RETRIBUTION AS ANCIENT ARTIFACT AND MODERN MALADY

by
Molly J. Walker Wilson[*]

One of the oldest and most entrenched goals of punishment is retribution, which is the idea that inflicting pain on someone who has committed a wrong is a worthwhile goal, regardless of any other benefits or harms that may result. Retribution has been the justification for increasingly punitive policies in the United States, the effect of which has decimated communities of color, strapped taxpayers with huge associated costs, and increased crime rates. It is difficult to understand why we perpetuate harmful policies based on "just deserts" until we consider that the foundation of these policies is moral outrage—a powerful, automatic, compelling response to witnessing social transgressions. Evidence from evolutionary biology, brain science, psychology, and anthropology has revealed the role of moral outrage in promoting social cooperation among early humans as social groups expanded. Moral outrage shares commonalities with other cognitive heuristics, or mental shortcuts that behavioral scientists have identified as leading humans to behave irrationally. While these automatic responses have historically served an adaptive function, they can lead to poor judgment in contemporary society. This Article employs scientific findings and theory from several disciplines to explore the origin and function of moral outrage, before examining the maladaptive consequences of retributivist objectives in modern times. Ultimately, all evidence suggests that retribution is an ancient artifact of human evolution only serving to create a foundation for harmful policies. As such, retribution should no longer be considered a legitimate punishment goal.

[*] Professor of Law and Psychology, Saint Louis University School of Law.

I. INTRODUCTION

An eye for an eye leaves the whole world blind.[1]

Just deserts, vengeance, an eye-for-an-eye—these concepts are widely touted as legitimate bases for punishing in response to an ever-expanding array of transgressions. Philosophers, psychologists, and scholars have described the retribution drive as part emotion, part reason. Austrian psychologist Fritz Heider called it a "force[]."[2] Born of anger, outrage, and blame, retributivist ideals are grounded in a desire for revenge and manifest in moral condemnation.[3] The justification for retribution is based on a hedonistic satisfaction resulting from the assuaging of moral outrage, which is its own end. A retributivist outcome may or may not result in optimal social benefit. No matter, the imposition of suffering upon someone who has committed a wrong is a good in and of itself.

In contrast to retribution is utilitarianism, a normative ethic stemming from the late eighteenth- and nineteenth-century English philosophers Jeremy Bentham and John Stuart Mill.[4] A utilitarian perspective judges an action as correct if it results in the greatest benefit for the largest number of people.[5] While utilitarianism can incorporate a notion of emotional satisfaction resulting from revenge, it

---

[1] THE YALE BOOK OF QUOTATIONS 269–70 (Fred R. Shapiro ed., 2006). This quote is famously attributed to Mahatma Gandhi, although the precise context of the notion has proven elusive. A similar sentiment was expressed by a Canadian lawmaker in 1914, arguing against the death penalty: "If in this present age we were to go back to the old time of 'an eye for an eye and a tooth for a tooth,' there would be very few hon. gentlemen in this House who would not, metaphorically speaking, be blind and toothless. "Canada, Parliament, *House of Commons Debates*, 12th Parl, 3rd Sess, Vol 113, No 4 at 496 (5 Feb. 1914) (statement of George P. Graham).

[2] FRITZ HEIDER, THE PSYCHOLOGY OF INTERPERSONAL RELATIONS 235 (1958).

[3] Rob Canton, *Crime, Punishment and the Moral Emotions: Righteous Minds and Their Attitudes Towards Punishment*, 17 PUNISHMENT & SOC'Y 54, 58–59 (2015).

[4] GEORGE CATLIN, THE STORY OF THE POLITICAL PHILOSOPHERS 342, 381 (1939).

[5] Jeremy Bentham believed that the usefulness in institutions and conduct "is that which makes for happiness." *Id.* at 368. Bentham specifically meant "the greatest happiness of the greatest number," meaning the greatest number of people; that is to say, Bentham is not concerned about the happiness of society, but rather, the claim of each individual to happiness. *Id.* at 369. Thus, in designing punishment, the legislature must seek the greatest happiness of the greatest number by "balancing" and by "social mechanics"—"one pressure against another." *Id.* at 369. Coming after Bentham, John Stuart Mill was also a utilitarian, but he wrote passionately about pleasure, arguing that there are both quantities and qualities of pleasure. *Id.* at 394. Mill argues that moral action is not only a matter of duty or of the intuition of conscience, but of the consequences for the world and its happiness. *Id.* at 395. Mill connected individual happiness to the happiness of society at large, meaning that the sum total of happiness or the "general happiness" is in "unity with our fellow-creatures," and thereby disregarding individualism. *Id.* Some scholars argue that Mill asserts that there is some kind of natural right or law that is superior to the sovereign's view of what is "socially useful." *Id*. at 396–97.

balances that good against any possible harm stemming from the punishment's infliction.[6] For utilitarianism, society's benefit must outweigh the aggregate costs of inflicting punishment to justify administering punishment.

