CRIME BECAUSE PUNISHMENT? THE INFERENTIAL PSYCHOLOGY OF MORALITY AND PUNISHMENT

*Jessica Bregant**
*Eugene M. Caruso***
*Alex Shaw****

Psychologically speaking, punishment may operate as a special case of social norm information, but what sets punishment apart from other norms is the moral weight punishment carries. Although norms other than punishment may also communicate moral messages, punishment seems to be unique in its relationship to morality, and especially to judgments of harm. Prior research demonstrates that potential punishers rely heavily on the degree of harm caused by wrongdoing when determining the appropriate level of punishment. In this Article, we show that the opposite is also true—information about punishment can influence the extent to which an act of wrongdoing is judged to have been harmful. Part I reviews existing research on the message of punishment, drawing on literatures from law, psychology, and philosophy. We also highlight closely related research on social norms and behavior. In Part II, we present four original experiments (total N = 890). The results suggest that punishment is indeed a signal of harm, and that, like social norm information, punishment can be an effective cue for moral judgment. Finally, in Part III, we discuss some of the implications of our findings, including their relevance to debates about corporate malfeasance, non-prosecution of financial crimes, and legal reactions to police misconduct.

TABLE OF CONTENTS

* Jerome Hall Postdoctoral Fellow, Indiana University Maurer School of Law. Address correspondence to jessica.bregant@gmail.com.

** Associate Professor of Management and Organizations and Behavioral Decision Making, UCLA Anderson School of Management.

*** Assistant Professor of Psychology, University of Chicago.

I. INTRODUCTION

Throughout the literatures of law, psychology, and philosophy, a great deal of attention has been paid to the question of why people seek to punish one another.[1] Amid all the discussion of what punishment *should* and *can* accomplish or communicate, however, relatively little thought has been given to what punishment actually *does* signal.[2] Neglect of punishment's signal is no small oversight; many theories of punishment, from deterrence to restorative justice, rely on the assumption that lay people will understand punishment in a particular way that is consistent with normative theory.[3] If this assumption is mistaken, it could undermine the strength and legitimacy of punishment policy.

In this Article, we present empirical evidence that speaks to the most basic way punishment may be understood by the lay public. We pose a simple research question: What do people infer about an action based on the fact that it is punished? Psychologically speaking, punishment may operate as a special case of social norm information, but we argue that what sets punishment apart from other

1. *See* Kenworthey Bilz, *The Puzzle of Delegated Revenge*, 87 B.U. L. REV. 1059 (2007); Joshua W. Buckholtz et al., *The Neural Correlates of Third-Party Punishment.*, 60 NEURON 930, 930–40 (2008); Kevin M. Carlsmith, John M. Darley & Paul H. Robinson, *Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment*, 83 J. PERS. SOC. & PSYCHOL. 284, 284–99 (2002); Fiery Cushman et al., *Accidental Outcomes Guide Punishment in a "Trembling Hand" Game*, 4 PLOS ONE 1 (2009); John M. Darley & Thane S. Pittman, *The Psychology of Compensatory and Retributive Justice.*, 7 PERS. & SOC. PSYCHOL. REV. 324, 324–36 (2003); Robert Kurzban, Peter DeScioli & Erin O'Brien, *Audience Effects on Moralistic Punishment*, 28 EVOLUTION & HUM. BEHAV. 75, 75–84 (2007); Uli Orth, *Punishment Goals of Crime Victims.*, 27 LAW & HUM. BEHAV. 173, 173–86 (2003); Michael E. Price, Leda Cosmides & John Tooby, *Punitive Sentiment as an Anti-Free Rider Psychological Device*, 23 EVOLUTION & HUM. BEHAV. 203, 203–31 (2002).

2. *But see, e.g.*, JEFFRIE G. MURPHY & JEAN HAMPTON, FORGIVENESS AND MERCY 1 (1988); Kenworthey Bilz, *Testing the Expressive Theory of Punishment*, 13 J. EMPIRICAL LEGAL STUD. 358, 358–392 (2016).

3. *See* Darley & Pittman, *supra* note 1, at 325.

norms is the moral weight punishment carries. Although norms other than punishment may also communicate moral messages, punishment seems to be unique in its relationship to morality, and especially to judgments of harm. Prior research demonstrates that potential punishers rely heavily on the degree of harm caused by wrongdoing when determining the appropriate level of punishment.[4] In this Article, we show that the opposite is also true—information about punishment can influence the extent to which an act of wrongdoing is judged to have been harmful.

In the first part of this Article, we discuss the existing research on the message of punishment, drawing on literature from law, psychology, and philosophy. We also highlight closely related research on social norms and behavior. Our review of this literature concludes with a summary of research on punishment, moral judgment, and harm. In the second part of the Article, we present original experimental evidence that punishment can be an effective cue for moral judgment, influencing such judgments in a way that is similar to social norm information. Interestingly, however, punishment seems to most effectively signal a specific moral concern—harmfulness—especially relative to social normative information. Finally, in Part III, we discuss some of the important implications of our findings, including their relevance to debates about corporate prosecution, financial crimes, and police misconduct.

II. PUNISHMENT, NORMS, AND MORAL PSYCHOLOGY

The importance of punishment to law is almost tautological: laws without law *enforcement* mechanisms are little more than aspirations or norms. Though the mechanisms by which laws are enforced vary, most such mechanisms can be broadly described as punishments. Legal punishment therefore marks, at least, the difference between a legal rule and a merely normative one. This difference is psychologically important; as we discuss below, the presence of a legal rule—even one which carries only nominal sanctions—seems to influence behavior. Punishment itself, however, also has a special psychological significance in forming an important building block of human moral reasoning and moral development.

In this Part, we briefly review three related bodies of research that each address an important aspect of the present studies. First, we describe the extant research on the so-called expressive function of law, which demonstrates the power of laws to influence behavior. We next examine research on moral and social norms outside of the legal context; although laws undoubtedly provide normative information, our discussion highlights some ways in which the analogy between norms and laws can break down. Finally, we review the importance of punishment to moral reasoning in particular, paying special attention to the feature of moral judgment that appears to be most related to punishment: harm.

4. Kevin M. Carlsmith, *The Roles of Retribution and Utility in Determining Punishment*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 437, 437–51 (2006); Cushman et al., *supra* note 1, at 6.

A. The Expressive Function of Punishment

A single act of punishment can attempt to accomplish many simultaneous ends. For example, the target of punishment may be deterred from future wrongdoing by the threat of future punishment,[5] other members of the community may be deterred from imitating the target's wrongdoing,[6] the target may be incapacitated (*i.e.*, through incarceration[7]) or rehabilitated (*i.e.*, through treatment[8]), or restitution may be made to the victims of wrongdoing.[9] More diffuse retributive interests, such as correcting the moral scales or meting out justice,[10] may also be pursued. Beyond these instrumental ends, however, punishment—or, maybe more precisely what and who we *choose* to punish—carries a communicative

5. *See* MARTHA C. NUSSBAUM, ANGER: DOWNRANKING, WEAKNESS, PAYBACK (2013); John M. Darley & Adam L. Alter, *Behavioral Issues of Punishment, Retribution, and Deterrence*, *in* THE BEHAVIORAL FOUNDATIONS OF PUBLIC POLICY 186 (Eldar Shafir ed., 2013); Darley & Pittman, *supra* note 1, at 329; Robert L. Rabin, *Pain and Suffering and Beyond: Some Thoughts on Recovery for Intangible Loss*, 55 DEPAUL L. REV. 359, 366 (2006).

6. B. Sharon Byrd, *Kant's Theory of Punishment: Deterrence in its Threat, Retribution in its Execution*, 8 LAW & PHIL. 151, 152–53 (1989); John S. Carroll et al., *Sentencing Goals, Causal Attributions, Ideology, and Personality*, 52 J. PERS. SOC. & PSYCHOL. 107, 107–18 (1987); Neil Vidmar & Dale T. Miller, *Socialpsychological Processes Underlying Attitudes Toward Legal Punishment*, 14 LAW & SOC'Y. REV. 565, 573 (1980); NUSSBAUM, *supra* note 5; Orth, *supra* note 1, 173.

7. John M. Darley, Kevin M. Carlsmith & Paul H. Robinson, *Incapacitation and Just Deserts as Motives for Punishment*, 24 LAW & HUM. BEHAV. 659, 661 (2000); David P. Farrington, *Age & Crime*, 7 CRIME & JUST. 189, 223 (1986); Paul H. Robinson & John M. Darley, *The Utility of Desert*, 91 NW. U. L. REV. 453, 464–65 (1997).

8. F. T. Cullen et al., *Public Support for Correctional Treatment: The Tenacity of Rehabilitative Ideology*, 17 CRIM. JUST. & BEHAV. 6, 6–7 (1990); R. C. McCorkle, *Research Note: Punish and Rehabilitate? Public Attitudes Toward Six Common Crimes*, 39 CRIME & DELINQ. 240, 240–41 (1993); Tony Ward & Russil Durrant, *Evolutionary Psychology and the Rehabilitation of Offenders: Constraints and Consequences*, 16 AGGRESSION & VIOLENT BEHAV. 444, 445–46 (2011).

9. Tony Ward & Robyn Langlands, *Repairing the Rupture: Restorative Justice and the Rehabilitation of Offenders*, 14 AGGRESSION & VIOLENT BEHAV. 205, 211–12 (2009); Charlotte V.O. Witvliet et al., *Retributive Justice, Restorative Justice, and Forgiveness: An Experimental Psychophysiology Analysis*, 44 J. EXPERIMENTAL SOC. PSYCHOL. 10, 12 (2008); Ellen A. Waldman, *Healing Hearts or Righting Wrongs?: A Meditation on the Goals of "Restorative Justice"*, 25 J. HAMLINE J. PUB. L. POL'Y 355, 355, 360–63 (2003).

10. *See* Michael T. Cahill, *Retributive Justice in the Real World*, 85 WASH. U. L. REV. 815, 870 (2007); Katrina M. Fincher & Philip E. Tetlock, *Brutality Under Cover of Ambiguity: Activating, Perpetuating, and Deactivating Covert Retributivism*, 41 PERSONALITY & SOC. PSYCHOL. BULL. 629, 629 (2015); Thomas Grisso, *Society's Retributive Response to Juvenile Violence: A Developmental Perspective*, 20 LAW & HUM. BEHAV. 229, 235–38 (1996); Jean Hampton, *Correcting Harms Versus Righting Wrongs: The Goal of Retribution*, 39 UCLA L. REV. 1659, 1700–01 (1992); Jan-Willem van Prooijen et al., *Power and Retributive Justice: How Trait Information Influences the Fairness of Punishment Among Power Holders*, 50 J. EXPERIMENTAL SOC. PSYCHOL. 190, 199–200 (2014).

weight.[11] This communicative aspect of law is often called its "expressive" function.[12]

Though sometimes given as an alternative to retributive or utilitarian theories of punishment, expressive functions of punishment are essentially orthogonal to these aims; the message communicated by a punishment act may itself be retributive, utilitarian, neither, or both of these. Expressive theories of punishment are theoretically similar to so-called "signaling" accounts that are prevalent in the literatures of evolutionary science and economics,[13] because both theories hold that an action can send a message, over and above the immediate consequences of the action itself.[14] However, empirical studies of signaling accounts are generally precise as to the content of the message being sent and received—for example, some gazelles engage in a kind of jumping called "stotting" that appears to send an honest signal to predators about the gazelle's health.[15] In contrast, empirical studies looking at the expressive functions of law tend to be vague about the content of the message sent by punishment.[16] Even more importantly, the content of the message *received* has been left virtually unexplored by empirical research.[17]

To our knowledge, only two experimental studies have examined the message(s) that are communicated by punishment;[18] both find support for a particular view of the expressive function that is sometimes called "expressive retributivism."[19] Under this theory, crimes are themselves expressive acts that send a message to a victim and to society about the standing of the victim relative to the offender.[20] Punishment, in contrast, sends the opposite message, rejecting the

11. *See* MURPHY & HAMPTON, *supra* note 2, at 1–13; 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, 168 (2002); Dan M. Kahan, *What Do Alternative Sanctions Mean?*, 63 U. CHI. L. REV. 591, 631–32 (1996); Cass R. Sunstein, *On the Expressive Function of Law*, 144 U. PA. L. REV. 2021, 2050–51 (1996).

