trained on the classification task and does not require labeled examples for training. Instead, we rely on the model's understanding of language and the given task definition to classify the text into the desired labels. By employing GPT-3.5 in this manner, we can perform classification tasks without the need for extensive supervised training or labeled datasets.

The study's "collection" consisted of Tweets posted in proximity to legal events pertaining to four homicide cases. As such, the "tweets" are treated as the unit of analysis. In particular, we investigate the social media discourse around the trials of Aaron Hernandez and Casey Anthony and the sentencing decisions in Kimberly Potter and Nikolas Cruz's cases. For each of these cases, we analyzed the tweets three days before and after the decisions (*i.e.*, verdicts/sentencing) were rendered (April 15, 2015; July 5, 2011; February 18, 2022; and October 13, 2022, respectively). While the TM methodology allows us to understand dominant themes related to culpability and punishment in the context of these cases, the text classification methodology directly classifies the tweets into different categories, that is, different justifications for punishment.

Given the exploratory nature of this study, we chose cases in different procedural stages: two in the trial stage and two in the sentencing stage. We started the analysis with the trial stage cases and then, based on our preliminary findings, moved to the sentencing stage cases. We chose to analyze the particular cases because they share similar traits: all cases were State prosecuted, received significant media attention, and revolved around similar offenses (*i.e.*, homicide). But the cases also have some differences, particularly with regard to the suspects' characteristics and the outcome; for example, Hernandez was found guilty of the homicide, and Anthony was found not guilty. While the focus of Hernandez's and Anthony's cases were naturally questions of guilt or innocence, the issue of punishment was at the core of Potter and Cruz's cases. As such, and in the context of analyzing attitudes regarding punishment, the comparison between the group of cases offered an opportunity to better understand social media conversations that directly revolve around sentencing as opposed to more general (and more common) conversations about criminal cases. We found these cases to be a good starting point for the exploratory analysis given that they are all homicide cases and received meaningful media attention, which increases the likelihood of spontaneous and emotional social media responses and thus offers a richer and potentially nuanced corpus for analysis. At the same time, similar rationales also expose the limitations associated with choosing these cases, as they do not represent the bulk of the traditional, day-to-day criminal cases.

We narrowed the scope of the analysis to tweets that include some form of *sentiment* based on sentiment-analysis software. [229] By doing so, we aspired to

[229] C.J. Hutto, & Eric Gilbert, *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, EIGHTH INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM-14) (2014).

drop news reporting and to capture only those texts that embed a form of a user's response (emotional or other).

To decide on the preferred number of topics for the TM methodology, we ran the algorithm for three, five, seven, ten, and twelve topics and visualized the results using pyLDAvis, a web-based interactive tool to visualize topics under the TM model.[230] Under pyLDAvis, good topic modeling includes big and nonoverlapping topics. Based on this criterion, we ended up choosing five topics for our final model.[231]

As mentioned above, for the text classification portion of the analysis, we used GPT 3.5 and offered four definitions of theories of punishment based on existing scholarship.[232] Given the novelty of using GPT 3.5 for the purposes of text classification, we validated the results by randomly selecting several tweets and manually classifying them by several individuals: three research assistants and one of the authors.

Based on the existing experimental findings related to theories of punishment, which have revealed a strong consensus that retributive motivations underly many people's decisions about whether and how much to punish, we hypothesized that reactions on Twitter to the cases—particularly in direct relation to punishment decisions—would mostly reflect retributivist intuitions. As we will discuss below, while we find some support for this hypothesis, our findings are much more nuanced.

A. Summary of the Cases

Before delving into the findings, we will offer a brief summary of the criminal cases that are the focus of this study.

[230] For a discussion of the LDAvis tool, see Carson Sievert & Kenneth E. Shirley, *LDAvis: A Method for Visualizing and Interpreting Topics*, *in* PROCEEDINGS OF THE WORKSHOP ON INTERACTIVE LANGUAGE LEARNING, VISUALIZATION, AND INTERFACES 63 (2014), https://aclanthology.org/W14-3110/ [https://perma.cc/A4A6-QYGL].

[231] The tweets were scrapped using the Snscrape package on Python. The LDA was the algorithm used for TM, which was fitted using the Gensim package on Python.

[232] These are the definitions used in the analysis:

Retribution: Imperative to punish derived from the goal of giving offenders what they deserve. Punishment appropriate when the severity of punishment is proportionate to the magnitude of harm and to the offender's criminal intent

Utilitarianism: Imperative to punish derived from the future consequences of the punishment: weigh the potential harm to offender against the benefits to society. Benefits: either deter offender, deter society at large, rehabilitate offender or incapacitate offender

Expressive: Punishment serves to define and reinforce important social norms of law-abiding behavior and relative crime seriousness

Restorative: A collaborative social process whereby parties with a stake in a specific offense collectively resolve how to deal with the aftermath of the offense

1. *Casey Anthony*

Casey Maria Anthony was charged in Florida with first-degree murder of her child, Caylee Marie Anthony, after Caylee's grandmother called 9-1-1 to report her granddaughter's absence.[233] Casey pled not guilty.[234] The trial lasted for six weeks between May and July 2011, and the prosecution sought the death penalty, claiming that Casey administered chloroform and applied duct tape to her daughter's nose and mouth.[235] The defense claimed that Caylee drowned accidentally in the family's swimming pool and that her grandfather, George, disposed of the body.[236] Casey did not testify.[237] On July 5, 2011, the jury found her not guilty of first-degree murder, aggravated child abuse, and aggravated manslaughter.[238] She was found guilty of four misdemeanor counts for providing false information to a law enforcement officer.[239]

2. *Aaron Hernandez*

Aaron Josef Hernandez was an American football tight end and played in the National Football League ("NFL") for three seasons.[240] His career came to an end in 2013 after his arrest for the murder of Odin Lloyd.[241] Lloyd himself was a semiprofessional who dated the sister of Hernandez's fiancé.[242] Lloyd's body was found with multiple gunshots in a park about a mile from Hernandez's home.[243] In June 2013, Hernandez was charged with first-degree murder for the murder of Lloyd, alongside five gun-related charges.[244] Hernandez pled not guilty.[245] On April 15, 2015, he was found guilty of first-degree murder, which in Massachusetts involved an automatic sentence of life without the possibility of parole.[246] Hernandez was also found guilty of the five firearm-related

[233] Tim Ott, *Casey Anthony: A Complete Timeline of Her Murder Case and Trial*, BIOGRAPHY (Dec. 2, 2020), https://www.biography.com/crime/casey-anthony-muder-trial-timeline-facts [https://perma.cc/N83K-UFE6].

[234] *Casey Anthony Trial Fast Facts*, CNN (June 22, 2022), https://www.cnn.com/2013/11/04/us/casey-anthony-trial-fast-facts/index.html [https://perma.cc/NR3X-VHGU] [hereinafter *Fast Facts*].

[235] Ott, *supra* note 233.

[236] *Id.*

[237] *Fast Facts*, *supra* note 234.

[238] Ott, *supra* note 233.

[239] *Id.*

[240] Colin Bertram, *Aaron Hernandez: Timeline of His Football Career, Murder Trials and Death*, BIOGRAPHY (Jan. 15, 2020), https://www.biography.com/athletes/aaron-hernandez-timeline [https://perma.cc/7MNT-UVGW].

[241] *Id.*

[242] *Id.*

[243] *Id.*

[244] *Id.*

[245] *Id.*

[246] *Id.*

offenses.[247] While Hernandez was on trial, he was also indicted for a 2012 double homicide but was acquitted of these charges after a 2017 trial.[248] After his acquittal from the double homicide, Hernandez was found dead in his cell in what was later determined as suicide.[249]

3. *Kimberly Potter*

On April 11, 2021, during a traffic stop and attempted arrest for an outstanding warrant in Brooklyn Center, Minnesota, Daunte Wright, a 20-year-old Black man, tragically lost his life when police officer Kimberly Potter fatally shot him.[250] Potter claimed that she had intended to deploy her service Taser, but instead, she fired her service pistol. The shooting ignited protests in Brooklyn Center and reignited ongoing demonstrations against police shootings in the Minneapolis–Saint Paul metropolitan area. Days after the incident, Potter resigned from her position. On December 23, 2021, Potter was convicted by the jury of first-degree manslaughter and second-degree manslaughter. On February 18, 2022, she was sentenced to two years in prison, serving sixteen months and eight months of supervised release.[251] On April 24, 2023, Potter was released from prison.[252]

4. *Nikolas Cruz*

On February 14, 2018, Nikolas Cruz, a former student at the Marjory Stoneman Douglas High School in Parkland, Florida, shot and killed seventeen students and staff at the school and injured an additional seventeen.[253] In October 2021, Cruz pleaded guilty to all charges (seventeen charges of first-degree murder and seventeen charges of attempted first-degree murder) in the deadliest high

[247] Ryan Wilson, *Aaron Hernandez Charged with Murder, Five Gun-Related Charges*, CBS SPORTS (June 26, 2013, 10:49 AM), https://www.cbssports.com/nfl/news/aaron-hernandez-charged-with-murder-five-gun-related-charges/ [https://perma.cc/433V-8XPX].