Retribution as a justification for the state to impose consequences upon wrongdoers is a curious phenomenon standing in stark opposition to good-maximizing objectives traditionally offered to prop up our democratic institutions.[7] Retribution shares characteristics with the kinds of powerful negative emotions that we typically associate with irrational and counterproductive thought and behavior. The retributive approach lacks the qualities we routinely seek in policy justifications, such as cost-benefit analysis, fact-driven empirical justifications, and efficiency. Retribution-based punishment also defies systematic regulation and promulgates inconsistent outcomes. Because the underlying basis for this type of punishment is an emotional response, quantification and standardization are difficult or impossible. Despite these realities, the notion of just deserts is entrenched in the U.S. legal system, particularly in the criminal justice architecture.

Although the widespread acceptance of a retributivist approach seems at odds with modern sensibilities, a close examination of human emotions' prehistoric roots reveals an ancient rationale for retributive methods. Recent work in evolutionary psychology suggests that retribution is an artifact of the development of human cooperation and altruistic behavior. The emotional response associated with retribution, moral outrage, is thought to serve as a heuristic that cues a punishing impulse.[8] Experimental psychologists have recently suggested that modern-day rapid intuitions are artifacts from many thousands of years ago when early social groups would respond to norm violators with immediate and direct infliction of physical harm: "you fail to cooperate, I impose pain."[9] This early behavior was shaped by the enlargement of human social groups, our human ancestors' evolution, and social learning that assured the human species' success.[10]

---

[6] *See generally* JEREMY BENTHAM, *Principles of Penal Law*, *in* 1 THE WORKS OF JEREMY BENTHAM 365, 396 (John Bowring ed., Russell & Russell Inc. 1962) (1843).

[7] For example, the Government Accountability Office issued a report to Congress evaluating the extent to which rules issued by federal agencies had expressly been evaluated using a cost-benefit analysis. *See* U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-14-714, FEDERAL RULEMAKING: AGENCIES INCLUDED KEY ELEMENTS OF COST-BENEFIT ANALYSIS, BUT EXPLANATIONS OF REGULATIONS' SIGNIFICANCE COULD BE MORE TRANSPARENT (2014).

[8] *See* Jillian J. Jordan & David G. Rand, *Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment in One-Shot Anonymous Interactions*, 118 J. PERSONALITY & SOC. PSYCHOL. 57, 58 (2020).

[9] *Id.* at 58; *cf.* Joshua D. Greene et al., *Pushing Moral Buttons: The Interaction Between Personal Force and Intention in Moral Judgment*, 111 COGNITION 364 (2009).

[10] *See generally* MORRIS B. HOFFMAN, THE PUNISHER'S BRAIN: THE EVOLUTION OF JUDGE AND JURY 14 (2014).

Survival instincts, passed on through generations, remain with us today. Several of these primitive instincts involve social aggression, necessitated by competition for mates and defense of resources.[11] Although these aggressive tendencies are as ancient as the first Homo sapiens, they have lasted over thousands of years of evolution, multiple phases of human development, and are present in human society today.

One important phase followed a significant shift that occurred tens of thousands of years ago, when human society moved from a small-group structure to larger cooperative communities. Forming increasingly larger social groups was an excellent survival strategy. Humans benefitted in important ways, as the species increased the ability to imitate other humans' successful behaviors, to reason, and to benefit from social experience.[12] However, the shift was not without challenges. Whereas initially, virtually all human interaction occurred among familiar others, this pattern changed as social groups became increasingly populous. Archaeological evidence reveals that contact between individuals in the same tribes decreased as human communities grew larger.[13] Less frequent contact significantly diminished early humans' ability to punish freeloaders and transgressors, diluting reputational effects that served as a check on wrongdoers.[14] Humans lost much of their power to predict when they would be victimized and to respond to bad behaviors directly. As a result, humans became more reliant on other members of society to keep transgressors in check.[15] This reliance gave rise to a form of cooperative altruism. All members of society were called upon to police all other members. However, altruistic behavior imposes costs on the individual: imposing punishment can result in a backlash and harm to the punisher. Accordingly, a strong and almost irresistible impulse was required to motivate the necessary action. This impulse evolved as the emotion of moral outrage, and all but assured a strong reaction and instinct to punish. Hence, retribution was born. Retribution assured that wrongdoers would be

---

[11] DONALD SHARPES, THE EVOLVING HUMAN PRIMATE: AN EXPLORATION THROUGH THE NATURAL AND SOCIAL SCIENCES 32 (2016); *see also* Frans B. M. de Waal, *Morality and the Social Instincts: Continuity with the Other Primates*, *in* TANNER LECTURES ON HUM. VALUES, at 5 (Nov. 19–20, 2003) (pointing out that our bodies and minds are not designed for life without others; humans become "hopelessly depressed" in the absence of company, and without social support, human health outcomes deteriorate).