12. *See* Bilz, *supra* note 2, at 358–59; Heather J. Gert et al., *Hampton on the Expressive Power of Punishment*, 35 J. SOC. PHIL. 79, 79 (2004); Jason Mazzone, *When Courts Speak: Social Capital and Law's Expressive Function*, 49 SYRACUSE L. REV. 1039, 1039 (1999); Sunstein, *supra* note 11, at 2021–25.

13. *See* Joseph Bulbulia & Richard Sosis, *Signalling Theory and the Evolution of Religious Cooperation*, 41 RELIGION 363, 364 (2011); Brian L. Connelly et al., *Signaling Theory: A Review and Assessment*, 37 J. MGMT. 39, 39 (2011); C.D. FitzGibbon & J.H. Fanshawe, *Stotting in Thomson's Gazelles: An Honest Signal of Condition*, 23 BEHAV. ECOLOGY & SOCIOBIOLIOGY 69, 69 (1988); Richard D. Morris, *Signalling, Agency Theory and Accounting Policy Choice*, 18 ACCT. & BUS. RES. 47, 48 (1987).

14. *See* Connelly et al., *supra* note 13, at 43–45; Gert et al., *supra* note 12, at 82.

15. FitzGibbon & Fanshawe, *supra* note 13, at 69. The gazelle's health, of course, influences its potential ability to escape. *See id.* at 73.

16. *See* Robert Cooter, *Do Good Laws Make Good Citizens? An Economic Analysis of Internalized Norms*, 86 VA. L. REV. 1577, 1581 (2000); Patricia Funk, *Is There an Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines*, 9 AM. L. & ECON. REV. 135, 155 (2007); Maggie Wittlin, *Buckling Under Pressure: An Empirical Test of the Expressive Effects of Law*, 28 YALE J. REG. 419, 421 (2011).

17. *But see* Bilz, *supra* note 2, at 358–62.

18. *See, e.g.*, Bilz, *supra* note 2; Jessica Bregant, Alex Shaw & Katherine D. Kinzler, *Intuitive Jurisprudence: Early Reasoning About the Functions of Punishment*, 13 J. EMPIRICAL LEGAL STUD. 693, 693.–717 (2016).

19. *See* Bilz, *supra* note 2, at 363–85; Bregant et al., *supra* note 18, at 698–712.

20. MURPHY & HAMPTON, *supra* note 2, at 24–25; Bilz, *supra* note 2.

offender's false claim and restoring the victim's position in society.[21] In a set of experiments testing this view, Kenworthey Bilz found that, across a variety of crimes, punishment decreases the social standing of the offender and—crucially—increases the social standing of the victim.[22] In a study of children aged five to eight years-old, Bregant, Shaw, and Kinzler similarly found that children liked the victim of a theft more if the thief who committed the act was punished, compared to when the thief went unpunished.[23]

The expressive retributivism argument centers on condemnation of the bad actor, especially relative to the victim, rather than on condemnation of the act itself. However, if punishment sends a message of condemnation, psychological evidence suggests the condemnation need not be limited to the actor.[24] For example, Bregant, Shaw, and Kinzler also found that children used punishment as a signal of how "bad" the *act* of stealing is—in a world where those who steal are "never punished," children between the ages of five and eight overwhelmingly reported that stealing was not "bad."[25]

This divergence in children's reactions is, in some ways, a microcosm of the bigger questions surrounding expressive punishment messages because it highlights two major themes that are relevant: social norms and moral condemnation. There are at least two possible explanations for children's beliefs that stealing is not "bad" when it is not punished. One possibility is that punishment information communicates that an action is "bad" in the same way that it is "bad" to eat with one's hands at dinner. That is, punishment may merely be communicating that the action in question is a conventional violation of social norms.[26] A second possibility is that punishment information communicates something about whether the action is immoral.[27] That is, that this action is wrong intrinsically and immutably, which might cause people to infer that the action is harmful or morally disgusting. These two possibilities—social norms and moral judgment—are both cited in the broader literatures as possible messages of punishment,[28] and we explore both below.

21. MURPHY & HAMPTON, *supra* note 2, at 1–4.

22. *See generally* Bilz, *supra* note 2.

23. Bregant et al., *supra* note 18, at 712. Recent research in social neuroscience further emphasizes the importance of the victim in moral judgments. *See* Indrajeet Patil et al., *The Behavioral and Neural Basis of Empathic Blame*, 7 SCI. REP. 1, 1–2 (2017) (finding that empathy for the victim contributes to moral blame, even when the harm is accidental).

24. *See* Bregant et al., *supra* note 18, at 708–12.

25. *See id.*

26. *See infra* Part II.B.

27. *See infra* Part II.C.

28. *See infra* Part II.B–C.

B. Normative Messages, Laws, and Punishment

Although empirical evidence of the messages of punishment is scarce, theories abound. One especially common characterization of the expressive function is that laws express social norms.[29] A vast literature in social psychology illustrates the power of social norms to influence behavior across a wide variety of contexts.[30] Experimentally, normative information has been used to reduce self-reported speeding,[31] increase energy conservation,[32] and curb college alcohol use.[33] When people think that "everyone else" is doing something, they are more likely to engage in that something themselves.[34]

If laws are perceived as the codification of social norms, then information about the legal status of an act could have a similar effect on behavior. Of course, laws may change behavior for other reasons as well; for example, the threat of punishment may deter people from engaging in the illegal act.[35] Nonetheless, a handful of studies have used changes in the law to argue in support of a normative expressive function.[36] One of the clearest is Patricia Funk's study of Swiss voting laws.[37] Funk analyzed voter turnout in several Swiss cantons during the last half of the twentieth century. During that period, five of the cantons repealed long-standing mandatory voting laws that had been accompanied by fines that Funk called "symbolic"—the fines varied from canton to canton but were usually equal to about $1.00 (US) or less. Funk's study found that repeals decreased voter turnout in those cantons by 6 to 10%.[38] Because the punishment was so small, Funk argues that this is support for an expressive theory of law: people's behavior seemed to be influenced by the mere presence of the law even in the absence

29. Robert Cooter, *Expressive Law And Economics*, 27 J. LEGAL STUD. 585, 585 (1998); Cooter, *supra* note 16, at 1601; Funk, *supra* note 16, at 137; Sunstein, *supra* note 11, at 2025.

30. *See generally* Brian Borsari & Kate B. Carey, *Descriptive and Injunctive Norms in College Drinking: A Meta-Analytic Integration*, 64 J. STUD. ON ALCOHOL 331 (2003); Robert B. Cialdini, *Descriptive Social Norms as Underappreciated Sources of Social Control*, 72 PSYCHOMETRIKA 263 (2007); Robert B. Cialdini et al., *Managing Social Norms for Persuasive Impact*, 1 SOC. INFLUENCE 3 (2006); Harold L. Cole, George J. Mailath & Andrew Postlewaite, *Social Norms, Savings Behavior, and Growth*, 100 J. POLIT. ECON. 1092 (1992); Alan S. Gerber & Todd Rogers, *Descriptive Social Norms and Motivation to Vote: Everybody's Voting and so Should You*, 71 J. POLIT. 178 (2009); F. Marijn Stok et al., *Don't Tell Me What I Should Do, But What Others Do: The Influence of Descriptive and Injunctive Peer Norms on Fruit Consumption in Adolescents*, 19 BRIT. J. HEALTH PSYCHOL. 52 (2014); Stanley Milgram, Leonard Bickman & Lawrence Berkowitz, *Note on the Drawing Power of Crowds of Different Size*, 13 J. PERSONALITY SOC. PSYCHOL. 79 (1969).

31. Patrick De Pelsmacker & Wim Janssens, *The Effect of Norms, Attitudes and Habits on Speeding Behavior: Scale Development and Model Building and Estimation*, 39 ACCIDENT ANALYSIS PREVENTION 6, 6, 13 (2007).

32. Hunt Allcott, *Social Norms and Energy Conservation*, 95 J. PUB. ECON. 1082, 1083 (2011).

33. *See generally* H. Wesley Perkins, *Social Norms and the Prevention of Alcohol Misuse in Collegiate Contexts.*, J. STUD. ON ALCOHOL. 164, 164 (2002); *see also* Borsari & Carey, *supra* note 30, at 331.

34. Robert B. Cialdini, Carl A. Kallgren & Raymond R. Reno, *A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior*, 24 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 201, 203 (1991); Ernst Fehr & Urs Fischbacher, *Social Norms and Human Cooperation*, 8 TRENDS COGNITIVE SCI. 185, 185 (2004); Milgram et al., *supra* note 30, at 79.

35. Funk, *supra* note 16, at 156.

36. *See supra* note 29 and accompanying text.

37. *See* Funk, *supra* note 16, at 135.

38. *Id.* at 138.

of meaningful punishment, suggesting that voters were not simply deterred from defecting out of fear of punishment.[39]

Similar studies have documented significant increases in compliance following the adoption of seatbelt laws, dog waste ordinances, and smoking bans, even when the penalty for violating the laws is minimal.[40] Although these natural experiments generally reveal only the end points of the process—that is, a change in law leads to changes in behavior—researchers often claim (or assume) that the mechanism behind this behavioral change is the expression of social norms.[41]

Of course, laws do carry normative weight. At the very least, legal prohibitions convey injunctive norms against the prohibited actions. For example, a law against speeding suggests that—at the very least—the legislature believes one should not speed. But formalizing a social norm through punishment can also lead to unexpected and counterintuitive changes in behavior. In a notable field study, for example, Gneezy and Rustichini introduced a new punishment for late parents at some Israeli day care centers.[42] Parents signed a contract at the beginning of the school year in which they agreed to pick their children up on time, but prior to the study, no enforcement mechanism was specified for the rule.[43] After measuring the number of late parents for four weeks, the experimenters introduced a financial punishment for being late at some of the day cares in the study.[44] The punishment was relatively small—just ten shekels (worth approximately $2.72 US at the time of the study) per child if the parent was more than ten minutes late.[45]

The introduction of the monetary punishment did change parental behavior at the day cares in the test group,[46] but the effect was surprising. Rather than decreasing lateness at the selected centers, the fines caused a steady increase in lateness.[47] After twelve weeks, day care centers where the fine had been introduced reported a near doubling of the number of late parents, and removing the fine at the end of the study did nothing to reduce this new, higher rate of lateness.[48] The experimenters argued that the introduction of the fine was equivalent to setting a price for late pickup: rather than deter late parents, the (small) fine changed the prevailing social norm from one of obligation ("parents should pick

39. *Id*. at 155.