[248] *Aaron Hernandez Acquitted in Double-Murder Trial*, NFL (Apr. 14, 2017, 8:35 AM), https://www.nfl.com/news/aaron-hernandez-acquitted-in-double-murder-trial-0ap3000000800192 [https://perma.cc/8MQ3-Q3K4].

[249] *Aaron Hernandez Found Dead After Hanging in Prison Cell*, ESPN (Apr. 19, 2017), https://www.espn.com/nfl/story/_/id/19191248/former-new-england-patriots-te-aaron-hernandez-found-dead-hanging-prison-cell [https://perma.cc/A6ZE-HVJZ].

[250] *The Killing of Daunte Wright*, MPR NEWS, https://www.mprnews.org/crime-law-and-justice/killing-of-daunte-wright (last visited Aug. 21, 2023) [https://perma.cc/62NP-KGWL].

[251] *Kim Potter Sentenced to 2 Years in Fatal Shooting of Daunte Wright*, N.Y. TIMES (Feb 18, 2022) https://www.nytimes.com/video/us/100000008217715/daunte-wright-kim-potter-sentencing.html [https://perma.cc/6WWE-NR4X].

[252] Adrienne Broaddus, *Former Minnesota Police Officer Kim Potter Released from Prison After Serving Time for Deadly Shooting of Daunte Wright*, CNN (Apr. 24, 2023) https://www.cnn.com/2023/04/24/us/kim-potter-release-prison-daunte-wright/index.html [https://perma.cc/J3FR-7BLZ].

[253] *Id.*

school shooting in the history of the United States.[254] The prosecution sought the death penalty.[255] On October 13, 2022, the jurors unanimously agreed that Cruz was entitled to the death penalty but not about whether it should be imposed.[256] The result was a recommendation for a sentence of life without the possibility of parole.[257] On November 2, 2022, in accordance with the jury's recommendation, Cruz was sentenced to life without parole.[258]

VI. FINDINGS

A. *Topic Modeling*

As a reminder, the TM methodology was deployed for all the tweets that mentioned "Casey Anthony," "Aaron Hernandez," "Kimberly Potter," and "Nikolas Cruz" three days before and three days after their verdicts (Hernandez and Anthony) or sentencing decisions, in Cruz and Potter's cases (the jury's recommendation in Cruz): July 5, 2011; April 15, 2015; February 18, 2022; and October 13, 2022, respectively. The algorithm was asked to analyze five topics.[259]

Overall, in Casey Anthony's case, we analyzed 28,672 tweets with sentiment: 1,729 before the verdict and 26,943 after the verdict. In Aaron Hernandez's case, we analyzed 6,741 tweets with sentiment: seventy-seven before the verdict and 6,664 after the verdict. In Kimberly Potter's case, we analyzed 161 tweets with sentiment: 31 before the sentencing and 130 after. In Nikolas Cruz's case, we analyzed 126 tweets with sentiment: four before the jury's recommendation and 122 after.[260]

Recall that Twitter poses some unique challenges in identifying distinct topics for both the algorithm and the researcher, given the limited textual space it offers and the relative similarity in messaging. With this limitation in mind, a careful reading of the words included in each topic, and their relative frequencies

[254] Elisha Fieldstadt, *Nikolas Cruz Pleads Guilty to 17 Counts of Murder in 2018 Parkland School Shooting*, NBC NEWS (Oct. 20, 2021, 11:04 AM), https://www.nbcnews.com/news/us-news/nikolas-cruz-pleads-guilty-17-counts-murder-2018-parkland-school-n1281961 [https://perma.cc/6DFF-H9RW].

[255] *See id.*

[256] *Non-Unanimous Florida Jury Sentences Nikolas Cruz to Life Without Parole for Parkland School Shootings*, DEATH PENALTY INFO CTR. (Oct. 13, 2022), https://deathpenaltyinfo.org/news/non-unanimous-florida-jury-sentences-nikolas-cruz-to-life-without-parole-for-parkland-school-shootings [https://perma.cc/PEH6-X7MM].

[257] *Id.*

[258] *Id.*

[259] Selecting the number of topics is one of the most problematic modeling choices. As mentioned, we have experimented with a number of options and five seems to offer the best fit that is able to distinguish between topics without too much of an overlap.

[260] It is clearly evident that the universe of analysis in Cruz and Potter's cases is much smaller.

within the topic, all complemented with a close reading of the actual tweets, allowed the identification of five distinct topics for each of the cases.[261]

For each case, we will offer a dynamic online visual that captures the semantic fields identified in each of the topics, will discuss the core meaning given to each topic, and offer a number of illustrative tweets and a visual that reflects the distribution of topics among the overall universe of tweets.

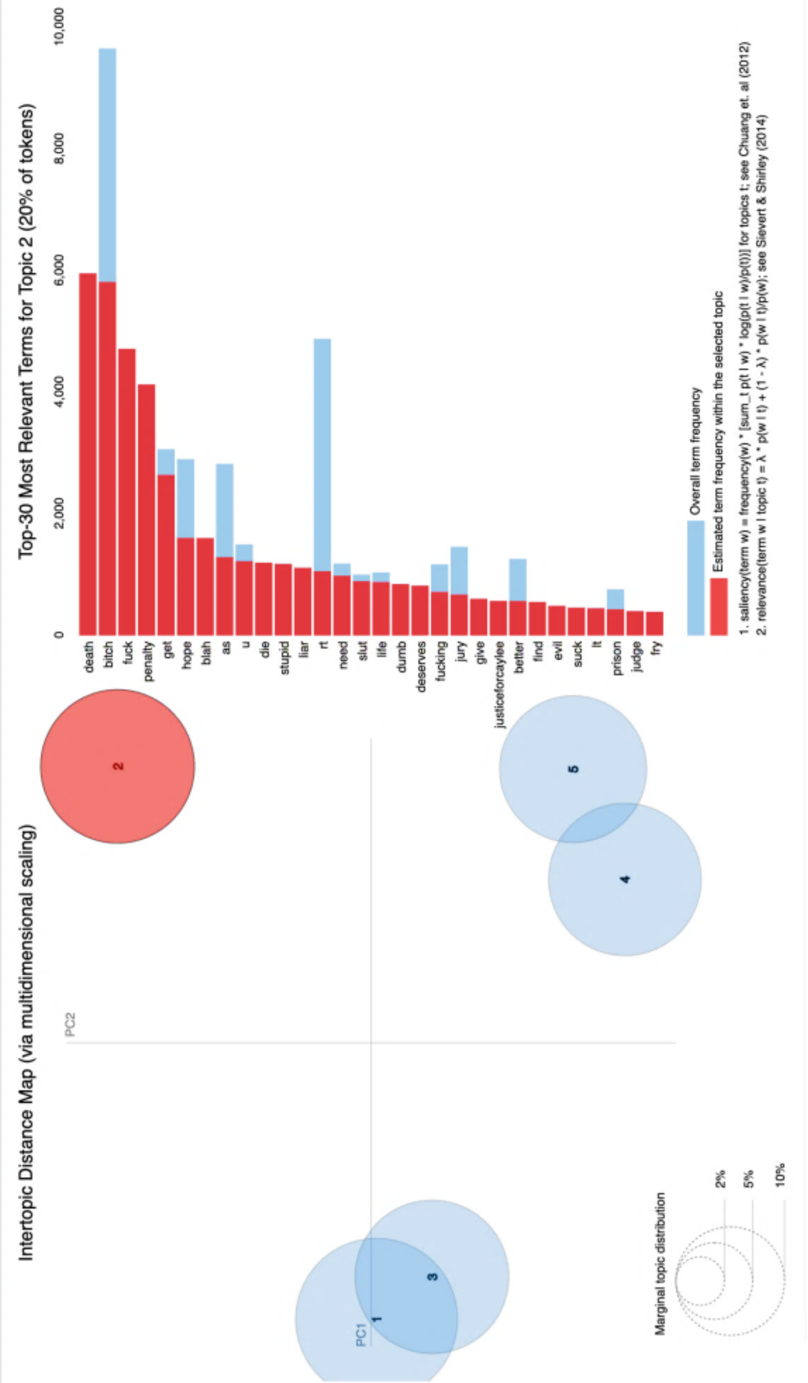
1. Casey Anthony

For illustration, Figure 1 captures the semantic field identified under Topic 2 (5,422 tweets). A link to a dynamic visualization of all topics can be found here.[262]

[261] An additional note of caution: given the nature of social media, some of the words used in the tweets, and thus identified by the algorithm, can be offensive or obscene. Some do represent emotional responses and are in common use in day-to-day language. As such, omitting these words would likely affect the robustness of the analysis. As such, these words will be presented as part of the visualization of the topics identified.