[12] SHARPES, *supra* note 11, at 28–29 (noting that companions offer advantages in locating food and avoiding predators); *see also* de Waal, *supra* note 11, at 5 (noting that humans "come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy").

[13] *See* R. Boyd & P.J. Richerson, *Culture and the Evolution of the Human Social Instincts*, *in* ROOTS OF HUMAN SOCIALITY: CULTURE, COGNITION, & INTERACTION 453, 471–72 (N.J. Enfield & Stephen C. Levinson eds., 2006).

[14] *Cf.* Dennis L. Krebs, *Morality: An Evolutionary Account*, 3 PERSP. ON PSYCHOL. SCI. 149, 159 (2008).

[15] *See* de Waal, *supra* note 11, at 16.

punished, even by those not directly affected by the wrongdoing.[16] As a result, all community members could reinforce the prevailing norms of behavior and rely on others to punish failures and to cooperate whenever and wherever they occurred.

Moral outrage's automatic emotional response was sufficiently compelling to motivate action, despite the potential costs to the punisher. Punishment researcher John Darley has described the "phenomenological characterization of the feelings persons have when they become aware of a transgression: a feeling of moral outrage," a feeling which "is produced by humans' intuitive systems and roughly computes what the transgressor justly deserves based on the moral wrongness of the transgression."[17] Social learning, driven by emotion, was an essential feature in developing moral outrage and punishment.[18] When early humans observed and imitated responses to the moral outrage experience, they learned from other group members, perpetuating previous generations' reactions and responses.[19]

Evidence that outrage is a cooperative social phenomenon comes from game research, revealing that when individuals see third parties victimized, they are motivated to punish the victimizer. Psychologists who study social cooperation have noted that "[t]he arousal of moralistic anger is not confined to injustices perpetrated against one's self. Witnessing the harming of a third party can also arouse strong feelings of anger and injustice."[20] Moreover, research has shown that individuals react most strongly when witnessing violations particular to their moral community.[21] The transgression represents an insult to the community's integrity, provoking moralistic anger, and an urge to punish the offender for the sake of all society.

Reactions to moral transgressions are automatic and powerful. Like other illogical impulses, the drive to punish is a primitive instinct. The human brain's prefrontal cortex has evolved to allow for higher thought and reason, but ancient predispositions are difficult to override. Within the past 20 years, behavioral scientists have come to understand that much of human behavior is influenced by heuristics or

---

[16] *See* Krebs, *supra* note 14, at 159.

[17] John M. Darley, *Morality in the Law: The Psychological Foundations of Citizens' Desires to Punish Transgressions*, 5 ANN. REV. L. & SOC. SCI. 1, 2 (2009).

[18] *See* NATALIE HENRICH & JOSEPH HENRICH, WHY HUMANS COOPERATE: A CULTURAL AND EVOLUTIONARY EXPLANATION 8 (2007) (describing social learning by portraying a human born in a band of hunter-gatherers trying to figure out what to eat). In this situation, a human would have the option to wander around the environment "sampling various potential foods" until it found something to eat. However, this is risky because there could be poisonous food in the environment. Instead, a social learner will "focus[ ] on the healthiest members of [the] group and eat[ ] whatever they eat." *Id.*

[19] *See id.* (describing generational cultural transmission as "an adaptive system of learned traits that accumulate over time").

[20] Dale T. Miller, *Disrespect and the Experience of Injustice*, 52 ANN. REV. PSYCHOL. 527, 535 (2001).

[21] *See* HEIDER, *supra* note 2, at 221; Miller, *supra* note 20, at 535.

mental "shortcuts" that follow a predictable pattern.[22] These heuristics are automatic, largely unconscious, and often defy logic. In the ancient world, shortcuts evolved to help our ancestors survive. For example, one of the most primitive impulses is our fight-or-flight response. When our ancient ancestors faced a charging lion, the automatic burst of adrenaline that speeds up our respiration, sends blood to our legs, and gives us a burst of energy was useful. Today, that same flood of adrenaline is often maladaptive in the context of modern performance. For example, rapid breathing, increased heart rate, and sweaty palms make public speaking more difficult. Like some other formerly adaptive instincts, the retribution impulse may have outlived its usefulness.[23] As evolutionary psychologists Eyal Aharoni and Alan Fridlund have argued, "[i]f human punishment is driven largely by retribution, and retribution is arrived at by way of heuristic judgments, we are forced to ask whether we can trust such judgments."[24]