40. Robert D. Cooter, *Three Effects of Social Norms on Law: Expression, Deterrence, and Internalization*, 79 OR. L.REV. 1, 10–11 (2000); Cooter, *supra* note 29, at 595; Dhammika Dharmapala & Richard H. McAdams, *The Condorcet Jury Theorem and the Expressive Function of Law: A Theory of Informative Law*, 5 AM. L. ECON. REV. 1, 4–5 (2003); Wittlin, *supra* note 16, at 422.

41. *See* Patricia Funk, *On the Effective Use of Stigma as a Crime-Deterrent*, 48 EUR. ECON. REV. 715, 717 (2004); Funk, *supra* note 16, at 137; Richard H. McAdams & Janice Nadler, *Testing the Focal Point Theory of Legal Compliance: The Effect of Third-Party Expression in an Experimental Hawk/Dove Game*, 2 J. EMPIRICAL L. STUD. 87, 88 (2005); Cooter, *supra* note 29, at 607; Wittlin, *supra* note 16, at 420.

42. Uri Gneezy & Aldo Rustichini, *A Fine Is a Price*, 29 J. LEGAL STUD. 1, 1 (2000).

43. *Id.* at 4.

44. *Id.*

45. *Id.* at 4–5.

46. *Id.* at 7.

47. *Id.* at 3.

48. *Id.* at 7.

up their children on time") to one of transaction ("parents can pay to pick their children up late").[49]

As the Gneezy and Rustichini study demonstrates, the surface-level similarities between the effects of social norms and enforced laws on behavior may conceal deeper psychological differences. Moreover, punishment can signal a meaningful shift in the nature of the underlying act that colors subsequent behavior.[50] This shift could be one from a social cooperation dynamic to a transactional dynamic—as occurred in the day care centers—but it could also be another kind of shift, such as one from a norm to a moral imperative.

C. Moral Psychology and Punishment

In contrast to the research noted above, which tends to treat punishment as a simple enforcement mechanism for social norms, philosophical approaches often emphasize the distinctly moral component of punishment.[51] Indeed, many legal scholars characterize the message of punishment—rather vaguely—as moral condemnation.[52] Dan Kahan argues, for example, that "[p]unishment . . . is a special social convention that signifies moral condemnation."[53] Although moral psychology has not yet approached our question directly, that literature provides many important connections between punishment and moral judgment that may be particularly relevant to understanding what, exactly, punishment signals.[54] Indeed, amid the vast body of research on moral judgments, one link emerges repeatedly: the link between punishment and harm.

Harm is the central feature of retributive theories of punishment.[55] Under a retributive view, punishment is morally justified—indeed, morally required—to balance the harm done by the offender.[56] In contrast to consequentialist or utilitarian theories of punishment, which advocate punishment only to stem the future risk posed by an offender, retributivism is concerned primarily (or, in the extreme,

49. *Id.* at 13–14.

50. *Id.*

51. *See* MURPHY & HAMPTON, *supra* note 2, at 4; Kenworthey Bilz, *We Don't Want to Hear It: Psychology, Literature and the Narrative Model of Judging*, 2010 U. ILL. L. REV. 429, 429 (2010); Gert et al., *supra* note 12, at 79; Hampton, *supra* note 10, at 1659; Dan M. Kahan, *Social Influence, Social Meaning, and Deterrence*, 83 VA. L. REV. 349, 358 (1997); Kahan, *supra* note 11, 652.

52. *See, e.g.*, Bilz, *supra* note 51, at 484; Kahan, *supra* note 11, at 593; Kahan, *supra* note 51, at 383.

53. Kahan, *supra* note 11, at 593.

54. *See, e.g.*, *id.* at 592–93.

55. *See* Kenworthey Bilz & John M. Darley, *What's Wrong with Harmless Theories of Punishment*, 79 CHI. KENT. L. REV. 1215, 1232 (2004); Byrd, *supra* note 6, at 191; Darley & Pittman, *supra* note 1, at 325; Hampton, *supra* note 10, at 1663; Amrisha Vaish, Manuela Missana & Michael Tomasello, *Three-Year-Old Children Intervene in Third-Party Moral Transgressions.*, 29 BR. J. DEV. PSYCHOL. 124, 124–25.

56. *See* MURPHY & HAMPTON, *supra* note 2, at 111; Bilz & Darley, *supra* note 55, at 1232; Cahill, *supra* note 10, 818; Byrd, *supra* note 6, at 155, 191; Carlsmith et al., *supra* note 1, at 284; Hampton, *supra* note 10, at 1663; Ian R. McKee & N. T. Feather, *Revenge, Retribution, and Values: Social Attitudes and Punitive Sentencing*, 21 SOC. JUST. RES. 138 143–45 (2008).

exclusively) with evaluating the harm already caused and ensuring that perpetrators get what they deserve, even if this does not lead to better consequences.[57]

Research in psychology also demonstrates the close relationship between punishment and harm. Empirical studies designed to compare the degree to which people rely on implicit theories of retributivism or consequentialism have found that the degree of harm caused is one of the most important pieces of information to (mock) punishers.[58] Of course, moral psychology research often includes examinations of punishment outside of the retributivism versus consequentialism debate, and that research also supports the idea that punishment judgments are closely related to harmfulness judgments.[59] For example, studies of the "outcome bias" in moral psychology demonstrate that harm caused can even be more important for judging blame and assigning punishment than the wrongdoer's intent.[60] Even more tellingly, studies of so-called "moral luck" have demonstrated that when an act causes harm, judgments of punishment and blame are increased relative to judgments of the same action when it does not cause harm.[61] In contrast, judgments of moral character and the wrongness of the act itself do not seem to rely as much on whether harm was done—it seems that outcomes matter for harm and punishment more than they matter for wrongness (we will return to this issue in our later studies).[62]

Developmental research has also long recognized the connection between harm, immorality, and punishment. Developmental morality scholars have repeatedly demonstrated that children and adults distinguish between rules that they see as conventional (*i.e.*, social norms), those they see as prudential or safety-related, and those that they see as moral.[63] Whereas moral rules like

57. *See* Carlsmith, *supra* note 4, at 437–38; Darley et. al, *supra* note 7, at 660; Paul H Robinson, *Competing Conceptions of Modern Desert: Vengeful, Deontological, and Empirical*, 67 CAMB. L.J. 145, 147–48 (2008); Robinson & Darley, *supra* note 7, at 454.

58. *See* Kevin M. Carlsmith, *On Justifying Punishment: The Discrepancy Between Words and Actions*, 21 SOC. JUST. RES. 119, 133 (2008); Carlsmith, *supra* note 4, at 438; Darley et al., *supra* note 7, at 668; Darley & Pittman, *supra* note 1, at 326.

59. Darley & Pittman, *supra* note 1, at 325.

60. *See* Fiery Cushman et al., *The Development of Intent-Based Moral Judgment*, 127 COGNITION 6, 15 (2013); Cushman et al., *supra* note 1, at 5; Norman J. Finkel, *But It's Not Fair! Commonsense Notions of Unfairness.*, 6 PSYCHOL. PUB. POLY & L. 898, 930–47 (2000); Francesca Gino, Don A. Moore & Max H. Bazerman, *No Harm, No Foul: The Outcome Bias in Ethical Judgments*, HARV. BUS. SCH. 1, 3–4 (2009).

61. Fiery Cushman, *Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment*, 108 COGNITION 353, 354 (2008); Justin W. Martin & Fiery Cushman, *The Adaptive Logic of Moral Luck*, COMPANION TO EXP. PHILOS. 190, 190 (2016).

62. *See infra* Part IV.

63. Richard A. Shweder, Elliot Turiel & Nancy C. Much, *The Moral Intuitions of the Child*, *in* SOCIAL COGNITIVE DEVELOPMENT: FRONTIERS AND POSSIBLE FUTURES 288 (John H. Flavell & Lee Ross eds., 1981); Alicia Ardila-Rey & Melanie Killen, *Middle Class Colombian Children's Evaluations of Personal, Moral, and Social-Conventional Interactions in the Classroom*, 25 INT. J. BEHAV. DEV. 246, 253 (2001); Cameron B. Richardson, Kelly Lynn Mulvey & Melanie Killen, *Extending Social Domain Theory with a Process-Based Account of Moral Judgments*, 55 HUM. DEV. 4, 4 (2012); Judith G. Smetana, *Reasoning in the Personal and Moral Domains: Adolescent and Young Adult Women's DecisionMmaking Regarding Abortion*, 2 J. APPL. DEV. PSYCHOL. 211, 212, 224 (1981); Judith G. Smetana, Melanie Killen & Elliot Turiel, *Children's Reasoning about Interpersonal and Moral Conflicts*, 62 CHILD DEV. 629, 639–43 (1991); Marie S. Tisak & Elliot Turiel, *Children's Conceptions of Moral and Prudential Rules*, 55 CHILD DEV. 1030, 1030 (1984); Elliot Turiel, *Social Regulations and Domains of Social Concepts*, 1978 NEW DIRECTIONS FOR CHILD DEV. 45, 46, 50 (1978).

"don't hit" are seen as universal and immutable, even fairly young toddlers are more flexible when it comes to rules based in social convention, like "don't wear pajamas to school."[64] A key distinction between rules that are perceived as moral and those that are perceived as conventional seems to be that the former—but not the latter—involve harm done to another person or creature.[65] In short, research indicates that moral transgressions—that is, acts that harm others—*demand* punishment, even if the surrounding social conventions are changed.[66]

The previous research makes it clear that harm leads to increased punishment, but we do not know if punishment leads people to infer that an action in question is harmful.[67] Are there other candidates for what punishment could signal about an action?

Despite the early theories of morality in social and developmental psychology that tended to treat moral violations as fairly homogenous and harm-based, many contemporary theories adopt a broader approach.[68] Moral foundations theory, for example, identifies several underlying themes, in addition to harm, that may help to explain why moral violations are perceived as moral in the first place.[69] In response to harm-centric theories of morality, Jonathan Haidt and others point to apparently harmless scenarios, such as a case of consensual incest with no negative consequences for either party.[70] That such scenarios provoke a negative moral reaction has been used to argue in favor of a "purity" or "sanctity" domain of morality.[71] Analogous hypotheticals led to Haidt's first categorizations of five moral foundations: harm, fairness, loyalty, authority, and purity.[72]

64. Ardila-Rey & Killen, *supra* note 63, at 248–54; Charles W. Kalish & Mark A Sabbagh, *Conventionality and Cognitive Development: Learning to Think the Right Way*, NEW DIR. CHILD ADOLESC. DEV. 1, 6–8 (2007); Kristin D. Neff & Charles C. Helwig, *A Constructivist Approach to Understanding the Development of Reasoning About Rights and Authority Within Cultural Contexts*, 17 COGN. DEV. 1429, 1430–31 (2002); Jared Piazza et al., *Authority Dependence and Judgments of Utilitarian Harm.*, 128 COGNITION 261, 261–62 (2013); Judith G. Smetana et al., *Preschool Children's Judgments about Hypothetical and Actual Transgressions*, 64 CHILD DEV. 202, 203 (1993); Tisak & Turiel, *supra* note 63, at 1037; Xin Zhao & Tamar Kushnir, *Young Children Consider Individual Authority and Collective Agreement When Deciding Who Can Change Rules*, 165 J. EXP. CHILD PSYCHOL. 101, 103 (2017).