[262] INTERTOPIC DISTANCE MAP, https://rtmdrr.github.io/Renegotiating_Theories_of_Punishment/Anthony.html#topic=0&lambda=1&term= (last visited Aug. 21, 2023) [https://perma.cc/6SVW-R2F6].

FIGURE 1: CASEY ANTHONY MOST SALIENT TERMS, TOPIC 2



Topic 1 includes terms like "guilty," "murder," "wow," "verdict," and "reached." We entitled it "*Guilt/Innocence*." It includes simple statements relating to the defendant's culpability but also a form of legal analysis of different informational pieces that can assess either guilt or innocence.
Illustrative tweets:

(1) " #CaseyAnthony found not guilty of 1st degree murder & aggravated child abuse. Found guilty of lying to a police officer.."

(2) "Emotions can be blinding, in the Casey Anthony trial. . . No DNA", no proof of murder, no proof of murder, no conviction. . ."

(3) "Felony murder 1st degree?? May have trouble getting the premeditation to stick.. thoughts?"

Topic 2 includes terms like "death," "b…," "penalty," and "get," and seems to focus on the "*punishment*" (or lack thereof in Anthony's case). Misogynist terms are the second most frequent terms used within this topic.
Illustrative tweets:

(1) "casey anthony you deserve a death sentence for killing your two year old daughter."

(2) "What an ironic thing to say. RT: @HLNTV: Jose Baez criticizes death penalty: "We need to stop killing our own people." #CaseyAnthony"

Topic 3 includes terms like "guilty," "wtf," "justice," "Caylee," "sad," and "verdict," focusing on *emotional reactions* to the verdict. This topic seems to be victim-focused.
Illustrative tweets:

(1) "Can't believe Casey Anthony was found guilty!!![263] *NO JUSTICE FOR BEAUTIFUL CAYLEE!"*

(2) "Cannot believe #caseyanthony verdict! I am absolutely disgusted and sadden no justice for that poor truly innocent little girl".

Topic 4 includes terms like "killed," "free," "b…," "daughter," "damn," "baby," "hate," and "sick" and also includes emotional responses to the trial, but this time with a focus on Anthony herself either approving or disapproving the decision in her case.
Illustrative tweets:

(1) "I hate Casey Anthony! She is SO guilty! @StavKaragiorgis"

[263] Should have been "not guilty."

(2) *"Free #CaseyAnthony! Let her go! Let justice finally happen"*

Topic 5 includes terms like "closing," "argument," "cry," "b…," and "prosecution," which we will entitle "*descriptive*" as it offers information about the course of events in court during the trial.
For example:

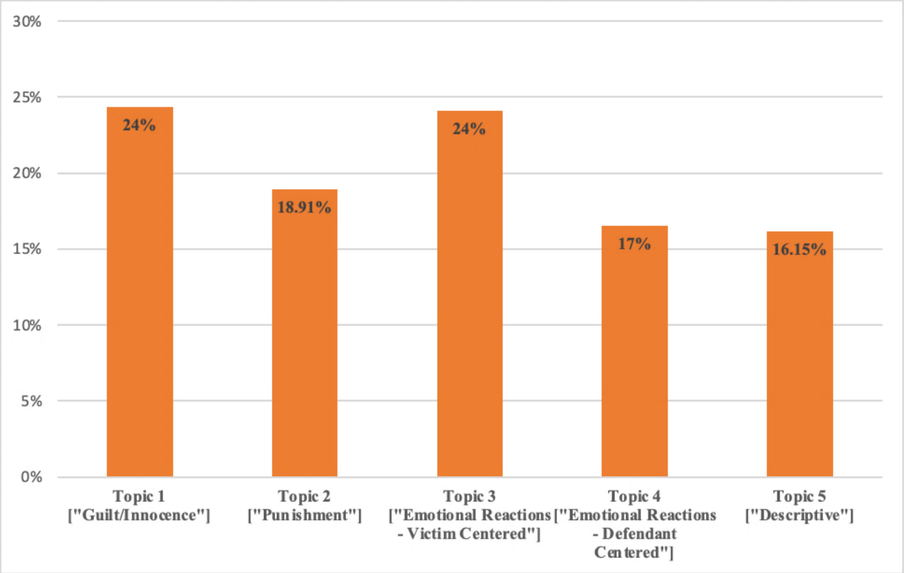
(1) *"Fantastic closing argument from the prosecution. #CaseyAnthony #ClosingArgument"*

(2) *"#caseyanthony broke down crying as prosecutors called her a liar who murdered her child #closing arguments after 33 days and 100 witnesses"*

(3)

Figure 2 below summarizes the distribution of topics in Casey Anthony's case:

FIGURE 2: CASEY ANTHONY TOPIC DISTRIBUTION

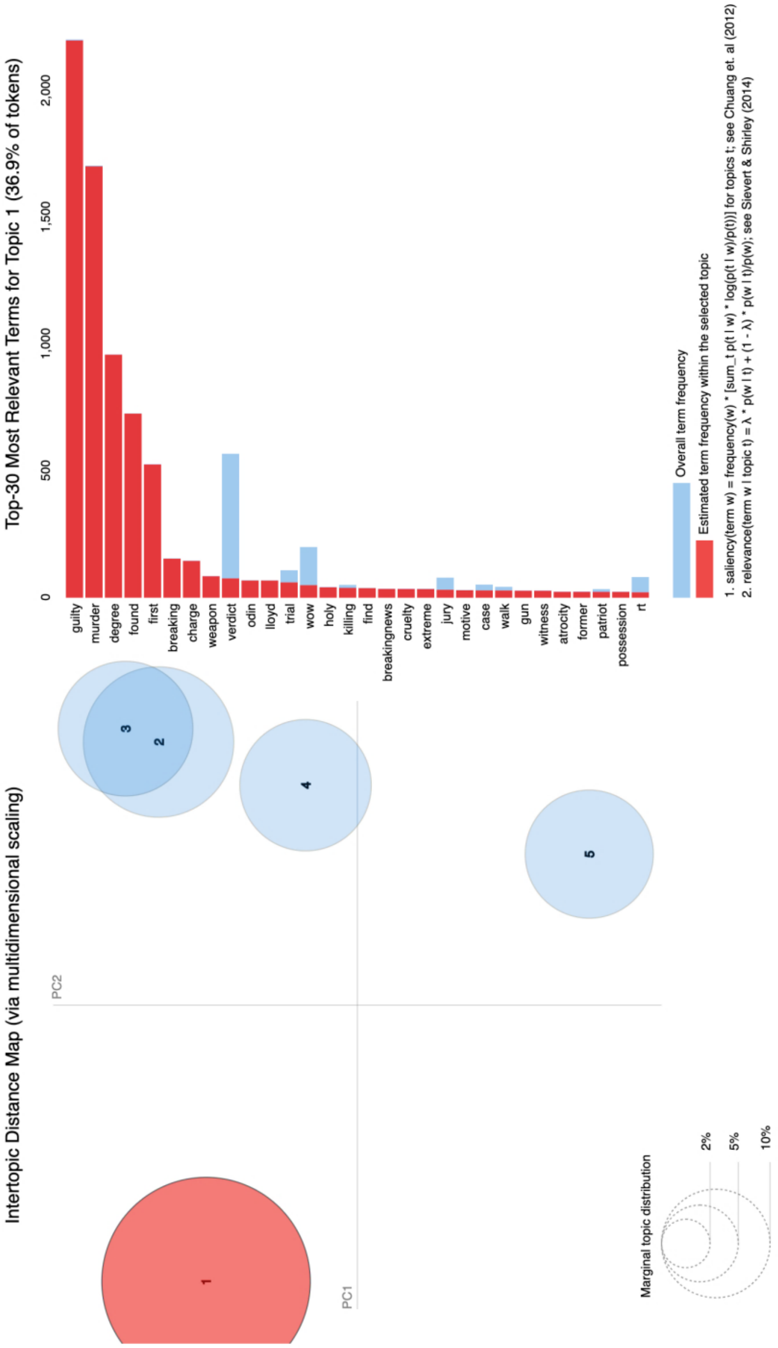


2. *Aaron Hernandez*

For illustration, Figure 3 captures the semantic field identified under Topic 1. A link to a dynamic visualization of all Topics can be found here.[264]

[264] INTERTOPIC DISTANCE MAP, https://rtmdrr.github.io/Renegotiating_Theories_of_Punishment/Hernandez.html#topic=0&lambda=1&term= (last visited Aug. 21, 2023) [https://perma.cc/SXV3-3DFM].

FIGURE 3: AARON HERNANDEZ MOST SALIENT TERMS, TOPIC 1



The left-hand side shows the number of topics and their spread across the semantic universe. The size of each cycle represents the relative size of the topic (*i.e.*, the number of tweets out of the overall tweets). The right-hand side depicts the relevant terms per topic and their frequency within that topic. Topic 1 is the largest topic and includes 2182 tweets.