When considering policies that serve our modern interests, we often talk about getting the most good for the least cost. This calculus is inherently utilitarian.[25] When it comes to punishment, the utilitarian goal of crime deterrence is widely touted. However, the current American system of incarceration cannot be endorsed using a utilitarian rationale. In contrast to a retribution model of criminal punishment, a utility-based theory asks which choice of various options will ultimately result in the greatest overall good to society. Under the utilitarian approach, the primary goal of punishment is to deter; pain is not exacted for the sake of causing pain—it is a means to an end.[26] Were utility the primary goal, our sentencing and penal practices would look very different. For example, it is impossible to explain current incarceration rates within a utilitarian framework. Incarceration exacts a heavy toll on taxpayers and communities. From a utilitarian perspective, the burdens argue in favor of incarceration *only* when there are significant benefits to offset the losses.[27] Studies of punishment and recidivism have shown that incarceration has not served to reduce crime and may *increase* recidivism.[28] Compared to a non-prisoner control group, offenders who received harsher sentences and prison time were

[22] *See* Gašper Štukelj, *On the Simplicity of Simple Heuristics*, 28 ADAPTIVE BEHAV. 261, 269 (2020).

[23] *See* Eyal Aharoni & Alan J. Fridlund, *Punishment Without Reason: Isolating Retribution in Lay Punishment of Criminal Offenders*, 18 PSYCHOL. PUB. POL'Y. & L. 599, 618 (2011).

[24] *Id.* (noting that research on infanticide behavior "shows how impulses that might have benefitted our ancestors' individual fitness may not be useful to societal groups").

[25] *See* RAYMOND G. GETTELL, HISTORY OF POLITICAL THOUGHT 340 (1924).

[26] *Id.* at 340, 345.

[27] *See* Francis T. Cullen et al., *Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science*, 91 PRISON J. 48S, 50S (2011).

[28] *See* Ojmarrh Mitchell et al., *Examining Prison Effects on Recidivism: A Regression Discontinuity Approach*, 13 J. EXPERIMENTAL CRIMINOLOGY 1, 19 (2017); *see also* Cullen et al., *supra* note 27, at 60S (finding that prisons do not reduce recidivism more than non-custodial

more likely to reoffend. Not only are harsh sentences failing to produce positive gains, but they are also creating significant collateral consequences.[29] In terms of impacts on individuals, society, and the cost to American taxpayers, the United States' harsh sentencing practices are anti-utilitarian.[30]

Although the data is clear, lawmakers have not moved decisively to change punishment policy. Rhetoric around the need to be "tough on crime" persists in the United States, although the United States is the world's most punitive country.[31] The rate of incarceration remains staggeringly high and is significantly higher than any other country.[32] As one commentator noted, "[i]t's a stark fact that the U.S. has less than 5 percent of the world's population, yet we have almost 25 percent of the world's total prison population."[33] Moreover, the trend has been inversely related to crime rates. Incarceration rates are "higher than they were 30 [to] 40 years ago, despite the fact that crime is at historic lows."[34] Lawmakers on both sides of the political aisle have lamented the effects of the U.S.'s draconian approach to punishment. One Republican senator pointed out that "[n]ot only does the current overpopulated, underfunded system hurt those incarcerated, it also digs deeper into the pockets of taxpaying Americans."[35]

---

sanctions, and that incarceration may increase recidivism in low-risk offenders). *But see* William Rhodes et al., *Relationship Between Prison Length of Stay and Recidivism: A Study Using Regression Discontinuity and Instrumental Variables with Multiple Break Points*, 17 CRIMINOLOGY & PUB. POL'Y 731, 733 (2018) (finding that lengthening a prison term does not increase recidivism, but rather reduces recidivism by a small amount); Daniel P. Mears et al., *Recidivism and Time Served in Prison*, 106 J. CRIM. L. & CRIMINOLOGY 83, 118–19 (2016) (finding that "there is an inverted U-shaped relationship between time served and recidivism, at least for inmates serving up to five to six years in prison"). This means that there may be no single effect of time served on recidivism, and that the effect of time served may vary depending on the specific amount of time served. *Id.* at 119.

[29] Mitchell et al., *supra* note 28, at 3.

[30] *See* John M. Darley, *On the Unlikely Prospect of Reducing Crime Rates by Increasing the Severity of Prison Sentences*, 13 J.L. & POL'Y 189, 193–95 (2005).

[31] *See Too Many Laws, Too Many Prisoners*, ECONOMIST (July 22, 2010), https://www.economist.com/briefing/2010/07/22/too-many-laws-too-many-prisoners; *see also United States Considered Most Punitive Country in the World*, EQUAL JUST. INITIATIVE (July 26, 2010), https://eji.org/news/united-states-is-most-punitive-country-in-the-world-study-says/.