65. Daniel Kelly et al., *Harm, Affect, and the Moral/Conventional Distinction*, 22 MIND LANG. 117, 121 (2007); Tisak & Turiel, *supra* note 63, at 1030; Philip David Zelazo et al., *Intention, Act, and Outcome in Behavioral Prediction and Moral Judgment*, 67 CHILD DEV. 2478, 2478 (1996).

66. Zelazo et al., *supra* note 65, at 2478–79.

67. *See supra notes* 64 and 65 and accompanying text.

68. James A. Dungan, Alek Chakroff & Liane Young, *The Relevance of Moral Norms in Distinct Relational Contexts: Purity Versus Harm Norms Regulate Self-Directed Actions*, 12 PLOS ONE 1, 1–2 (2017).

69. Kurt Gray & Jonathan E Keeney, *Disconfirming Moral Foundations Theory on Its Own Terms: Reply to Graham (2015)*, 6 SOC. PSYCHOL. PERSONALITY SCI. 874, 874 (2015).

70. Jonathan Haidt, *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*, 108 PSYCHOL. REV. 814, 814 (2001).

71. Dungan et al., *supra* note 68, at 1.

72. Dungan et al., *supra* note 68, at 11; Jonathan Haidt, *The New Synthesis in Moral Psychology.*, 316 SCIENCE 998, 1001 (2007); Jonathan Haidt, *Morality*, 3 PERSPECT. PSYCHOL. SCI. 65, 70 (2008).

In addition to generating some of the most memorable hypotheticals for researchers,[73] purity violations are often cited as a rebuttal to critics of moral foundations theory.[74]

The purity (or "sanctity") domain also attracts attention because it has the clearest connection to a specific emotional response—namely disgust. Numerous studies have linked disgust reactions to moral judgments,[75] and even incidental feelings of disgust (such as those caused by a foul smell in the experiment room) can increase the harshness of moral evaluations and the desire to punish, especially for perceived violations in the purity/sanctity domain.[76] Although these scenarios strongly minimize or eliminate obvious harms, they are nonetheless viewed as *morally* wrong (and therefore deserving of punishment) by participants.[77]

It is worth noting that critics of moral foundations theory, the most prominent of whom argue that harm can adequately explain moral judgments without the need for other foundations,[78] have responded with a variety of explanations. Psychologist Kurt Gray has argued that these apparently harmless violations are not really perceived as harmless at all.[79] Instead, Gray argues, subjective harm is imputed even when the scenarios are written to foreclose the possibility of objective harm.[80]

The present project is not designed to resolve the debate between moral foundations theory and its critics by adjudicating whether morality is driven solely by harm or by other concerns beyond harm, or even to address it directly, though we discuss some possible implications of this research on this debate in the general discussion. However, the close association between harm and punishment led us to predict that punishment would communicate messages of harm

73. *See, e.g.*, Haidt, *supra* note 72, at 999; Kurt Gray, Chelsea Schein & Adrian F. Ward, *The Myth of Harmless Wrongs in Moral Cognition: Automatic Dyadic Completion from Sin to Suffering.*, 143 J. EXP. PSYCHOL. GEN. 1600, 1600 (2014); Kelly et al., *supra* note 65, at 123; Cass R. Sunstein, *Moral Heuristics.*, 28 BEHAV. BRAIN SCI. 531 540–42 (2003).

74. Dungan et al., *supra* note 68, at 11–13.

75. Beatrice H. Capestany & Lasana T. Harris, *Disgust and Biological Descriptions Bias Logical Reasoning During Legal Decision-Making*, 9 SOC. NEUROSCIENCE 265, 266 (2014); Yoel Inbar & David Pizarro, *Grime and Punishment: How Disgust Influences Moral, Social, and Legal Judgments*, 21 JURY EXPERT 11, 14 (2009); Jorge Moll et al., *The Moral Affiliations of Disgust: A Functional MRI Study*, 18 COGNITIVE & BEHAV. NEUROLOGY 68, 69 (2005); David Pizarro, Yoel Inbar & Chelsea Helion, *On Disgust and Moral Judgment*, 3 EMOTION REV. 267, 267 (2011); Joshua Rottman & Deborah Kelemen, *Aliens Behaving Badly: Children's Acquisition of Novel Purity-Based Morals*, 124 COGNITION 356, 357 (2012); Jessica M. Salerno & Liana C. Peter-Hagene, *The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments*, 24 PSYCHOL. SCI. 2069, 2069–70 (2013); Simone Schnall et al., *Disgust as Embodied Moral Judgment*, 34 PERSONALITY & SOC. PSYCHOL. BULL. 1096, 1105 (2008).

76. Pizarro et al., *supra* note 75, at 268.

77. *See id.*

78. *See*, *e.g.*, Kurt Gray, *Harm Concerns Predict Moral Judgments of Suicide: Comment on Rottman, Kelemen and Young (2014)*, 133 COGNITION 329, 329 (2014); Kurt Gray & Jonathan E Keeney, *Disconfirming Moral Foundations Theory on Its Own Terms: Reply to Graham (2015)*, 6 SOC. PSYCHOL. & PERSONALITY SCI. 874, 874 (2015); Gray et al., *supra* note 73, at 1600–01; Chelsea Schein, Ryan S. Ritter & Kurt Gray, *Harm Mediates the Disgust-Immorality Link.*, 16 EMOTION 862, 871 (2016).

79. *See* Gray et al., *supra* note 73, at 1601.

80. *Id.* at 1609.

particularly well, and the current debate in moral psychology provides us with an interesting alternative possibility. Perhaps the apparent relationship between harm and punishment is not so unique, but instead is an artifact, provoked by researchers who treat harmfulness as synonymous with morality.[81] In that case, the disgustingness (*i.e.*, the lack of purity) of an action might also be communicated by punishment information. Indeed, this possibility also finds support in the literatures of psychology and law.[82] As noted above, disgust can increase the harshness of moral judgments; feelings of disgust have also been associated with more frequent and more severe punishment in vignette studies, mock juries, and economic games.[83]

III. EXPERIMENTAL EVIDENCE

Drawing on the literatures discussed above, we set out to look for evidence of what messages people actually receive from learning about punishment. We adopt a simple experimental paradigm in which participants were told about a novel action. In the first condition of both studies, participants are told only that the novel act is or is not punished. Participants are then asked to rate the action on several dimensions. To test whether punishment signals moral condemnation, for example, participants are asked to judge the "moral wrongness" of the action.

Across Studies 1 and 2, we compare normative information to punishment information on three dimensions that are suggested by our review of the literature: moral wrongness, harmfulness, and disgust.[84] In Studies 3 and 4, we extend our findings from controlled, but artificial, alien actions to familiar, but messier real-world actions.[85]

A. Study 1: Punishment, Harm, and Moral Wrongness

Study 1 tests whether information about punishment leads people to make inferences about the moral status of an action and, if so, whether those inferences are specific to a particular moral dimension, such as wrongness or harm. Although it is not obvious that people will make any inferences, especially in such a simplified and artificial context, even if they do, such an inference is not very informative without additional comparisons. Is there anything special about punishment, or would any information about others' negative reactions give rise to the same inferences? To address this issue, our paradigm compares punishment with normative information, which we operationalized as telling participants that an action either causes or does not cause the actor to be disliked by others. By

81. *See* Schein, Ritter & Gray, *supra* note 78, at 871.

82. *See supra* Section II.C.

83. Capestany & Harris, *supra* note 75, at 267; Inbar & Pizarro, *supra* note 75, at 16–17; Bunmi O. Olatunji et al., *Who Am I to Judge? Self-Disgust Predicts Less Punishment of Severe Transgressions*, 12 EMOTION 169, 169 (2012).

84. Study 1: Punishment, Harm, and Moral Wrongness (on file with author).

85. Study 3: Inferences of Harm in the Real World & Study 4: Inferences about Punishment in the Real World (on file with author).

focusing the normative information on the actor, we can keep the information in the "dislike" conditions parallel to the information in the punishment conditions.

In Study 1, participants in all conditions were first introduced to the novel actions "blicking" and "gomping." Participants then received limited information about each action; the type of information varied by condition, as described in more detail below. In the punishment information condition, participants were told that one action was generally punished and one was generally not punished. In the normative information condition, participants were told that one action generally caused the actor to be disliked, and the other action did not. The third condition—the conflicting information condition—pitted the punishment and normative information against each other. One action is described as punished but not likely to cause dislike of the actor, whereas the other action is described as *not* punished but generally causing dislike. This condition allows us to gauge whether punishment of the action or dislike of the action is a stronger signal of moral wrongness or harm.

1. Methods

Participants. Participants were 270 adults (100 female), ages 19-65 (M = 37.28, SD = 15.63), recruited from Amazon's Mechanical Turk ("MTurk") and paid for their participation.[86]

Procedure. Participants were randomly assigned to one of three conditions: punishment information (n = 91), normative information (n = 88), and conflicting information (n = 91).[87]

In all three conditions, participants first read very brief instructions in which they were told to imagine an alien planet populated with aliens. Participants were also told that on this planet, "some things are quite similar to Earth, and some things are quite different." To minimize the degree to which participants incorporated their pre-existing moral beliefs into their responses, we used nonce words to describe unknown and novel actions; participants were told that these were two things that people on Earth "do not do." One was called "blicking," and the other was called "gomping."

In the punishment information condition, participants learned only whether the actions were or were not punished; *i.e.*, they read that while blicking is punished, gomping is not. In the normative information condition, participants were told: "An alien who blicks another alien is generally disliked. An alien who gomps another alien is generally not disliked." Finally, in the conflicting information condition, participants received all of the information provided in the other two conditions: "An alien is disliked when she blicks another alien, but she

86. In all of the studies in this Article, MTurk workers were only allowed to view and complete the study if they had not already participated in a previous version. This exclusion was accomplished with the help of TurkPrime.com, a third-party platform that provides additional features for researchers using MTurk. *See generally* Leib Litman et al., *TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences*, 49 BEHAV. RES. METHODS 443 (2017) (dsicussing MTurk).

87. While we intended to recruit approximately ninety participants for each condition, we allowed the numbers to vary slightly due to random assignment.

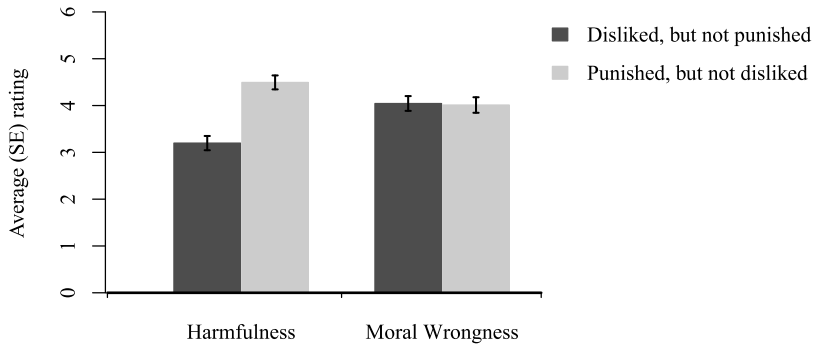
is generally not punished. An alien is not disliked when she gomps another alien, but she is generally punished."

After participants were given the condition-specific information about blicking and gomping, they were asked to choose which action was "morally worse." On a separate screen, participants also rated how morally "good or bad" they believed each action was on a scale ranging from "very bad" to "very good." Participants were also asked to rate how harmful each action was, and to choose which act was the more harmful. The moral wrongness questions were always presented together (though on separate screens), and the harmfulness questions were always presented together (though on separate screens), but the order of moral wrongness and harmfulness questions was randomized between participants.