Topic 1, which we entitled "*Guilt/Innocence*," resembles the Guilt/Innocence topic identified in Casey's case. It includes terms like "guilty," "murder," "degree," and "found" and represents a group of tweets that focus on the question of whether the defendant is guilty or innocent. It includes simple statements relating to the defendant's culpability (*e.g.*, "Hernandez was found guilty"), but also a form of legal analysis of different informational pieces that can assess either guilt or innocence.

Illustrative tweets:

(1) "I'm about 80% sure Aaron Hernandez is going to beat this murder case. No murder weapon, no clear motive. He might walk"

(2) #Aaron Hernandez No witness No motive No confession No weapon. Guilty of first degree murder. OJ: Blood, DNA, gloves, motive. Not guilty…

(3) #Aaron Hernandez case included, no witnesses and no murder weapon. . . But he got first degree murder.

Topic 2, which we entitled "*punishment*," includes terms like "life," "parole," and "sentenced." For the most part, it includes descriptions of the punishment imposed on Hernandez, with some references to its justification, either supporting or rejecting the punishment imposed.

Illustrative tweets:

(1) "Damn Aaron Hernandez Got Life In Prison 😔"

(2) "Aaron Hernandez should get the death penalty, if you kill someone you should be killed as well"

Topic 3, which we entitled "*support*," includes terminology like "free," "damn," "hope," "justice," "n . . . ," and "innocent." It includes tweets either supporting Aaron Hernandez (and thus rejecting the verdict or sentence) or supporting the verdict and the sentence. Importantly, tweets that fall under this topic include the majority of race-related comments in Hernandez's context (either positive or negative).

Illustrative tweets:

(1) "Free my n… @AaronHernandez"

(2) *"no motive, no witnesses, no weapon. Aaron Hernandez found guilty.. Yet a cop choking a man to death on tape gets acquitted.. interesting."*

(3) *"Looks like #AaronHernandez luck finally ran out. Justice was served."*

Topic 4, which we entitled "*Disappointment*," includes terms like "smh,"[265] "damn," "wow," "waste," and "talent." Many tweets that fall under this topic express frustration or disappointment from the case but not from the direct legal issues surrounding it. Instead, these tweets seem more focused on the fact that Aaron Hernandez—as a symbol of a promising life trajectory—was implicated in criminal matters and was supposed to end his life in prison.

Illustrative tweets:

(1) *"Damn Aaron Hernandez.. Guilty!!! What a promising career gone to waste!!*

(2) *"Aaron Hernandez, wasted talent, wasted greatness smh"*

Topic 5, which we entitled "*emotional reactions to verdict*," includes terms like "verdict," "reached," "good," "hell," "great," and "sad." The responses here are more general and do not directly express disappointment related to the potential Hernandez had as an athlete, as seen in Topic 4.

Illustrative tweets:

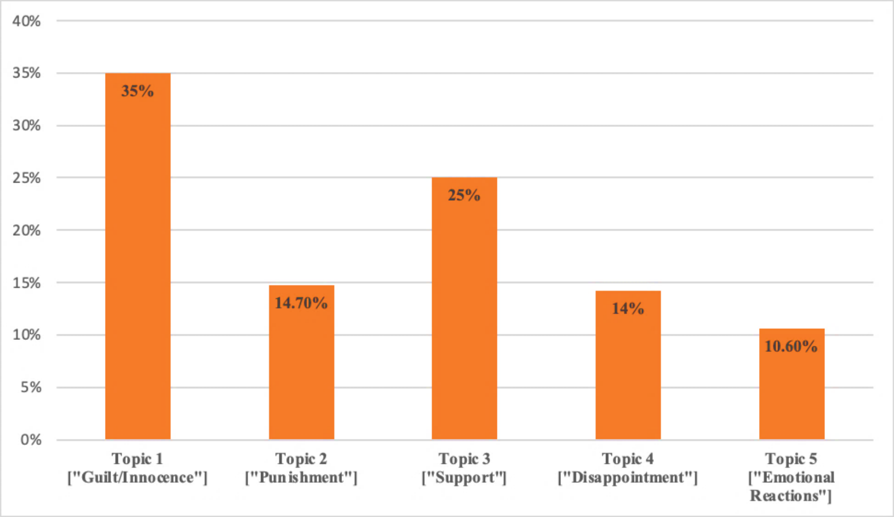
(1) *"Aaron Hernandez was actually pretty good too 😕 smh"*

(2) *"super shocked at the Aaron Hernandez verdict . . ."*

Figure 4 below summarizes the distribution of topics in Aaron Hernandez case:

[265] SMH=Shaking my head. *See What Does* SMH *Mean?*, MERRIAM-WEBSTER, https://www.merriam-webster.com/words-at-play/what-does-smh-mean-shaking-my-head (last visited Aug. 21, 2023) [https://perma.cc/2TJY-5SK8].

FIGURE 4: AARON HERNANDEZ TOPIC DISTRIBUTION



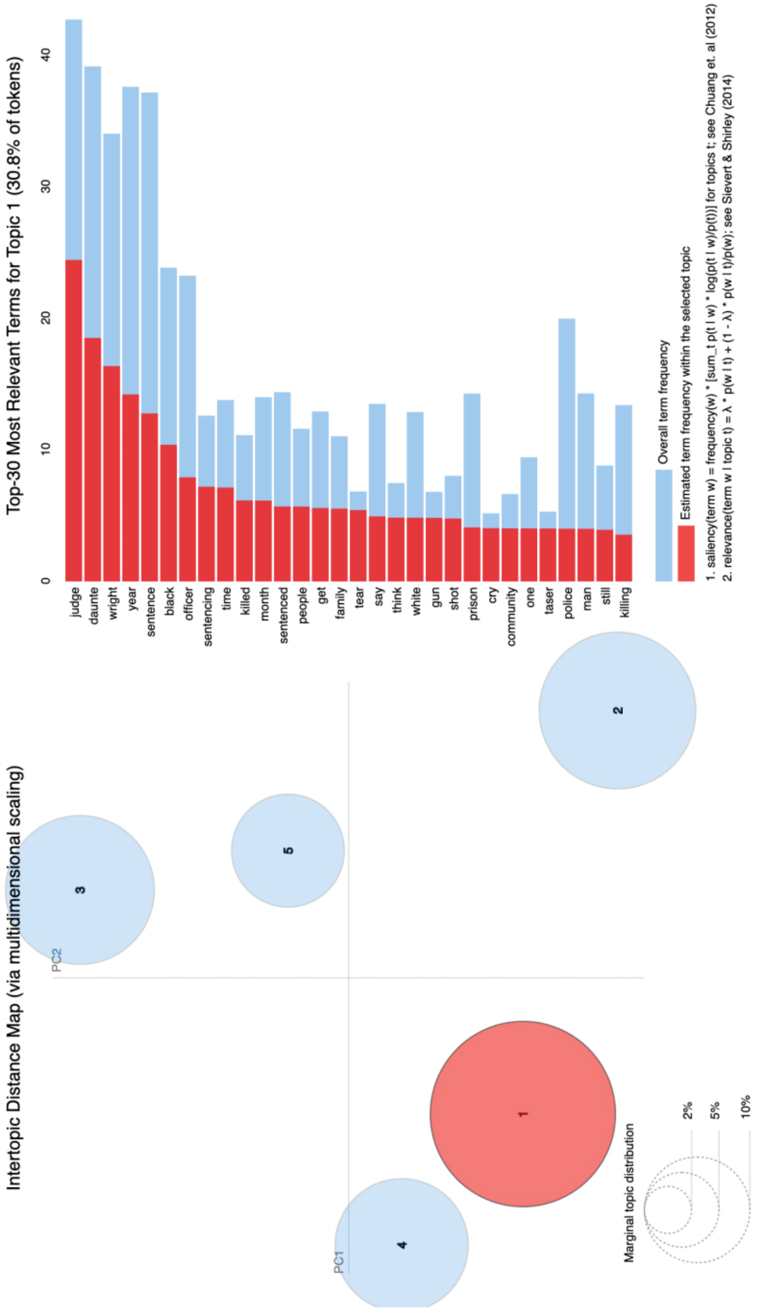
3. Kimberly Potter

Overall, we analyzed 160 tweets with sentiment in Potter's case. The small number of tweets affected the ability to tease out nuances in differences in terminology used in each topic, and while the distribution of words in each cluster varied, under most topics it was difficult to identify the topic based on the words alone. The supplemented qualitative analysis, however, contributed to our ability to tease out some differences between the topics, but we remain cautious regarding the allocation of topics.

For illustration, Figure 5 captures the semantic field identified under Topic 1. A link to a dynamic visualization of all topics can be found here.[266]

[266] INTERTOPIC DISTANCE MAP, https://rtmdrr.github.io/Renegotiating_Theories_of_Punishment/Potter.html (last visited Aug. 21, 2023) [https://perma.cc/66NJ-HBCM].