[32] *See* Michelle Ye Hee Lee, *Yes, U.S. Locks People Up at a Higher Rate than Any Other Country*, WASH. POST (July 7, 2015, 12:00 AM), https://www.washingtonpost.com/news/fact-checker/wp/2015/04/30/does-the-united-states-really-have-five-percent-of-worlds-population-and-one-quarter-of-the-worlds-prisoners/.

[33] Hillary Clinton, *Remarks at Columbia University*, C-SPAN (Apr. 29, 2015), https://www.c-span.org/video/?325657-1/hillary-clinton-remarks-criminal-justice-reform.

[34] *Id.*

[35] Michelle Ye Hee Lee, *Does the United States Really Have 5 Percent of the World's Population and One Quarter of the World's Prisoners?*, WASH. POST (Apr. 30, 2015, 7:00 AM),

The effects of retribution-based punishment policies suggest that moral outrage creates policy that is antithetical to American interests. If lawmakers can identify moral outrage as an ancient survival strategy that no longer serves a useful function, they might be motivated to move away from policies that are not justified by any cost-benefit analysis. A utility-based model would eliminate harmful practices while creating space for restorative responses to crime. Restorative practices emphasize healing the victim, as well as reconciliation and reintegration of the offender. Models based exclusively on imposing pain fail to address the long-term effects on society resulting from the widespread, long-term incarceration of millions of Americans, including the consequences for poorer communities and the effects of the eventual release of inmates. A utilitarian approach would require quantification and measurement of costs and benefits of various punishment approaches. As crime rates fluctuate, and behavioral science and technology advance, social scientists and statisticians can provide innovative ways to measure and respond to crime. The best utilitarian approach offers an opportunity for flexibility and change, responding to shifting social inputs and requirements. The time has come to recognize the need for a new approach that leaves behind our outdated, harmful, gut-driven motivations in favor of a sophisticated approach with an eye toward the good of all members of society.

Part I of this Article first discusses the main tenets of retributivist and utilitarian punishment and then explores behavioral science findings that can shed light on how and why people seek to punish others. Part II delves into the evolutionary history of the emotion of moral outrage to explore its ancient role in promoting social cooperation. Part III explains how modern science, including behavioral research and brain-scan data, confirms the evolutionary roots of retribution. Part IV discusses the similarity between the human retribution instinct and other automatic heuristic cues that lead to poor judgments. Finally, Part V describes the myriad ways in which the type of draconian punishment that flows from retributivist goals violates the pragmatic and ethical objectives of a civilized, organized, and free society.

II. THEORIES OF PUNISHMENT AND HUMAN BEHAVIOR

Many, if not most, first-year criminal law classes begin with a discussion of punishment theories. The main theories justifying punishment are the utilitarian motives of deterrence and incapacitation, and the non-utilitarian goal of retribution or just deserts. Those who debate the merits of these different approaches vary widely in the justifications for their preferred method. Most agree that different punishment goals will support different processes, considerations, and results.

https://www.washingtonpost.com/news/fact-checker/wp/2015/04/30/does-the-united-states-really-have-five-percent-of-worlds-population-and-one-quarter-of-the-worlds-prisoners/.

Therefore, the debate on the merits of various punishment theories is not merely academic; it has important implications for real-world decisions about sentencing policy.

A. The Retribution Goal

"Retributive justice is a system by which offenders are punished in proportion to the moral magnitude of their intentionally committed harms."[36] This goal is famously associated with philosopher Immanuel Kant, who argued that those who violate laws deserve to be punished according to their "internal wickedness."[37] This perspective, related to the deontological approach, derives not from the future consequences of the punishment, but rather from a universal moral obligation to give wrongdoers what they "deserve."[38] Kant defended this stance as a "categorical imperative" to punish not for any practical end goal, but for its own sake.[39] The severity of punishment in a just deserts framework is presumed to be measured against offenders' moral blameworthiness for their transgression, without regard to their future dangerousness, the potential for rehabilitation, costs to society of punishment, or other possible consequences.[40] Deontologists who follow Kant's reasoning

---

[36] Kevin M. Carlsmith & John M. Darley, *Psychological Aspects of Retributive Justice*, 40 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 193, 194 (2008); *see also* Rob Canton, *Crime, Punishment and the Moral Emotions: Righteous Minds and Their Attitudes Towards Punishment*, 17 PUNISHMENT & SOC'Y 54, 57 (2015) (describing the principle within positive retributivism that affirms that "someone who is guilty ought to be punished" while "the negative principle asserts that someone who is not guilty must not be punished").

[37] Carlsmith & Darley, *supra* note 36, at 197.