2. Results

Punishment information condition. In the punishment information condition, 90% (n = 82) of participants reported that the act that was punished was more morally wrong than the act that was not punished; 3% (n = 3) responded that the non-punished act was more morally wrong, and 7% (n = 6) said the acts were about the same in terms of moral wrongness, $\chi^2(2, N = 91) = 132.15, p < .001$.[88] Participants in the punishment information condition also rated the punished act (M = 5.02, SD = 1.20) as significantly more morally wrong than the non-punished act (M = 2.54, SD = 1.28), $t_{paired}(90) = 12.55, p < .001$.

FIGURE 1. RATINGS OF HARMFULNESS AND MORAL WRONGNESS, CONFLICTING INFORMATION CONDITION



Note: Error bars represent standard error of the mean

Similarly, 92% (n = 84) of participants in the punishment information condition reported that the punished act was the more harmful. Two percent (n = 2) answered that the non-punished act was more harmful, and 6% (n = 5) said the acts

88. Unless otherwise noted, all reported chi-square results are chi-square tests for goodness of fit.

were about the same, $\chi^2(2, \mathrm{N} = 91) = 142.57, \mathrm{p} < .001$. Participants also rated the punished act (M = 5.16, SD = 1.09) as significantly more harmful than the non-punished act (M = 2.49, SD = 1.17), $t_{paired}(90) = 13.72, p < .001$.

Normative information condition. In the normative information condition, 83% (n = 73) of participants chose the disliked act as the more morally wrong. Another 2% (n = 2) responded that the act which does not cause the actor to be disliked is more morally wrong, and 15% (n = 13) said that they were about the same, $\chi^2(2, N = 88) = 99.57, p < .001$. Ratings of moral wrongness reflected a similar pattern. Participants rated the disliked act (*M* = 4.50, *SD* = 1.58) as significantly more morally wrong than the act that was not disliked (*M* = 2.40, *SD* = 1.44), $t_{paired}(87) = 10.16, p < .001$.

Again, responses about harmfulness were similar. The disliked act was chosen by 88% of participants (n = 77) as the more harmful, while 2% (n = 2) chose the other action and 10% (n = 9) said the actions were about the same in terms of harmfulness, $\chi^2(2, N = 88) = 117.02, p < .001$. Participants also rated the disliked act (*M* = 4.48, *SD* = 1.41) as significantly more harmful than the not disliked act (*M* = 2.27, *SD* = 1.25), $t_{paired}(87) = 10.16, p < .001$.

Conflicting information condition. When participants were given conflicting information about whether an act caused an actor to be disliked and whether an act was punished, 43% (n = 39) of participants chose the act that is disliked but not punished as the more wrong, and 42% (n = 38) chose the act that is punished but not disliked, while 15% (n = 14) responded that the acts were about the same in terms of moral wrongness, $\chi^2(2, N = 91) = 13.21, p = .001$. There was no significant difference in ratings of moral wrongness for the punished (but not disliked) act (*M* = 4.01, *SD* = 1.59) and the disliked (but not punished) act (*M* = 4.04, *SD* = 1.51), $t_{paired}(90) = 0.12, p = .91$.

In the same condition, however, 65% of participants (n = 59) said that the punished (but not disliked) action was more harmful than the disliked (but not punished) action. Of the rest, 19% of participants (n = 17) chose the disliked act as more harmful, and 17% (n = 15) responded that the two acts were about the same. The punished (but not disliked) action was also rated as significantly more harmful (*M* = 4.49, *SD* = 1.42) than the alternative (*M* = 3.20, *SD* = 1.45), $t_{paired}(90) = 5.75, p < .001$.

Discussion. Even in this minimal paradigm, with little context and no additional information, participants in the punishment information condition believed that a punished act was more harmful and less moral than a non-punished act. Our results also reaffirm that normative information—in the form of dislike—can act as a signal of harmfulness and moral wrongness: here too participants believed that a disliked action was more harmful and less moral than a non-disliked action.[89] As a first step, these results confirm a necessary assumption for the current research—*i.e.*, that people are willing to make inferences about an act based solely on information about whether it is punished.[90]

89. *See* discussion on Study 1 results, *infra* Section III.A.2.

90. This finding is also consistent with prior research demonstrating that children will use punishment as a cue to the moral "badness" of an act. *See* Bregant et al., *supra* note 18, at 712.

More importantly, however, when punishment information conflicted with information about what was disliked by others, participants regarded the punished action as more harmful but *not* more morally wrong than the disliked action. Thus, although punishment and dislike appear to be equally good at expressing that an action is morally wrong, punishment appears to be a better cue that an action is harmful. When asked about harm, the same participants who decline to distinguish between the wrongness of a punished act and a disliked one report that a punished (but not disliked) action is significantly more harmful than a disliked (but not punished) one. This reasoning is also robust to the type of question asked; participants made this distinction in both scaled ratings and forced choice responses.[91]

Although we take these findings to be evidence that punishment may contribute uniquely to judgments of harmfulness, another possibility is that punishment is simply a particularly intense variety of dislike or disapproval; when this dislike is strong enough, people assume an action is not only wrong, but also harmful. That is, punished actions are not different from disliked actions in kind, but only in degree. If this is true, then we should always find that punishment is taken as stronger evidence for the negative qualities of an action than is normative dislike. Our finding in Study 1 that punishment is not taken as stronger evidence of general moral wrongness casts some doubt on the simplest version of this explanation, but it is nonetheless possible that harmfulness—and not punishment—is the distinguishing factor. In other words, it could be that harmfulness judgments are especially sensitive to the degree of dislike or disapproval expressed, while moral wrongness judgments relatively insensitive, so that it is only harmfulness ratings that pick up the difference in degree between punishment and normative dislike. To test this alternative explanation, we can see whether punishment is also a stronger signal of moral concerns other than harmfulness. As we noted above, harmfulness is just one of several important psychological aspects of morality.[92] In Study 2, we turn to another important aspect: disgust.

B. Study 2: Harm and Disgust

Study 1 suggests that participants treat both punishment and dislike by others as a cue that an action is immoral, but that when the two types of information conflict (when one action is punished and the other is disliked), punishment is taken as a particularly strong indication that the action is harmful.[93] Participants interestingly think that both punished and disliked actions are equally morally wrong. At first glance, these results are puzzling. If harmfulness is an important component of moral wrongness, and punishment is a strong signal of moral

91. To corroborate these results, we conducted a separate replication of the conflicting information condition only; as in Study 1, participants in the replication were significantly more likely to answer that the punished but not disliked act was the more harmful act, but they were only marginally more likely to choose the punished but not disliked act as the more morally wrong.

92. *See infra* Section III.A.2.

93. *See infra* Section III.A.2.

wrongness, such that the punished act is more harmful than the disliked act, then why isn't a punished act also seen as more morally wrong than a disliked act? The answer, of course, could be that dislike communicates one or more different moral concerns more strongly than punishment does.

To test this possibility, we sought to identify a second moral dimension on which to compare punishment and dislike. As discussed above, the moral psychology literature has largely focused on two primary moral concerns in recent years: harm and purity.[94] Beyond this focal relevance, however, the literature provides some reasons to think that purity might be a good candidate. Though both harm and purity concerns are often moralized, researchers have demonstrated a number of striking differences between the two.[95] Brain imaging studies suggest that concerns about harm and purity may have significantly different neural origins,[96] be influenced by different situational and social factors,[97] lead to different emotional and behavioral reactions,[98] and ultimately lead to different inferences about the actors involved.[99]

Purity violations are often associated with feelings of disgust.[100] While the precise nature of the relationship between disgust and moral judgment is unclear, some speculate that moral disgust provides an incentive to reject and distance oneself from the moral offender, just as non-moral disgust prompts one to reject a potential contaminant.[101] Indeed, experiments have repeatedly demonstrated that people who feel disgusted will physically distance themselves from the source of the disgust.[102] Dungan, Chakroff, and Young argue that moral purity concerns may have evolved as a way of identifying group members whose behavior does not conform to group norms.[103] Thus, they note, although harm-based moral violations often seem to signal that an actor is a bad person, purity-based violations seem to signal instead that a person is a bad group member.[104]

In the current studies, our social normative information that an alien who blicks or gomps is disliked by other aliens essentially implies that aliens seek to put social distance between themselves and the offending alien. In other words, the social norm information we have provided may be signaling a moral concern more akin to disgust than to harm. Thus, in Study 2, we measure participant's inferences about the *disgustingness* of the underlying action in addition to its

94. Gray, *supra* note 78, at 1; Gray & Keeney, *supra* note 78, at 874; Haidt, *The New Synthesis in Moral Psychology*, *supra* note 72, at 999.

95. *See infra* notes 96–99 and accompanying text.

96. *See, e.g.*, *Moll* et al., *supra* note 75, at 69.

97. *See* Liane Young & Rebecca Saxe, *When Ignorance Is No Excuse: Different Roles for Intent Across Moral Domains*, 120 COGNITION 202, 202 (2011).

98. Catherine Molho et al., *Disgust and Anger Relate to Different Aggressive Responses to Moral Violations*, 28 PSYCHOL. SCI. 609, 609 (2017); Paul Rozin et al., *From Oral to Moral*, 323 SCI. 1179, 1179–80 (2009).

99. *See generally* Dungan et al., *supra* note 68.

100. *See, e.g.*, Pizarro et. al, *supra* note 75, at 267.

101. Paul Rozin & April E. Fallon, *A Perspective on Disgust*, 94 PSYCHOL. REV. 23, 39 (1987); Rozin et al., *supra* note 98, at 1180.

102. For a review, see Rozin & Fallon, *supra* note 101.

103. Dungan et al., *supra* note 68, at 11.

104. *See id.* at 11–12.

harmfulness. We predict that we will again find that punishment will be seen as a better indication of harm than will dislike. But we also posit that dislike may be taken as better evidence than punishment that an action is disgusting.

1. Methods

The paradigm for Study 2 was substantially identical to the paradigm used in Study 1, with changes noted below.

Participants. Participants were 125 adults (61 female), ages 18-73 (M_{age} = 34.57, *SD* = 10.74), recruited from Amazon's Mechanical Turk and paid for their participation.

Procedure. As before, participants were randomly assigned to one of three conditions: normative information (n = 41), punishment information (n = 43), and conflicting information (n = 41).[105] The conditions were identical to those used in Study 1, such that participants in the conflicting information condition read that one of the alien acts causes the actor to be disliked but not punished and the other act is punished but does not cause the actor to be disliked. In contrast, participants in the normative information condition read that one act caused dislike and the other did not, whereas participants in the punishment information condition read that one act was punished and the other was not.

In Study 2, however, we added a new set of "disgust" measures. Participants still rated the harmfulness of each act (on a seven-point scale) and chose which was the more harmful, but then we asked participants to rate the degree to which each act was "disgusting" and to choose which act was the more disgusting (forced choice, including an option for "about the same").

2. Results

Punishment information condition. In the punishment information condition, participants overwhelmingly (73%, n = 30) reported that the punished act was more harmful than the non-punished act, $\chi^2(2, N = 43) = 68.98, p < .001$; of those who did not choose the punished act, 5% (n = 2) chose the non-punished act and 2% (n = 1) chose "about the same". On the scale response, participants also rated the punished act as significantly more harmful (*M* = 4.63, *SD* = 0.98) than the non-punished act (*M* = 1.28, *SD* = 1.32), $t(40) = 7.53, p < .001$.