FIGURE 5: KIMBERLY POTTER MOST SALIENT TERMS, TOPIC 1



On the left-hand side, one can find the number of topics and their spread across the semantic universe. The size of each cycle represents the relative size of the topic (*i.e.*, number of tweets out of the overall tweets). On the right-hand side, one can find the relevant terms per topic and their frequency within that topic. Topic 1 is the largest topic and includes 46 tweets.

Topic 1 included terms like: "judge," "Daunte," "Wright," "year," and "sentence," focused on the implications of the case on the Black community as a whole, particularly leniency toward police brutality. We entitled this topic: "critique – carte blanche for police brutality."

Illustrative tweets:

> *(1) "The sentencing of Kimberly Potter is ridiculous. Judge is so very wrong. I am tired of seeing black men killed by police with such blatant systemic and racist bias. Judge: "Daunte mattered, but Cop was trying to do the right thing. Oh well. Downward departure from guidelines"*

> *(2) To hear that judge in the Daunte Wright case reference the Derrick Chauvin killing of George Floyd as part of her rationale to give Kim Potter a wrist-slap sentence was sickening. IMO, The msg she sent to sociopath cops was shoot rather than choke if lethal force is needed. SMH!*

Topic 2 included terms like "year," "police," "officer," "sentence," and "judge" and offered some support for the decision. This is the only topic in which blaming the victim (Wright) was used to support the defendant (Potter). We entitled this topic: "support—blame the victim."

Illustrative tweets:

> *(1) "Daunte Wright had a history of violent crime including allegedly shooting an 18 year old in the head who is now brain damaged. Police knew of the warrant for his arrest for carrying a gun without a permit when he was stopped. Kim Potter made a tragic mistake. It's complicated."*

> *(2) "Everyone is focused on the media narrative of "murder of a young black man" and not on the totality of the situation. A young black man in a felony arrest stop who was resisting arrest and attempting to flea with 2 other officers partially in the car with him. She made a mistake."*

Topic 3 included terms like "judge," "sentence," "Daunte," "Wright," and "justice," and offered a mix of views about the case, but was more facts heavy, that is, connecting a view to a set of facts related to the case. We entitled this topic: "facts based."

Illustrative tweets:

(1) "You noticed that AFTER handcuffing and banging the young black kid head and both kneeing him on his back both white police officers l will call Derek Chauvin and Kim Potter that officer Kim Potter walked over to the white kid and tapped him on his arm saying its ok."

(2) "Sure it was. She murdered somebody. Right afterwards, before she really would have ad time to process the killing as an accident, she fell to her knees and confessed. When her first few words were fear of punishment she showed her priorities. She didn't even run over to check him"

Topic 4 included terms like "sentence," "Daunte," "Wright," "year," and "Killing" and focused on criticizing the sentencing decision based on the systemic racism in the criminal legal system. The critique here was either general or focused on comparing other homicide cases in which the defendants were Black. We entitled this topic: "critique – systemic racism."

Illustrative tweets:

(1) "I tweeted comparing the sentences of white woman Kim Potter for killing a Black man (2 yrs) to Black woman Crystal Mason for casting a provisional vote while inelligible (5 yrs). To those who say it's anecdotal, it's NOT because this kind of injustice is EVERYWHERE in America."

(2) "Students, Kim Potter shot and killed a Black man at a traffic stop. Crystal Mason was on release from prison and cast a provisional ballot, that wasn't even counted, with help from a poll worker." "Who got 2 years, who got 5 years?" "One hint, Kim is white, Crystal is Black."

Topic 5 included terms like "judge," "prison," "mistake," "month," and "sentence," and is close to Topic 2 but focused more directly on the defendant's guilt/innocence arguing that what she did was a mistake and thus not a crime. We entitled this Topic: "Guilt/Innocence."

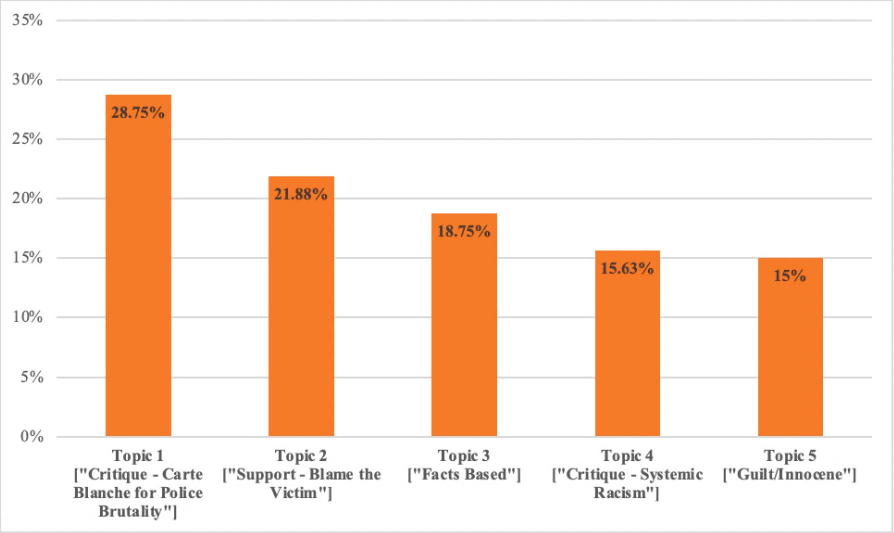
Illustrative tweets:

(1) "I may have to rethink my position. I agree with @DineshDSouza??? To me, the analogy is a surgeon making an HONEST mistake, doesn't put her in JAIL, unless it was caused by negligent behavior. For example it was shown she rushed for a tee time. But honest mistake? $$ not jail"

> (2) *"It wasn't murder, it was an accident, and you know that. Kimberly Potter should not be in prison at all"*

Figure 6 below summarizes the distribution of topics in Kimberly Potter's case:

FIGURE 6: KIMBERLY POTTER TOPIC DISTRIBUTION



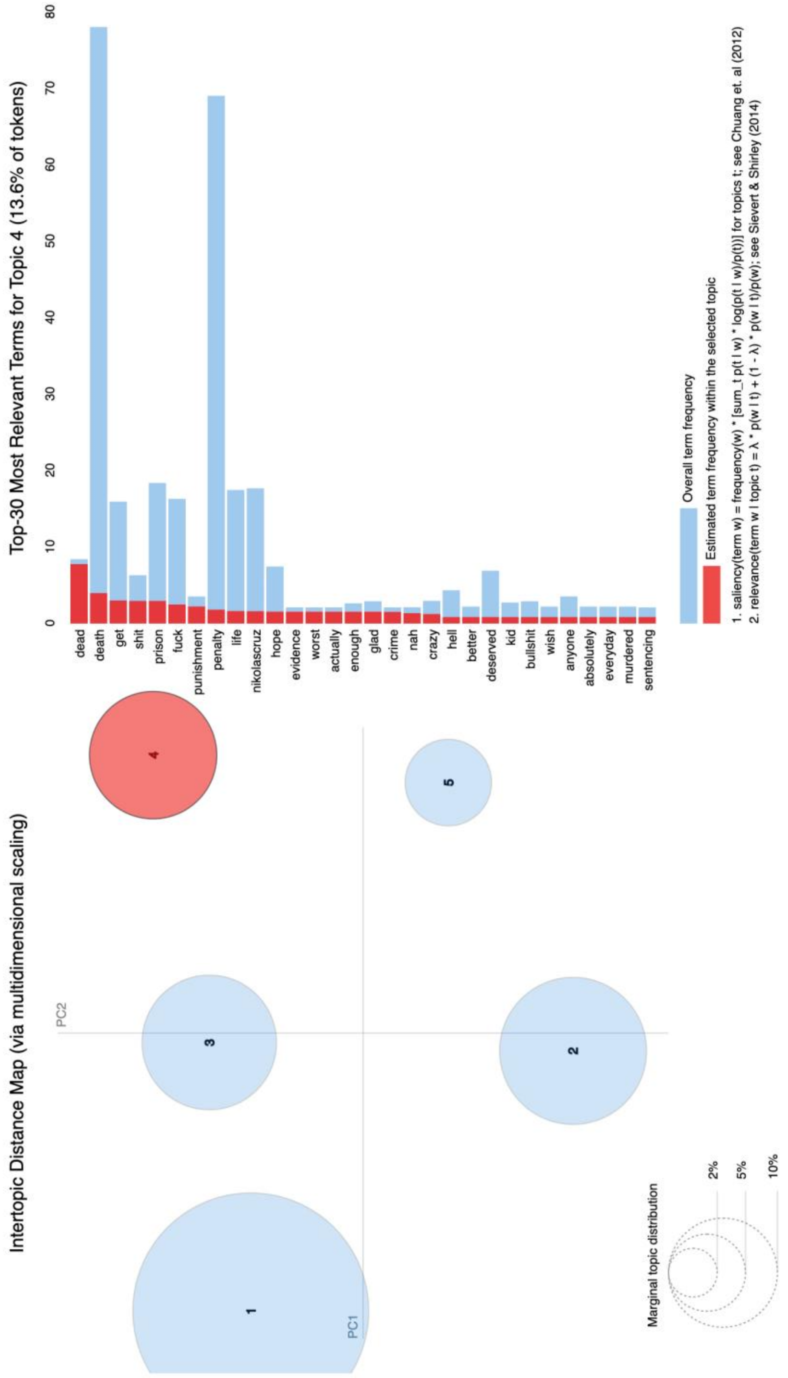
4. *Nikolas Cruz*

Analyzing the Twitter conversations around the jury's sentencing recommendation for Nikolas Cruz posed similar challenges to Potter's, as the universe of tweets was also much smaller than in the cases of Hernandez or Anthony. Overall, we analyzed only 126 tweets with sentiment. The small number of tweets affected the ability to tease out nuances in differences in terms used in each topic, as described below.