[38] *Id.*; *see also* Canton, *supra* note 36, at 66 (noting that "[t]he idea that an unavenged wrong debases and pollutes the community may persist" and that "*cleansing* through punishment is a recurrent metaphor in justifications of retribution," as evidenced by Kant's insistence that "even if a civil society were to dissolve itself by common agreement . . . the last murderer remaining in prison must first be executed so that . . . the blood guilt thereof will not be fixed on the people").

[39] IMMANUEL KANT, THE SCIENCE OF RIGHT (1790), *reprinted in* 42 GREAT BOOKS OF THE WESTERN WORLD 392, 446 (W. Hastie trans., Robert Maynard Hutchins et al. eds., Encyc. Britannica, Inc. 1952). *But see* David Dolinko, *Some Thoughts About Retributivism*, 101 ETHICS 537, 542 (1991) (calling a bold retributivist someone who "asserts both that [the] lawbreakers deserve punishment and that this, all by itself, constitutes a good or sufficient reason for the state to inflict punishment on [the person]").

[40] *See* KANT, *supra* note 39, at 446–47; *see also* Dan Markel, *Are Shaming Punishments Beautifully Retributive? Retributivism and the Implications for the Alternative Sanctions Debate*, 54 VAND. L. REV. 2157, 2158–59 (2001) (noting that a "weak sense" of retributivism asserts that "a criminal may be punished because . . . he 'deserves' that punishment, and that punishment should be meted out in proportion to the wrong committed and the blameworthiness of the offender" while "strong" retributivism "incorporates the same desert and proportionality assertions [while] also impos[ing] an obligation: the criminal *must* be punished, regardless of the consequences").

have articulated several justifications for the retributivist approach, including proposing that inflicting pain on transgressors serves to restore a moral balance.[41] Kant and other strict retributivists reject various utility-based rationales, maintaining that "punishment can never be administered merely as a means for promoting another good."[42]

B. The Utilitarian Goal

For a utilitarian, punishment "gives force to the law" and "is necessary because it preserves social order": like any action in a utilitarian system, punishment is not valuable in itself, but only because it augments happiness.[43] As commentators have noted, the utilitarian's vision of punishment is "forward-looking to crime prevention, rather than backward-looking to moral guilt."[44] Prominent utilitarian Jeremy Bentham argued that punishment could only be justified because it would result in a net benefit to society.[45] Bentham argued that the right of one person to punish another is limited: "If the apparent magnitude, or rather value of [the] pain be greater than the apparent magnitude or value of the pleasure or good he expects to be the consequence of the act, he will be absolutely prevented from performing it."[46] This consequentialist philosophy holds that any decision to punish individuals must weigh the harm to the transgressor against the potential for future harm to society.[47] According to this approach, the justification for punishment is almost exclusively to

---

[41] *See* John Cottingham, *Varieties of Retribution*, 29 PHIL. Q. 238, 238–45 (1979); *see also* Dolinko, *supra* note 39, at 538 (pointing out that retributivism has traditionally provided the primary basis of support for the death penalty in the United States, and that Americans who support the death penalty do so based on retributive grounds). One death penalty scholar "has said that even if execution had no extra deterrent effect he would support it 'on grounds of justice alone.'" *Id. See generally* HERBERT L. PACKER, THE LIMITS OF THE CRIMINAL SANCTION 38 (1968).

[42] KANT, *supra* note 39, at 446.

[43] Michelle H. Kalstein et al., Comment, *Calculating Injustice: The Fixation on Punishment as Crime Control*, 27 HARV. C.R.-C.L. L. REV. 575, 579–80 (1992).

[44] *Id.* at 580. *See* Carissa Byrne Hessick & Douglas A. Berman, *Towards a Theory of Mitigation*, 96 B.U. L. REV. 161, 183 (2016) (discussing deterrence theory, which "seeks to decrease crime by using the threat of punishment to produce law-abiding behavior").

[45] *See* Carlsmith & Darley, *supra* note 36, at 197 (discussing Bentham's work).

[46] BENTHAM, *supra* note 6, at 396.

[47] *See* Carlsmith & Darley, *supra* note 36, at 197; s*ee also* Aharoni & Fridlund, *supra* note 23, at 600; Robert Justin Lipkin, *The Moral Good Theory of Punishment*, 40 U. FLA. L. REV. 17, 18 n.2 (1988) (explaining that a consequentialist theory "holds that a theory of the right is determined by a theory of the good," and that "consequentialist theories maintain that an act is right if it brings about more good than an alternative course of conduct").

promote social welfare by controlling future harmful behavior.[48] The utilitarian approach eschews consideration of moral guilt in favor of a cost-benefit analysis and proactive minimization of bad outcomes.[49]

Critics of utilitarianism argue that, followed to its logical end, pure consideration of utility might lead to widely condemned acts like lying and stealing. Proponents of utilitarian theory have pointed out that in the aggregate, the utilitarian approach cannot support minor moral transgressions that could be justified by small utilitarian gains, because if promulgated widely, they would be destructive. For instance, the occasional small "white" lie might be beneficial in the short-term, but if repeated over and over, lying could lead to widespread distrust, contribute to the breakdown of social order, and erode our institutions.[50] Therefore, proponents of this approach argue, the larger goal of societal order and happiness would keep these minor transgressions in check.