When asked which was more disgusting, 74% of participants (n = 32) chose the punished act, 9% (n = 4) chose the non-punished act, and 16% (n = 7) chose "about the same" $\chi^2(2, N = 43) = 32.98, p < .001$. Participants also rated the punished act as significantly more disgusting on the scale measure (*M* = 3.53, *SD* = 1.65) than the non-punished act (*M* = 1.53, *SD* = 1.33), $t_{paired}(42) = 6.01, p < .001$.

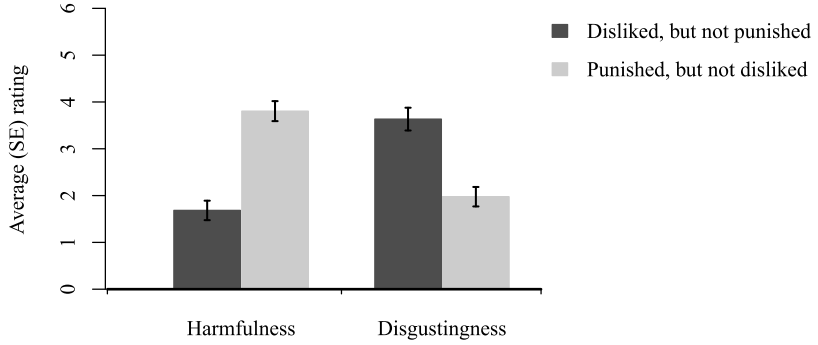
105. As in Study 1, the exact numbers in each condition were allowed to vary as a function of random assignment. We recruited 125 participants so that, even with this variation, each condition would have at least forty participants.

Normative information condition. Results in the normative information condition were also as predicted. On the forced-choice measure, 73% (n = 30) of participants chose the disliked act as the more harmful, 12% (n = 5) chose the non-punished act, and 15% (n = 6) said they were about the same, $\chi^2(2, N = 41) = 29.32, p < .001$. Participants also rated the disliked act as significantly more harmful (*M* = 3.63, *SD* = 1.48) than the non-punished act (*M* = 1.20, *SD* = 1.33), $t(40) = 7.53, p < .001$.

Similarly, 83% of participants (n = 34) chose the punished act as the more disgusting, 7% (n = 3) chose the non-punished act as the more disgusting, and 10% (n = 4) responded that the acts were about the same, $\chi^2(2, N = 43) = 45.41, p < .001$. On average, participants also rated the punished act as significantly more disgusting (*M* = 4.49, *SD* = 1.21) than the non-punished act (*M* = 1.37, *SD* = 1.26), $t_{paired}(40) = 10.38, p < .001$.

Conflicting information condition. When asked to compare an action that is punished but not disliked to an action that is disliked but not punished, 56% of

FIGURE 2. RATINGS OF HARMFULNESS AND MORAL WRONGNESS, CONFLICTING INFORMATION CONDITION



the participants in this condition (n = 23) chose the punished act as the more harmful of the two, $\chi^2(2, N = 43) = 9.56, p < .01$. The remaining participants split evenly between the other two choices: 22% (n = 9) chose the disliked act as the more harmful, and 22% (n = 9) responded that they were about the same. Participants also rated the punished act as significantly more harmful (*M* = 3.80, *SD* = 1.36) than the disliked act (*M* = 1.68, *SD* = 1.33), $t_{paired}(40) = 6.16, p < .001$.

However, the results were different for the disgust measures. A majority of the participants in this condition (63%, n = 26) chose the disliked act as the more disgusting of the two, $\chi^2(2, N = 41) = 17.02, p < .001$, while 15% (n = 6) said the punished act was more disgusting and 22% (n = 9) responded that they were about the same. Participants also rated the disliked act as significantly more disgusting (*M* = 3.63, *SD* = 1.56) than the punished act (*M* = 1.98, *SD* = 1.33), $t(40) = 4.62, p < .001$.

Discussion. This study replicates several key findings from Study 1. First, participants were again willing and able to make inferences about an action based

solely on knowing that the action was punished (in the punishment information condition) or that the action was disliked (in the normative information condition). Both pieces of information again caused participants to rate the actions as more harmful than the actions that were not punished or not disliked. The same held for an action that was punished but not disliked (in the conflicting information condition). As in Study 1, the punished act was viewed as more harmful than the non-punished but disliked act.

The results of Study 2 also suggest that punishment information is particularly informative about harm, and that dislike information is particularly informative about disgust. Consistent with Study 1, we again found that participants thought a punished action (that is not disliked) is more harmful than a disliked action (that is not punished). Dislike, though a weaker signal of harmfulness than competing punishment information, is a stronger signal of the disgustingness of an action than is punishment information. This fact may also help to explain why the two actions in the conflicting information condition were not seen as differing in terms of moral wrongness even though the punished action was seen as more harmful. If punishment has a relatively targeted effect on judgments of harm, and dislike has a similar effect on judgments of disgust, then the two effects may effectively cancel each other out in the broader moral judgment. Of course, neither punishment information nor normative information is necessarily limited to influencing a single moral domain; the results from the punishment information and normative information conditions show that both kinds of information are able to influence broad moral judgments in some circumstances. However, as we discussed at the outset, the theoretical landscape corroborates our argument that the harm-punishment relationship is special, and the results of Studies 1 and 2 further support this view.

Although Studies 1 and 2 have the advantage of simplicity, allowing us to inquire directly about the moral constructs we are interested in, we can draw only limited conclusions about how information about punishment may influence moral judgments in everyday life. If punishment information does indeed lead to increased inferences of harmfulness, then we should be able to see that effect outside the minimalistic alien worlds that we created for Studies 1 and 2. In Studies 3 and 4, we look for evidence of this effect in the real world, asking participants to rate realistic actions—described as being either punished or illegal but unpunished—in terms of their harmfulness. Of course, real world actions often carry with them pre-existing ideas about the morality and harmfulness of the action, as well as increased noise from social context. Nonetheless, we predicted that participants would view actions as more harmful when they were led to believe the actions were punished than when they were not.

C. Study 3: Inferences of Harm in the Real World

Studies 1 and 2 provide evidence that people will infer the harmfulness of a novel action if they learn that it is punished. These results are interesting from a psychological perspective, but the artificial nature of the scenario used in the studies leaves open the question of whether people make these inferences in the

"real world." In other words, although punishment may lead to inferences of harm (in particular) when no other information about the action or the world in which it occurred is known, we do not yet know whether this carries over into non-novel acts. Study 3 tests for inferences of harm on real-world actions.

1. Methods

In this study, we presented participants with two real-world actions, one of which we claimed was generally punished and the other we claimed was generally not punished. We then asked participants to rate the harmfulness of each act. If, as Studies 1–3 suggest, people infer that a punished act is more harmful than a non-punished act, then participants' ratings of the harmfulness of each act could change, depending on the (purported) presence or absence of punishment.

Participants. Participants were 161 adults (70 female), ages 19 to 72 (M = 34.21, SD = 10.21) recruited from Amazon's Mechanical Turk and paid for their participation.

Design and Procedure. Participants were asked to evaluate the harmfulness of two ostensibly illegal acts: (1) "Bringing firewood from another part of the country into a state park" and (2) "Gambling on professional sporting events (outside a licensed casino or gambling facility)." They were told that the items were drawn from a larger pool of items that were "illegal in most places," but whose enforcement varied. In fact, we pre-tested thirty-eight items to obtain pre-existing beliefs about the harmfulness of each action, as well as pre-existing beliefs about whether the action "should be illegal." We then selected two items that had average and modal ratings near the neutral point of the scale; *i.e.*, these items were chosen because the pretest ratings suggested the harm they cause was ambiguous.

Participants were then randomly assigned to one of two conditions. In one condition, they were told that the firewood action was "punished in most places," and in the other condition participants were told that the firewood action was "not punished in most places." Each participant was given the opposite punishment information for the gambling action; thus, each participant was told that one of the acts was generally punished and one of the acts was generally not punished. The order of the acts themselves was randomized across participants.

For each of the two acts, participants were asked to rate how harmful the act was by moving a slider along a scale marked "Not at all harmful" at one end (coded as 0) and "Extremely harmful" at the other end (coded as 100). The coded numerical value out of 100 was computed by the survey software and not displayed to participants.

2. Results

Across both actions, participants rated the punished action as more harmful than the non-punished action, $M_{\text{punished}} = 38.31$, $M_{\text{not punished}} = 28.72$, $t(320.98) = 3.05$, $p < .01$. However, participants' ratings of the individual actions varied. Participants rated the firewood action as significantly more harmful when they

were told it was punished than when they were told it was generally not punished, $M_{\text{punished}} = 45.46$, $M_{\text{not punished}} = 29.25$, $t(158.68) = 3.47$, $p < .001$. Participants did not rate the gambling action as significantly more harmful when told that it was punished, $M_{\text{punished}} = 31.25$, $M_{\text{not punished}} = 28.35$, $t(158.98) = 0.71$, $p = .48$.

Discussion. These results, though not conclusive, suggest that information about punishment can influence participants' inferences about the harmfulness of an action in the real world. Participants rated transporting firewood as more harmful when they believed the act was punished, although that difference did not occur in the gambling action. Taken together with the results of Studies 1 and 2, this is further evidence that punishment can convey unique information about the harmfulness of an act, both in abstract cases and in familiar actions.

However, caution is warranted in interpreting these results. Although participants rated transporting firewood as more harmful when they believed it to be punished, the lack of a difference for gambling is notable. We noted at the outset of this study that the real world was likely to be noisier than the artificial alien world used in Studies 1 and 2; the null result for gambling may reflect this additional noise and complexity. Moreover, prohibitions on gambling are undoubtedly more familiar to many participants than are prohibitions on transporting firewood. Prior to the study, participants may have had clearer ideas and preconceptions about gambling and its harmfulness.

Of course, it could also be that something about the firewood prohibition made it particularly susceptible to this effect. If that is the case, then our results would have very limited generalizability. A replication of the effect and a demonstration that it applies to more than just transporting firewood is necessary before making any further conclusions. In Study 4, we repeat this experiment using the firewood action and three other new actions. To ensure that this replication is transparent, we also pre-registered the planned data collection and analyses for Study 4.

D. Study 4: Inferences about Punishment in the Real World

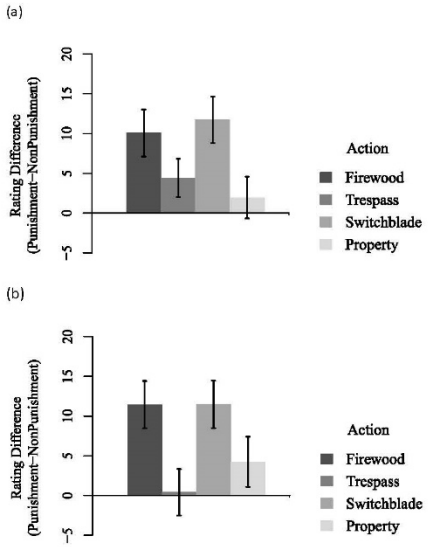
Study 3 provides some preliminary evidence that people judge even real-world actions as more harmful if they are punished. The design of that study, using just two real-world actions, has the advantage of simplicity, but the results are far from definitive. The effect of punishment information was consistent in direction, in that a punished act was rated as more harmful than the same act when not punished, but the difference was only significant for transporting firewood into a state park. As noted above, there are a number of possible explanations for this discrepancy. To address these possibilities, Study 4 examines a broader range of real-world actions and employs a larger sample in a pre-registered replication of Study 3. The purpose was two-fold: first, to replicate and confirm the effect of punishment information on harmfulness ratings for the firewood action and second, to better assess whether the effect is consistent across a range of actions.