For illustration, Figure 7 captures the semantic field identified under Topic 4. A link to a dynamic visualization of all Topics can be found here.[267]

[267] INTERTOPIC DISTANCE MAP, https://rtmdrr.github.io/Renegotiating_Theories_of_Punishment/Cruz.html (last visited Aug. 21, 2023) [https://perma.cc/S39R-XDYK].

FIGURE 7: NIKOLAS CRUZ MOST SALIENT TERMS, TOPIC 4



Topic 1 includes terms like “death,” “penalty,” “life,” “prison,” “parkland,” “jury,” “deserves,” and “failed,” which we entitled “*Systemic Failure*” as it reflects the social media users’ views of a system that does not punish according to what they believe to be “just.”

Illustrative tweets:

(1) *“The Jury FAILED the children of America killed by mass shooters! They FAILED future mass shooting victims! They FAILED America and they FAILED justice! #NikolasCruz”*

(2) *“Nikolas Cruz deserves the death penalty. And I don’t believe in the death penalty. I’m so sad for the Parkland families today.”*

Topic 2 includes terms like “death,” “penalty,” “f…,” “sentence,” and “get,” which seems to reflect immediate emotional reactions to the sentencing recommendation, focusing on the *offender himself*. This topic seems to focus on more immediate responses and less on broader implications from the perspective of the criminal legal system.

Illustrative tweets:

(1) *“Nikolas Cruz (the parkland shooter) didnt get the death penalty, insane truly insane”*

(2) *“Oh wow. Nikolas Cruz didn’t get the death penalty* 😐*”*

Topic 3 includes terms like “death,” “penalty,” “get,” “prison,” “hope,” and “family.” It also seems to reflect responses to the sentencing recommendation, but this time through a *victim-centered angle*. The distinction between Topic 2 and 3 is close to the distinction identified in Anthony’s analysis between Topics 3 and 4.

Illustrative tweets:

(1) *“Tragedy on top of tragedy for the #Parkland families. Unjust end to the failure of sentencing Nikolas Cruz to death.”*

(2) *“WTF!!! They really didn’t give Nikolas Cruz the death penalty?!?! That’s straight bullshit. I’m so sorry to families and victims . So unfair he gets to live.”*

Topic 4 includes terms like “dead,” “death,” “get,” “shit,” “prison,” “punishment,” and “penalty,” which offers some justifications for the punishment users believe should be imposed. We entitled this topic “*justifications*.”

Illustrative tweets:

(1) *“I’m glad #NikolasCruz will spend the rest of his life in prison. And I’m also glad a brave juror refused to give him the death penalty.*

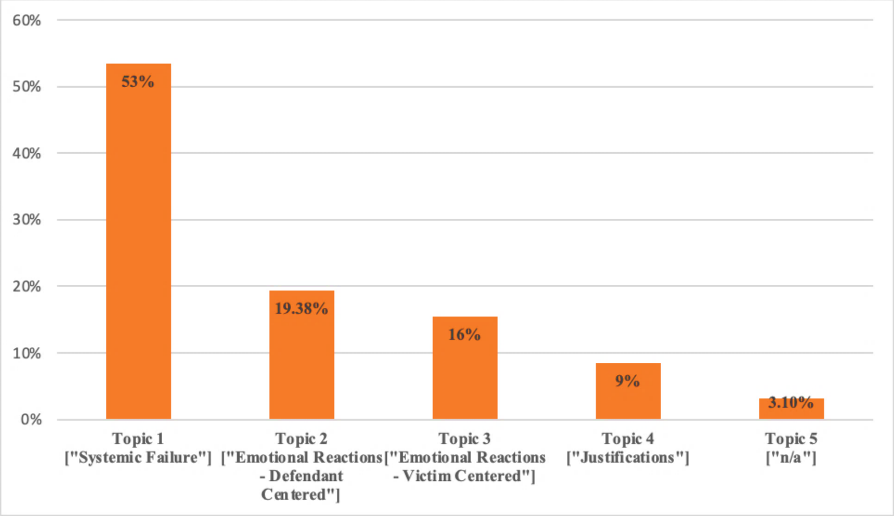
Capital punishment is wrong no matter how heinous the crime. #AbolishTheDeathPenalty"

(2) *"Wtf? Nikolas Cruz didn't get death? Fuck the system! The punishment should be greater than the crime. Life isn't enough!"*

Topic 5 includes terms like "life," "f…," "live," "jury," "wtf," "really," "one," and "kid," but includes only four tweets, which makes an assessment of this topic challenging.

Figure 8 below summarizes the distribution of topics in Nikolas Cruz's case:

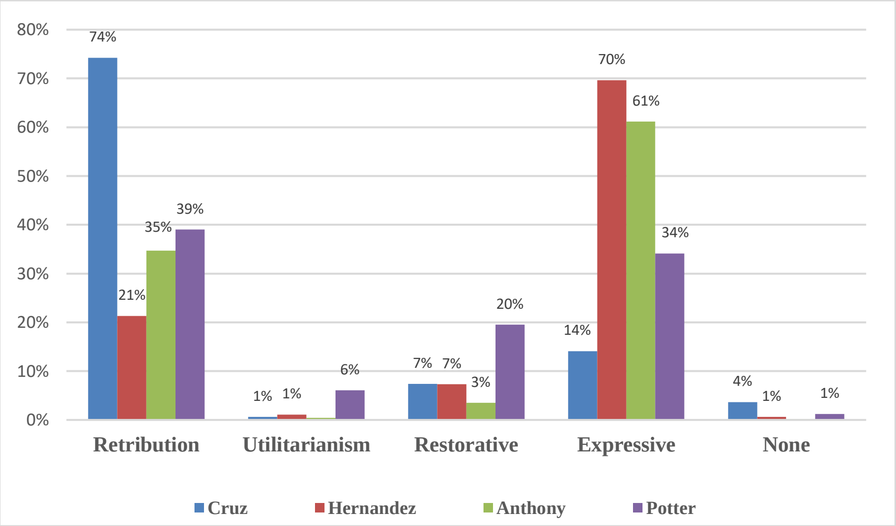
FIGURE 8: NIKOLAS CRUZ TOPIC DISTRIBUTION



B. Text Classification

Under the text classification model, we requested GPT 3.5 to classify the tweets based on four predefined categories aligned with four of the most dominant theories/alternatives for punishment discussed above: retributivism, utilitarianism, expressive, and restorative. We also requested the algorithm to explain the classification decision. For the purposes of this Article, we requested GPT to analyze all the tweets in Cruz and Potter's case and 30% of the tweets in Anthony and Hernandez's cases. The findings of the text classification are summarized in Figure 9 below:

FIGURE 9: TEXT CLASSIFICATION RESULTS—ALL CASES



The chart indicates, and aligned with previous studies, that in the context of sentencing and punishment, most views among the public reflect retributivist ideas. This was clearly the case in almost 75% of Cruz's tweets. Retributivism was also meaningfully identified in tweets surrounding Hernandez, Casey, and Potter's cases (21%, 35%, and 39% respectively). In both Anthony and Hernandez cases, however, GPT classified most of the tweets under the definition of expressive justifications for punishment.

While these findings can potentially shed light on the questions of interest at the core of this Article, we remain cautious as to the extent to which we should draw inferences from them, particularly regarding the "expressive" classification. This is mostly due to the novelty of utilizing GPT 3.5 as a text classifier and the validation process we undertook. As mentioned, to validate the results generated by the algorithm, we sampled several tweets and requested three research assistants to use the same set of categories to classify the tweets. One of us similarly analyzed the selected tweets. We then ran several interrater reliability tests to assess the level of agreement between the algorithm and the research assistants and the algorithm and the researcher. Overall, inter-rater reliability rates were low (under 0.2 Fleiss Kappa), limiting the ability to extract meaningful conclusions from the figure above. The Percent Agreement between the researcher and the algorithm regarding the classification of tweets as *retributivists* was, however, relatively high (85%). Percent Agreement does not take into account any agreement that occurs by chance, and as such, given the still low Cohen and Fleiss's Kappa statistics, we remain cautious regarding these findings. Having said that, we additionally performed a manual analysis of the algorithm's classification and explanations to several randomly selected tweets analyzed under the retributivist approach and revealed an overall reasonable grasp of the retributivist concept. As such, we are inclined to give more weight to GPT 3.5's classification under this category, as we will address in the discussion section below.