Another objection levied—particularly against Jeremy Bentham's hedonistic emphasis—holds that the value of life is more important than a balance of pleasure over pain. Proponents of utilitarianism address these objections by arguing that non-hedonistic values can be accounted for by the theory. For example, British philosophers like G.E. Moore regarded many kinds of consciousness—including love, knowledge, and the experience of beauty—as intrinsically valuable, independent of pleasure, a position labeled "ideal utilitarianism."[51]

Consequentialist goals are served through various methods, targeted to achieve maximum benefit and social wellbeing. The greatest utility may be achieved through incapacitation (e.g., imprisonment or execution), specific deterrence, and general deterrence.[52] Specific deterrence focuses on an individual offender; the imposition or threat of punishment creates a disincentive for that individual to reoffend by providing a painful consequence for committing a crime.[53] General deterrence is the process of discouraging the commitment of offenses by members of the general public.[54] This form of deterrence differs from "specific deterrence" in that it reaches

---

[48] *See* Carlsmith & Darley, *supra* note 36, at 197.

[49] *See id.*

[50] Another objection to utilitarianism is that it would prove unworkable. Individuals have idiosyncratic preferences; most people, if given the choice between either saving a loved one or saving two strangers from impending disaster would choose the loved one. And yet, saving two lives clearly has more utility than saving one. The answer to this objection is that when the State is choosing policy, it should operate out of a concern for the utility of a policy to broader society. In a democratic system, utilitarian objectives should supplant the individual preferences of policy makers.

[51] GEOFFREY SCARRE, UTILITARIANISM 114–20 (1996).

[52] *See* Carlsmith & Darley, *supra* note 36, at 200–02; *see also* Kalstein et al, *supra* note 43, at 580–81.

[53] Carlsmith & Darley, *supra* note 36, at 200.

[54] *Id.* at 201.

a broader pool of potential offenders, and theoretically prevents a wider range of infractions among a larger number of people.[55] A single act of punishment can accomplish both specific and general deterrence when the punishment is publicized so that people in the community see the individual offender's suffering and are discouraged from committing the offense for fear of receiving the same consequences.[56] Bentham emphasized the importance of general deterrence, arguing that "prevention ought to be the chief end of punishment, as it is its real justification."[57]

C. Psychological Studies of Punishing Behavior

The subject of punishment has received a great deal of attention from philosophers, politicians, educators, behavioral scientists, and others. There is a sizable body of multidisciplinary literature addressing the justifications for imposing punishment.[58] Animal and human behaviorists have written extensively on the consequences of receiving punishment—often in empirical studies on learning.[59] Psychologists have recently turned their attention to the social and cognitive factors influencing people's desire to punish.[60] In recent writing, psychologist Neil Vidmar proposed a six-stage model of the social-psychological dynamics of retribution:

> (1) there is a perceived rule or norm violation; (2) the rule violator's intention is perceived as blameworthy; (3) the combination of (1) and (2) threatens or actually harms values related to the perceiver's personal self, status, or internalized group values; (4) the negative emotion of anger is aroused; (5) the

---

[55] *Id.* at 201–02.

[56] *See id.* at 201 (pointing out that "[t]aken to its logical extreme, philosophers have argued that the key features of retribution—severity of the crime, responsibility, mitigating factors, etc.—are entirely irrelevant to a theory of deterrence. In fact, even guilt is irrelevant."). The difficulty with this supposition is that it ignores the fact that individual members of society recognize that in a functioning democracy, punishment is limited by the moral obligations of its citizenry, and this sets practical limits upon how sanctions are imposed.

[57] BENTHAM, *supra* note 6, at 396.

[58] *See* ARTHUR SHUSTER, PUNISHMENT AND THE HISTORY OF POLITICAL PHILOSOPHY: FROM CLASSICAL REPUBLICANISM TO THE CRISIS OF MODERN CRIMINAL JUSTICE 90–91 (2016) (discussing Italian philosopher Cesare Beccaria's theory of punishment). *See generally* BENTHAM, *supra* note 6; KANT *supra* note 39; Gertrude Ezorsky, *How Many Lives Shall We Save?*, 3 METAPHILOSOPHY 156 (1972).