1. Methods

Participants. Four hundred and four participants (149 female, 1 non-binary, 1 gender fluid), ages 19–77 (M = 34.87, SD = 11.18), recruited from Amazon's Mechanical Turk, participated in exchange for payment.

Procedure. All procedures and analyses for this study were preregistered on AsPredicted.org.[106] As in Study 3, participants were told that they would see a series of illegal actions, some of which were punished in most places and some of which were not. In fact, each participant saw the same four actions described. Participants were then told that two of the acts (assigned at random) were punished and the other two were not. Using a slider scale identical to the measure used in Study 3, participants then rated the harmfulness of the act and whether the action was morally wrong on a scale that was coded from 0 (Not at all harmful, Not at all morally wrong) to 100 (Extremely harmful, Extremely Morally Wrong). The four actions were: (1) taking home for personal use something your employer plans to throw away ("Employee act"); (2) carrying a switchblade knife ("Switchblade act"); (3) taking a shortcut through private property, where "no trespassing" signs are posted ("Trespass act"); and (4) bringing firewood from another part of the country into a state park ("Firewood act"). The order of the four actions was randomized for each participant.

FIGURE 3. DIFFERENCE BETWEEN PUNISHED AND NON-PUNISHED AVERAGES FOR (A) HARMFULNESS RATINGS AND (B) MORAL WRONGNESS



106. *Expressive Punishment–Large Study of Real Actions, July 2017 (#4886)*, ASPREDICTED (July 18, 2017, 12:35 PM), https://aspredicted.org/g5xb3.pdf.

2. Results

Overall, participants rated punished actions (M = 36.18, SD = 29.96) as significantly more harmful than non-punished actions (M = 29.14, SD = 26.64), $t(1592.2) = 4.99, p < .001$. Participants also rated the actions as more morally wrong when they believed the actions were punished (M = 40.25, SD = 31.70) than when the actions were described as not punished (M = 33.39, SD = 29.76), $t(1607.6) = 4.48, p < .001$. For each individual action, we also conducted an ANOVA to compare the ratings of participants who read that the act was punished to the ratings of those who were not punished.

Employee act. For taking home an employer's discarded property for personal use, there was no significant difference between the harmfulness ratings of participants who were told the action was generally punished (M = 23.32, SD = 26.88) and those who were told the action was generally not punished (M = 21.36, SD = 25.62), $F(1,402) = 0.561, p = .45$. Similarly, the acts were not rated differently on moral wrongness, (M_{punished} = 34.67, SD = 32.63; $M_{\text{not-punished}}$ = 30.43, SD = 30.96), $F(1,402) = 1.79, p = .18$.

Switchblade act. Participants rated carrying a switchblade as significantly more harmful when told that doing so was generally punished (M = 46.57, SD = 31.15) than when they were told it was generally not punished (M = 34.86, SD = 27.26), $F(1,402) = 16.16, p < .001$. Participants also rated the punished version as more morally wrong (M = 40.66, SD = 32.70) than the non-punished version (M =29.18, SD = 27.02), $F(1,402) = 14.76, p < .001$.

Trespass act. Trespassing through private property was rated as marginally more harmful when it was described as punished (M = 32.73, SD = 26.09) than when it was described as not punished (M = 28.31, SD = 22.33), $F(1,402) = 3.35, p = .07$. However, there was no significant difference between ratings of wrongfulness between the punished (M = 45.33, SD = 29.67) and non-punished (M = 44.92, SD = 29.25) versions, $F(1,402) = 0.02, p = .89$.

Firewood act. As in Study 3, participants in Study 4 rated transporting firewood across state lines as significantly more harmful when they believed such transportation was punished (M = 42.10, SD = 30.11) than when they believed it was generally not punished (M = 32.03, SD = 29.13), $F(1,402) = 11.66, p < .001$. They also rated the punished version as more morally wrong (M = 40.41, SD = 30.99) than the non-punished version (M = 28.97, SD = 28.81), $F(1,402) = 14.76, p < .001$.

Discussion. Pooling across all of the actions we studied, we found a main effect of punishment such that the punished action is seen as more harmful and also more wrong. Indeed, for all four acts used in this study, participants rated them as directionally more harmful and more morally wrong when they were described as being punished than when they were described as not punished, even though the actions were described as being illegal in all cases. But it was clear that the effect of punishment was stronger in some cases than others. With respect to harm, this difference was statistically significant for the switchblade act

and the firewood act and marginally significant for the trespass act. In comparison, the difference in moral wrongness was significant only for firewood and the switchblade act.

Taken together with the results of Study 3, these data demonstrate the power of punishment to communicate information about morality, and especially about harmfulness. As in Study 3, when participants were led to believe that an action is punished, they rated the action as consistently more harmful, at least for a subset of the actions we tested. We replicated the effect of punishment information on harmfulness judgments for transporting firewood, but also found that the effect holds for carrying a switchblade knife and, to a lesser extent, trespassing on private property.

It is worth noting that in both Study 3 and Study 4, the significant differences were much smaller than the differences we found in Studies 1 and 2. As we predicted, real world actions are not as clean as the novel actions we created in the first two studies, and participants likely brought to the latter studies their own ideas about many of the real acts. This additional noise underscores the value of the highly stylized methods we used in the earlier studies, however. Those studies allowed us to precisely explore the effect of punishment information in isolation, which in turn informed the predictions we made in the more realistic studies.

IV. GENERAL DISCUSSION

Across four studies, we find that people use information about punishment to make meaningful inferences about the punished act. In particular, our results show that punished acts are viewed as more harmful than identical actions that are not punished. Our results not only provide strong psychological support for expressive and communicative theories of punishment, but also add an important new component to our understanding of such theories by shedding light on the *content* of punishment's expressive message. In our studies, harm seems to be the strongest message of punishment, but it is not the only message; in the absence of other information, people also infer that a punished act is more morally wrong and more disgusting than an act that is not punished. Overall, these findings suggest that punishment can serve as an important psychological cue. In this Section, we first review the key findings from our empirical studies and then discuss how those findings may inform law and policy and increase our understanding of moral and legal psychology.

In Study 1, learning that an act is punished leads people to infer that it is more harmful and more morally wrong than an act that is not punished, even in a minimal and artificial context. When a non-punished action also causes the actor to be disliked, however, people do not make the same distinction between the two acts in terms of moral wrongness—both the punished but not disliked and the disliked but not punished actions are rated as equally morally wrong. However, participants do infer that the punished but not disliked act is more harmful than the disliked but non-punished act.

Study 2 confirmed that, in the absence of other information, participants will use the fact that an act is punished as a cue to harmfulness, but it also showed that participants will use the same information to infer that a punished act is more disgusting than a non-punished act. Study 2, however, also showed that the special contribution of punishment information to harmfulness judgments that we observed in Study 1 does not carry over onto all sub-domains of morality. When punishment and dislike information conflicted in Study 2, the punished action was still chosen as the more harmful, but the disliked (and not punished) action was chosen as the more disgusting. Thus, punishment information seems to lead to inferences that an action is harmful over and above any inferences that the action is morally wrong.

In Studies 3 and 4, we extended our findings into more real-world contexts, and we found that people will make the inference that a punished action is more harmful than a non-punished (but illegal) action. Although we found evidence of this inference in only some of the cases we tested, these studies nonetheless demonstrate that the inference is not limited to the barebones scenarios we used in Studies 1 and 2.

In both artificial and real-world contexts, punishment seems to lead people to make a number of meaningful inferences about the action that is being punished, at least when other cues are not available. In other words, punishment has informational value. This is consistent with prior research on the expressive function of law[107] and work finding that punishment can convey information about the victims of harm,[108] although, to our knowledge this is the first evidence that punishment also conveys nuanced information about the morality of the punished action.

Our results also provide an intriguing starting point for a broader discussion about the role of punishment in society. In law and policy, the inference that a non-punished act is somehow less harmful than a comparable punished act may have troubling consequences. When punishment varies in the real world, some crimes or victims of crimes may be perceived as more or less important, especially if the presence and absence of punishment is repeated or systematic.[109] Here, we highlight a few areas where such inferences may be of particular interest.

Following the 2008 financial crisis, many people took a renewed interest in the prosecution of corporate malfeasance.[110] The Securities and Exchange Commission ("SEC") and Department of Justice ("DOJ") investigated many allegations of criminal activity and breaches of trust on the part of financial institutions,[111] but the government also developed a number of somewhat unusual ways of dealing with the results of their investigations. Rather than pursuing civil or criminal suits against the (allegedly) offending institutions, the government

107. *See* Funk, *supra* note 16, at 136 n.2.

108. Bilz, *supra* note 1, at 1081; Bregant et al., *supra* note 18, at 699.

109. *See* Bregant et al., *supra* note 18, at 699.

110. *See, e.g.*, Jessica Bregant & Jennifer K. Robbennolt, *Neither Admit nor Deny*, 44 AM. PSYCHOL. ORG. 26 (2013).

111. *Id.*

reached agreements with them that allowed them to avoid official sanctions.[112] Although many SEC settlements required the institutions to submit to increased federal monitoring, pay fines, or both to avoid litigation, many also allowed the institutions to agree to such measures while still maintaining that they did nothing wrong.[113] The so-called "neither admit nor deny" statements came under heavy fire from the public and from judges, though the SEC maintained that they encouraged fast and efficient resolutions to important cases.[114] The DOJ has also created a number of ways for corporations to save face while still cooperating with government investigations and oversight; among the most notable is the deferred prosecution agreement ("DPA").[115] DPAs, like "neither admit nor deny" settlements, represent an agreement between a corporation or corporate employee and the government.[116] The former avoids a criminal prosecution (at least temporarily), and the latter gets to set terms—often quite stringent—to which the corporation must adhere if it is to remain unprosecuted.[117]

Even if the wide use of these non-punishment strategies has allowed the government to tighten corporate oversight and more directly control corporate affairs following malfeasance,[118] our results may suggest that the costs of these agreements could be more than previously believed. The idea that, as some have quipped, a financial institution may be "too big to jail,"[119] even when it is accused of serious wrongdoing, may resonate in the public psyche. When the government declines to prosecute or punish such an institution through a DPA, or when it imposes sanctions but allows those sanctions to be couched in terms that are not condemnation, through a "neither admit nor deny" settlement, the public perception of the institution's actions may change. Our results suggest, moreover, that the change in perception might be predictable: an act which is not punished is viewed as less harmful. In other words, the government's decision not to punish corporate wrongdoing could lead people to infer that the corporation's acts were less harmful than previously believed.[120]

112. *Id.*

113. *Id.*

114. *Id.*; Priyah Kaul, *Admit or Deny: A Call for Reform of the SEC's "Neither-Admit-Nor-Deny" Policy*, 48 UNIV. MICH. J.L. REFORM 535, 542 (2015).

115. Verity Winship & Jennifer K. Robbennolt, *An Empirical Study of Admissions in SEC Settlements*, 60 ARIZ. L. REV. 1, 6 (2018).

116. *See id.* at 28; Verity Winship & Jennifer K Robbennolt, *Admissions of Guilt in Civil Enforcement*, 102 MINN. L. REV. 1077, 1103 (2018).