Given the high percentage of tweets classified by the algorithm as "expressive," we carefully investigated the analysis offered in this context as well, but as opposed to the classification under the retributivist definition, we didn't find that the explanations offered by the algorithm were necessarily aligned with the expressive ideas (the Percent Agreement between the researcher and algorithm was low, and so was Cohen's Kappa statistic). For example, we identified situations where the algorithm classified emotions toward offenders as expressive, or tweets that included expressions such as "wow," even when no references to punishment were found in the tweet. We would have classified many of these tweets as "none," that is do not reflect any of the main justifications. We also suspect—but this is speculative—that while the algorithm did determine that a small number of tweets do not fall under either category, the algorithm generally prefers offering a classification over determining that a tweet cannot be classified, which yielded some of the results. All of these observations might explain some of the results we saw under the "expressive" classification.

We do believe, however, that utilizing GPT 3.5 and more advanced models as a text classifier is methodologically promising, especially considering the ongoing research in NLP that aims to develop methods for improving the reliability and minimizing the hallucinations of these models.[268] Despite the low inter-rater agreement observed between GPT 3.5 and the human annotators, employing automated tools for text classification enabled us to analyze a substantial volume of tweets that would have been impractical to process within a reasonable timeframe and budget constraints with human annotators. While acknowledging the possibility of classification errors, this approach facilitates the identification of user opinion trends, which we can rely upon by validating a subset of the results through human expert validators. Nonetheless, further improvement and refinement are necessary for its optimal utilization.

VII. DISCUSSION AND IMPLICATIONS

As the findings indicate, issues of crime and punishment receive meaningful attention on social media. Users report, engage, analyze, debate, and express emotional responses to a case or its outcomes and even connect particular cases to broader themes related to the criminal legal system. As such, there is at least some value in studying social media data to better understand lay intuition about questions of crime and punishment. A different question, to be addressed below,

[268] Chandrashekhar S. Pawar & Ashwin Makwana, *Comparison of BERT-Base and GPT-3 for Marathi Text Classification*, *in* 936 FUTURISTIC TRENDS IN NETWORKS AND COMPUTING TECHNOLOGIES: SELECT PROCEEDINGS OF FOURTH INTERNATIONAL CONFERENCE ON FTNCT 2021 (2022). Jaromir Savelka, *Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts*, arXiv preprint arXiv:2305.04417 (May 8, 2023), https://arxiv.org/abs/2305.04417 [https://perma.cc/6NXQ-N9XE].

is how and for what purposes such data should be used. According to the empirical desert theory, lay attitudes about culpability and punishment should inform criminal law and policy. This Article, however, advocates approaching any deduction of such attitudes from social media data with caution, both on methodological and normative grounds, as we will discuss below.

Before delving into our substantive findings, a preliminary observation we categorize as a methodological challenge is warranted. Our data suggest that social media users are much more interested in questions of guilt or innocence than questions of punishment. We experimented with Twitter conversations around different procedural stages: the trial stage and the sentencing stage. The number of responses on social media to the events surrounding the trial stage (Anthony and Hernandez) was significantly higher than those surrounding the sentencing stage (Potter and Cruz); thousands of responses versus a bit more than a hundred. As such, it could be that data availability issues might hinder the ability to analyze a large corpus of social media data to assess laypeople's intuition of justice for the purposes of empirical desert, which focuses on questions of punishment. The analysis itself offered substantial support for these concerns: exploring the topics identified in Hernandez's and Anthony's cases indeed indicated that in both cases, the questions of punishment received less attention than questions of criminal culpability, the initial questions of guilt or innocence. Social media users addressed these questions of guilt or innocence either descriptively (*e.g.*, Hernandez was found guilty of murder), legally (*e.g.*, Hernandez shouldn't have been found guilty because there is no motive or evidence), or emotionally (*e.g.*, some found it so sad that the talented Hernandez was found guilty). As such, and if the scope and content of tweets related to Hernandez and Anthony are indicative of the dominant social media conversations surrounding criminal punishment, social media might not be a very helpful tool to assess the community's justice judgments.

Several observations from our analysis, however, might mitigate these concerns. Indeed, in both Hernandez's and Anthony's cases, the questions of guilt or innocence were at the core of the social and media discourse. As such, one can expect that less attention will be given to questions of punishment and its justification in this context, and these cases received much more attention from social media users. At the same time, as reflected in some of the topics identified under the TM analysis, some conversations related to these cases, in fact, included references to punishment (*e.g.*, Anthony's Topic 2 we revealingly entitled "punishment"). As such, one can identify conversations about punishment even when the focus is on the guilt (or innocence) of the suspect. The analysis of Cruz and Potter's cases might offer some additional insights with regard to this methodological challenge. Unlike Hernandez and Anthony, Nikolas Cruz pleaded guilty to all of his charges, and Potter was already charged when we ran the analysis, and as such, questions of culpability were not at the heart of the media frenzy revolving around their case. The main question of interest to the public was exactly whether—assuming guilty—they should be punished and, if so, what should be the punishment: in Cruz's case, whether he should be sentenced to death (or to life in prison without parole), and in Potter's case whether her punishment was appropriate. While the TM analysis did not necessarily reflect such an outcome, triangulating the TM analysis with a qualitative analysis indeed revealed that *ALL* social media conversations around Cruz's case revolved around questions of "just punishment," mostly whether he should be sentenced to death and, if so, why. The analysis of Potter's case, however, was less conclusive. While we indeed identified tweets discussing the appropriateness of her punishment, many tweets continued conversations about Potter's guilt, for example, those tweets that fall under Topic 5 ("guilt/innocence"). In sum, there seems to be a potential trade-off between focusing on Twitter conversations related to the trial stage and the sentencing stage. While substantially more data can be found in the first group, the second group is likely more relevant for directly assessing questions related to lay people's attitudes regarding punishment. Given these challenges, a combination of NLP tools is likely to be most helpful in answering the questions of interest. With this in mind, we move on to discussing some of our substantive findings and their meanings.

First, while the cases we analyzed were different, they shared a common thread: a combination of the themes identified in the TM analysis and the qualitative analysis suggested that in all four cases, and particularly in Casey Anthony's and Nikolas Cruz's cases, a meaningful portion of the users on social media were *disappointed* with the decisions rendered by the representatives of the criminal legal system because they found them in tension with the users' vision of what is "right" or "just." For some, as indeed argued by proponents of "empirical desert" as a distributive principle, such dissatisfaction was attached to the legitimacy of the system as a whole (*e.g.*, Topic 1 in Cruz's case, entitled "systemic failure" is indicative of such connections, and so are Topics 1 and 4 in Potter's case, discussing issues of systemic racism and support for police brutality as a consequence of the sentencing decision). In cases like Hernandez or Potter, in which issues of racial inequality were salient, dissatisfaction with the decisions (verdict/sentencing) was translated into dissatisfaction from a racist criminal legal system (as expressed, *e.g.*, in Hernandez Topic 3, entitled "support"). Interestingly, users were often disappointed in the system when it was *not* punitive enough: when the juries found Anthony not guilty, recommended not to impose the death penalty on Cruz, or when the judge sentenced Potter for what they perceived as a lenient sentence. In Hernandez's case, there was a meaningful portion of voices that were not pleased with the verdict (*i.e.*, finding him guilty), but others believed it was the right decision. Many of the users who expressed negative sentiments with respect to the outcome did so not because they believed the decision was legally wrong but because they were disappointed with Hernandez himself, choosing the criminal path over what seemed to be a potential life of financial and professional success. In Potter's case as well, some voices reemphasized the view that she shouldn't have been found guilty to begin with.

As such, just by exploring broad patterns in the data, one can find some support for a proposition already identified in previous research: the connections between the legitimacy of the system in the eyes of its constituents and their alignment with the system's perceptions of justice. Indeed, the alignment of these findings with previous studies can first offer some validation to the methodology utilized. Furthermore, they can start to shed light on this Article's first question of interest, that is, whether social media analysis can contribute to our understanding of laypeople's intuition of justice and suggest there is some potential in studying social media data in this context. However, as we elaborate below, drawing inferences from social media data to society more broadly should be done carefully and with proper safeguards.

Additionally, a deeper look into the findings generated by the TM algorithm, and to a lesser extent the text classification, offers additional answers to the questions in hand, particularly the potential contribution of social media research to the assessment of lay intuitions regarding punishment.

Another conclusion to be drawn from the analysis is that social media users do not thoroughly explain their intuitions of justice. Posts like "Casey Anthony deserves to die!" were most common, with some connecting the killing of her child to the equation (ignoring the fact she was found not guilty on the killing charges). But given previous studies, this is not surprising, as people's views regarding punishment are first and foremost intuitional rather than reasoned[269] and the nature of social media often exposes such intuitive responses. Moreover, Twitter, as a tool that offers limited textual opportunities, seems to limit, and clearly not encourage, reasoned responses. As such, the limitations of Twitter might, in fact, align with what research has shown regarding human judgments of justice, that is that they are more gut reactions rather than reasoned justification.