117. *See generally* Donald L. Ferrin et al., *Silence Speaks Volumes: The Effectiveness of Reticence in Comparison to Apology and Denial for Responding to Integrity- and Competence-Based Trust Violations*, 92 J. APPL. PSYCHOL. 893 (2007); Kaul, *supra* note 114, at 555. This procedure is quite similar, at least conceptually, to criminal prosecutions of individuals in which the defendant pleads "no contest." *See* North Carolina v. Alford, 400 U.S. 25, 33–36 (1970) (discussing the "variety of different ways" courts have described nolo contendre pleas).

118. *See, e.g.*, Winship & Robbennolt, *supra* note 115, at 11.

119. Nizan Geslevich Packin, *Breaking Bad: Too-Big-to-Fail Banks Not Guilty as Not Charged*, 91 WASH. U. L. REV. 1089, 1092 (2014); Henry N. Pontell, William K. Black & Gilbert Geis, *Too Big to Fail, Too Powerful to Jail? On the Absence of Criminal Prosecutions After the 2008 Financial Meltdown*, 61 CRIME, L. SOC. & CHANGE 1 (2014).

120. Ferrin et al., *supra* note 117; Kaul, *supra* note 114.

We can only speculate about the further implications of such an inference, but one possibility is that the blame for such acts may be relocated. After all, the reasoning could go, if the actions of the corporations that led to the financial collapse were not actually as harmful as people believed, then perhaps the "real" blame lies more on the victims of the corporate actions (*e.g.*, "well, they should not have taken out mortgages they could not afford"). Indeed, some prior research demonstrates that failing to punish a wrongdoer can have negative consequences for how a victim is viewed.[121]

The effects of non-punishment on the perception of victims could be further exacerbated if the failure of punishment is systematically linked to certain victims or certain crimes. The Black Lives Matter Movement, for example, reflects a line of thinking that is consistent with our findings. When people perceive that violence by police officers goes unpunished, they may infer that the police action was less harmful—even if that action resulted in someone's death.[122] Thus, as activists argue, if officers are punished less often (or appear to be punished less often) when the injured party is black, it signals that injuring and killing black people is less harmful than injuring and killing non-black individuals.[123] These concerns also apply in other contexts. For example, the same logic can be applied to crimes against women, including domestic violence and sexual assault, which are often thought to be under-reported and under-punished.[124] Such crimes may be viewed as less harmful when they are not punished, which could in turn reflect poorly on victims and lead to even lower rates of reporting and punishment.

Of course, the broader context in which a given example of punishment or non-punishment occurs will be an important factor in how it is interpreted. In our studies, we state the presence or absence of punishment as a descriptive fact, *i.e.*, "[a]liens who blick are generally not punished", "[transporting firewood] is generally punished." In Studies 3 and 4, when we described apparently real criminal offenses, we told participants that all of the actions were illegal "in most places." We were careful not to give any explanation for why enforcement and punishment might vary. In contrast, when a high-profile case ends in punishment or non-punishment, the reasons likely matter a great deal to people and to the inferences they make. Very different inferences might arise when the underlying action is not punished because it is simply not illegal, because it is not reported, or

121. Bilz, *supra* note 2; Bregant, Shaw & Kinzler, *supra* note 18, at 707.

122. Of course, our findings show that such an effect would likely be small; unlike moving firewood across the country, most people have a fairly strong idea of how harmful a killing is.

123. Mark S. Brodin, *The Murder of Black Males in a World of Non-Accountability: The Surreal Trial of George Zimmerman for the Killing of Trayvon Martin*, 59 HOW. L.J. 765–786 (2016); Corinthia A. Carter, *Police Brutality, the Law & Today's Social Justice Movement: How the Lack of Police Accountability Has Fueled #Hashtag Activism*, 20 CUNY L. REV. 521, 525 (2017); Linda Sheryl Greene, *Before and after Michael Brown-Toward an End to Structural and Actual Violence*, 49 WASH. U. J.L. POL'Y 1, 4 (2015).

124. Michelle J. Anderson, *Campus Sexual Assault Adjudication and Resistance to Reform*, 125 YALE L.J. 1940, 1949 (2016); Susan Bandes, *Victim Standing*, UTAH L. REV. 331, 335 (1999); Christine E.W. Bond & Samantha Jeffries, *Similar Punishment?: Comparing Sentencing Outcomes in Domestic and Non-Domestic Violence Cases*, 54 BRIT. J. CRIMINOLOGY 849, 849 (2014); KJ Steinman, *Sex Tourism and the Child: Latin America's and the United States' Failure to Prosecute Sex Tourists*, 13 HASTINGS WOMENS L.J. 53, 67 (2002); Elisabeth Jean Wood & Nathaniel Toppelberg, *The Persistence of Sexual Assault Within the US Military*, 54 J. PEACE RES. 620, 621 (2017).

because it is not proven. The injustice, however, that people may feel after an instance of non-punishment may have far-reaching effects that go beyond the particular context at hand. That sense of injustice may lead to a kind of unintentional backlash: research shows that when a perceived wrongdoing goes unpunished, people's anger may lead them to act as "intuitive prosecutors," unconsciously transferring their anger and sense of injustice to future, *unrelated* transgressions.[125]

Though not our primary focus in this project, our results do add some interesting new information to the ongoing debate in moral psychology over the centrality of harm to moral judgments. As we alluded to above, there are debates about whether people truly find actions to be immoral in the absence of demonstrated harm.[126] Despite the vast body of research showing that information about harm influences punishment judgments,[127] to our knowledge this is the first to show the opposite is also true: information about punishment leads to increased inferences about harm. This relationship between punishment and harm adds further complexity to these ongoing debates. On one hand, this finding underscores the importance of harm judgments in moral reasoning, which may lend some support to the arguments that all moral judgments are, at their core, based on perceived harms.[128] On the other hand, these data do present something of a puzzle for an account in which condemnation and punishment are predicated on intuitions about harm. If, as our results demonstrate, this path can also be reversed, then the relationship between harm and punishment must be, at a minimum, bi-directional. Perhaps a kind of over-learning model[129] could account for this discrepancy, but more work is necessary to determine whether this can be squared with harm-only models of morality.

125. Julie H. Goldberg et al., *Rage and Reason: The Psychology of the Intuitive Prosecutor*, 29 EUR. J. SOC. PSYCHOL. 781, 783 (1999).

126. Gray & Keeney, *supra* note 78, at 875; *see* Haidt, *Morality, supra* note 72, at 69.

127. *See generally* Alek Chakroff et al., *Harming Ourselves and Defiling Others: What Determines a Moral Domain?*, 8 PLOS ONE e74434 (2013); Alek Chakroff et al., *From Impure to Harmful: Asymmetric Expectations About Immoral Agents*, 69 J. EXP. SOC. PSYCHOL. 201 (2017); Fiery Cushman et al., *The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm*, 17 PSYCHOL. SCI. 1082 (2006); Molly J. Crockett et al., *Harm to Others Outweighs Harm to Self in Moral Decision Making*, 111 PROC. NAT'L ACAD. SCI. 17320 (2014); Gino, Moore & Bazerman, *supra* note 60; Hampton, *supra* note 10; Charles C. Helwig, et al., *Children's Judgments of Psychological Harm in Normal and Noncanonical Situations*, 72 CHILD DEV. 66 (2001); Richard Murphy, *The Significance of Victim Harms: Booth v. Maryland and the Philosophy of Punishment in the Supreme Court*, 55 U. CHI. L. REV. 1303–1333 (1988); Stephen J. Schulhofer, *Harm and Punishment: A Critique of Emphasis on the Results of Conduct in Criminal Law*, 122 U. PA. L. REV. 1497 (1974); Thomas R. Shultz et al., 13 *Judgments of Causation, Responsibility, and Punishment in Cases of Harm-doing*, CAN. J. BEHAV. SCI. 238 (1981); Laurence Stern, *Deserved Punishment, Deserved Harm, Deserved Blame*, 45 PHIL. 317 (1970); Marie S. Tisak, *Preschool Children's Judgments of Moral and Personal Events Involving Physical Harm and Property Damage*, 39 MERRILL–PALMER. Q. 375 (1993); Amrisha Vaish et al., *Young Children Selectively Avoid Helping People With Harmful Intentions*, 81 CHILD DEV. 1661 (2010).

128. Gray, Schein & Ward, *supra* note 73; Gray & Keeney, *supra* note 78.

129. Such a model could posit, for example, that people rely so completely and automatically on their judgments of harmfulness to intuit on the appropriate level of punishment that they eventually come to associate harm with punishment even when punishment information comes *before* the harm judgment.

More broadly, these results also add to a growing body of research addressing the intuitive underpinnings of legal thinking.[130] We therefore join others in this field of research who seek to understand when the law aligns and misaligns with human psychology. Identifying the causes and consequences of misalignment is important for understanding how the law operates in people's lives and—where possible—addressing the mismatch. It is perhaps more important, however, in shaping how people react to the law. Although legal rules cannot (and should not) always reflect lay intuitions, research suggests that when policies and procedures make sense to people, they believe the system is more just, more legitimate, and more trustworthy.[131] By providing a deeper understanding of people's intuitive beliefs, research like ours can help policymakers find and address the gaps that might otherwise undermine these beliefs.

130. *See generally* Bregant et al., *supra* note 18; Cushman, *supra* note 61; Norman J. Finkel et al., *Equal or Proportional Justice for Accessories? Children's Pearls of Proportionate Wisdom*, 18 J. APPLIED DEV. PSYCHOL. 229 (1997); Ori Friedman et al., *First Possession, History, and Young Children's Ownership Judgments*, 84 CHILD DEV. 1519 (2013); Matthew R. Ginther et al., *The Language of Mens Rea*, 67 VAN. L. REV. 1327 (2014); Joshua Knobe, *Folk Judgments of Causation*, 40 STUD. HIST. PHILOS. SCI. 238 (2009); John Mikhail, *Moral Grammar and Intuitive Jurisprudence: A Formal Model of Unconscious Moral and Legal Knowledge*, 50 PSYCHOL. LEARNING & MOTIVATION 27 (2009); Melinda S. Mull & E. Margaret Evans, *Did She Mean to Do it? Acquiring a Folk Theory of Intentionality*, 107 J. EXP. CHILD PSYCHOL. 207 (2010); Richard E. Redding, *How Common-Sense Psychology Can Inform Law and Psycholegal Research*, 5 U. CHI. L. SCH. ROUNDTABLE 107 (1998); Tom R. Tyler & Robert J. Boeckmann, *Three Strikes and You Are Out, but Why? The Psychology of Public Support for Punishing Rule Breakers*, 31 L. SOC. REV. 237 (1997).

131. *See generally* Robert J. Boeckmann & Tom R. Tyler, *Commonsense Justice and Inclusion Within the Moral Community When Do People Receive Procedural Protections From Others ?*, 3 PSYCHOL. PUBLIC POL'Y & L. 362 (1997); Tom R. Tyler, *Conditions Leading to Value-Expressive Effects in Judgments of Procedural Justice: A Test of Four Models*, 52 J. PERSONALITY & SOC. PSYCHOL. 333 (1987); Tom R. Tyler, *Policing in Black and White: Ethnic Group Differences in Trust and Confidence in the Police*, 8 POLICE Q. 322 (2005); Tom R. Tyler & Kenneth Rasinski, *Procedural Justice, Institutional Legitimacy, and the Acceptance of Unpopular U.S. Supreme Court Decisions: A Reply to Gibson*, 25 LAW SOC'Y REV. 621 (1991).