As for the actual views expressed, our assessment is that these social media conversations are indicative of a general support for retributivist ideas or "just deserts." They suggest, for example, that social media users assess the punishment based on the severity of the criminal act. Particularly, so suggest many of the tweets, killing should be answered with a killing. This was particularly evident in Cruz's case but also in Anthony's. However, even in Hernandez's case, where the question of the death penalty was not dominant, similar views of "just deserts" were presented (*e.g.*, claiming that Hernandez should be sentenced to death because he killed a human being). In Potter's case, we found a large number of tweets suggesting she should be punished more severely because she took someone's life. The text classification analysis—though limited—also offered some support to this conclusion, particularly in the context of Cruz's case (in which 75% of the tweets were categorized as retributivist by the text classifier), but also in Anthony, Hernandez, and Potter's cases.

[269] Robinson, *supra* note 7, at 1569–70.

It should be noted that a small number of tweets offered a different version of "just deserts," mostly based on ideological objections to the death penalty. These voices agreed, however, that defendants (like Hernandez) *should* be punished *because* they took someone else's life.

Furthermore, in order to assess whether other perspectives and viewpoints to justify punishment might have some presence on social media, we manually searched for additional references to punishment justifications, using key phrases and concepts used in previous experimental studies to investigate utilitarian viewpoints. Some minor references to Hernandez's capability for remorse were identified, suggesting some additional factors might be considered by users in inflicting punishment. In general, however, and based on our analysis, we concluded that retributivist notions of justice were most dominant among social media users addressing questions of punishment.

Looking into the limited universe of tweets related to Cruz and Potter—that focused on punishment—indicated that retributivist ideas, alongside moral apprehension from Cruz's act and, to a lesser extent, Potter's, keep controlling the conversations. In Cruz's case, this led to the conclusion that the jury should have recommended the death penalty and, in Potter's, that she should have been punished more severely. These views indeed mimic similar propositions to those extracted from the tweets related to Hernandez and Anthony, all suggesting that the majority of social media users lean toward retributivist notions of punishment and, to some extent, in the most extreme version of proportionality, that is, that one be sentenced to death for the life one took. As discussed in Part III, previous studies have also indicated that individuals make punishment decisions based on retributivism more than any other theoretical justification. As such, despite the meaningful concerns regarding selection bias and the representativeness of social media users, these alignments with previous studies not only validate the methodology used but suggest that this methodology might do better than expected predicting general views in society. This is particularly true given studies suggesting that "people may agree on the relative seriousness of many aspects of core wrongdoings," what Robinson considers the core "community view."[270] Note, however, that our analysis (for example, in Cruz Anthony's cases) identified a tendency to impose harsher punishments (even if one was found not guilty). Such a tendency seems to align with popular "tough on crime" ideas and thus might contradict previous studies discussed earlier, which suggest that individuals tend to impose punishments that are less severe than those imposed by the criminal law.[271]

But even if we accept that social media analysis might have value in understanding that thin layer of agreement related to the relative seriousness of the criminal act and the punishment that should be thus imposed, it is most likely not nuanced enough. Particularly, social media narratives do not seem to provide a

[270] *Id.* at 1573.

[271] *Id.* at 1575–1579.

sufficient account of the differences between communities based on cultural, demographic, or other variables. Exploring the data related to Hernandez, Anthony, and Potter exposes such potential disagreements based on race or gender. In fact, particularly in the Hernandez case, intuitions related to his guilt and blameworthiness more broadly seemed affected by racial predispositions (for example, tweets calling for harsh punishments were often accompanied by racist slurs.) Tweets calling for his release, however, were often accompanied by references reflecting group solidarity such as "brother" or "my man." In Potter's case, many of the tweets that raised frustration from the punishment tied it with broader themes related to systemic racism and police brutality against the Black community. In Anthony's case, where the majority seemed to support conviction (and then harsh punishment), many of the references to her punishment included misogynist claims. Scholarship on the nature of exposure to social media content might explain such findings, and a deeper understanding of the communities dominating social media is required before we use social media data as a source for assessing "lay intuition" of justice.[272] Indeed, our findings offer some support to claims raised by Mary Anne Franks and others suggesting that social media serves as a site to preserve the dominant American social hierarchy. As such, social media as a platform might exacerbate the critique introduced against "empirical desert" and the objections offered by other scholars to the potentially immoral views of the community at large.[273]

These conclusions lead to an additional question: even if we are able to establish that social media might offer some form of direction in identifying general community views regarding criminal justice, what is the meaning, or the value, of identifying attitudes toward culpability and punishment on social media? From the perspective of empirical desert, lay attitudes can and should be taken into account when deciding criminal laws and policies.[274] Should data analyzed from social media be equally utilized? That is, and let us assume for a second that we are able to overcome some of the methodological concerns raised earlier, *should* social media data be used to inform criminal law and policy? Given the racist, misogynist, and other offensive expressions that are so tightly intertwined with the views expressed by social media users, our initial response is probably not. As others have indicated, much of what we see on social media does not reflect a race to the center, the "core" agreement Robinson refers to in his writing, but instead, we witness on social media a process of polarization and radicalization, in which sub-groups keep reiterating their views and perspectives, not in an attempt to advance dialogue but rather in a combative manner.[275] Furthermore, discourse on social media often preserves structural social hierarchies

[272] Franks, *supra* note 199.

[273] Robinson, *supra* note 4, at 29; Rudyak, *supra* note 113.

[274] *See e.g.*, Paul Robinson, *The Ongoing Revolution in Punishment Theory: Doing Justice as Controlling Crime*, 42 ARIZONA ST. L.J. 1089–90 (2010) (illustrating how community views were taken into account in criminal law and policies across various countries).

[275] Travis L. Dixon, *Understanding How the Internet and Social Media Accelerate Racial Stereotyping and Social Division, in* RACE AND GENDER IN ELECTRONIC MEDIA 161 (Ann Lind Rebecca, ed., 2017).

and "reinforce[s] radically unequal distributions of social, economic, cultural, and political power."[276] In that sense, social media is, in fact, not inclusive or democratic. As such, and given that empirical desert aspires to democratize criminal law, it can be argued that social media is an inherently flawed source of information to achieve that goal. In fact, social media seems to end up muting diverse voices of non-majoritarian communities or at least creating a space that preserves power imbalances within our society. Furthermore, the arguments regarding the immorality of lay intuition should receive special attention in the context of social media because, as others have shown, and the data analyzed for this study supports as well, social media is a fertile ground for offensive expressions that give rise to biased judgments against individuals.[277] As such, assessing community views of justice for the purposes of designing criminal law and policy based on these expressions, among others, is problematic as it normalizes these offensive views. Under these circumstances, some guidance from moral philosophers might not be so bad.

VIII. CONCLUSION

This Article offers new directions to explore and engage with old questions: how can we justify criminal punishment, and if we can, what should be the punishment imposed on a criminal wrongdoer? Empirical desert scholarship argues that the best answers to these questions can be found by assessing the shared community view of justice, later to be used in the design of criminal law and policy. We argued that empirical scholars have not yet explored an additional domain that can contribute to our understanding of such shared community view: social media. By reviewing interdisciplinary scholarship that utilizes social media for understanding human behavior in other settings, the Article engaged with the promise alongside the challenges of adopting this methodological approach. Utilizing developments in machine learning and natural language processing, the Article further offered a novel, if exploratory, analysis of social media discourse around culpability and punishment through two different methodologies. We identified some methodological challenges in the analysis of social media for purposes of assessing lay people's attitudes regarding punishment and offered directions for future research. We also discussed our substantive findings. On the one hand, some of these findings aligned with previous experimental scholarship assessing lay attitudes regarding criminal culpability and punishment. On the other hand, they exposed some concerns—likely exacerbated in the social media domain—that offer support to the existing critique of empirical desert theory's calls to consider lay people's attitudes in criminal law and policy design. The Article concluded by discussing the potential contribution of analyzing social media discourse around punishment to our understanding of community views of justice, alongside the practical, methodological, and normative limitations of

[276] Franks, *supra* note 199, at 429.

[277] John Suler, *The Online Disinhibition Effect*, 7 CYBERPSYCHOLOGY & BEHAV. 321, 321 (2004).

following such an approach. Although social media appears to hold the potential to contribute to our understanding of laypeople's attitudes toward punishment, the challenges we have discussed may ultimately outweigh its suitability, especially for the purposes of criminal law and policy design. If the objective is to democratize criminal law by embracing the community's perceptions of justice, it becomes crucial to amplify the voices of marginalized communities who bear the brunt of the criminal legal system. We harbor doubts, however, about whether social media can effectively serve as a platform to accomplish this aim.

* * *