[59] *See generally* B.F. SKINNER, ABOUT BEHAVIORISM (1974); B.F. SKINNER, THE BEHAVIOR OF ORGANISMS: AN EXPERIMENTAL ANALYSIS (1938); JOHN B. WATSON, BEHAVIORISM (1924).

[60] *See* Kevin M. Carlsmith et al., *The Paradoxical Consequences of Revenge*, 95 J. PERSONALITY & SOC. PSYCHOL. 1316, 1316 (2008); *see also* Carlsmith & Darley, *supra* note 36, at 203–11.

> cognitions and aroused emotions foster reactions against the violator; (6) during or following punishment the anger dissipates, cognitions return toward homeostasis, and the rule or norm is perceived to be vindicated.[61]

Empirical findings on altruism, revenge, empathy, social consensus, perspective-taking, ego maintenance, and emotion have played a role in informing our understanding of when, why, and how human beings punish.[62] Increasingly, policy advocates have advocated the importance of using empirical findings to inform policy choices. The broad body of research is vital to our understanding of the legitimacy of the various punishment philosophies.

In scholarship, some projects are *descriptive* and some are *prescriptive* or *normative*. A descriptive project aims to explain how something is, while a prescriptive project posits how things *should* be.[63] Behavioral science can illuminate how people make decisions and what factors influence their choices. Law, philosophy, religion, and other disciplines take up the normative piece. In the arena of punishment, the behavioral scientist's role may seem superfluous because we have hundreds of pages explaining the reasons for various penalties. For instance, politicians argue about the merits of various crime response approaches, whereas activists and religious leaders take messages to the streets, houses of worship, and the public. Moreover, principles and practices relating to crime response forge international treaties.

Most Americans have opinions on the *when*, *why*, and *how* of punishment, and most of us are comfortable explaining and even arguing for our preferred positions.[64] However, while people usually feel like they understand the reasons for their attitudes and behaviors, social scientists have empirically demonstrated a disconnect between true and perceived motivations. An enormous amount of what influences us is unconscious.[65] For example, we are biased against certain groups based upon

---

[61] Neil Vidmar, *Retribution and Revenge*, *in* HANDBOOK OF JUSTICE RESEARCH IN LAW 31, 43 (Joseph Sanders & V. Lee Hamilton eds., 2001); *see also* Neil Vidmar & Dale T. Miller, *Socialpsychological Processes Underlying Attitudes Toward Legal Punishment*, 14 L. & SOC'Y REV. 565, 568 (1980) (arguing that social norms—and their violations—are at the heart of retributive justice).

[62] *See generally* Canton, *supra* note 36; Carlsmith et al., *supra* note 60; Roger Giner-Sorolla & Hanah A. Chapman, *Beyond Purity: Moral Disgust Toward Bad Character*, 28 PSYCHOL. SCI. 80 (2017); Omar Tonsi Eldakar et al., *Emotions and Actions Associated with Altruistic Helping and Punishment*, 4 EVOLUTIONARY PSYCHOL. 274 (2006); Felix Warneken, *Altruistic Behaviors from a Developmental and Comparative Perspective*, *in* COOPERATION AND ITS EVOLUTION 399 (Kim Sterelny et al. eds., 2013).

[63] As usual, I attempt both in this work.

[64] *See* Aharoni & Fridlund, *supra* note 23, at 614–15.

[65] *See, e.g.*, Daniel M. Wegner & David J. Schneider, *The White Bear Story*, 14 PSYCHOL. INQUIRY 326 (2003).

assumptions we make and attitudes that we hold.[66] This bias persists even when we are unaware of its existence or when we explicitly *reject* it.[67] We make mistakes because we think too highly of ourselves, although we rarely acknowledge this fact.[68] We are motivated by many situational factors that never even cross the threshold of consciousness.[69] Moreover, we experience emotion as natural and correct in certain circumstances, without realizing how easily our emotions can be manipulated, and how profoundly our affective reactions influence our choices.[70]

These are just some examples of myriad influences on our behavior, of which we are blissfully unaware, even as our actions reflect them and we justify and defend

---

[66] *See* Alex Madva, *Implicit Bias, Moods, and Moral Responsibility*, 99 PAC. PHIL. Q. 53, 53 (2018) (describing John Dovidio's influential 2002 study which determined that white college student participants had "anti-racist explicit attitudes but racially-biased implicit attitudes"); *see also* Lilia M. Cortina et al., *Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact*, 39 J. MGMT. 1579, 1594–96 (2013) (finding that women, people of color, and particularly African American women reported more uncivil treatment at work than other groups); Neil Levy, *Am I a Racist? Implicit Bias and the Ascription of Racism*, 67 PHIL. Q. 534, 534–35 (2017) (noting that "when otherwise identical CVs of minority and majority applicants are submitted to potential employers . . . the minority candidates get fewer callbacks from potential employers and fewer invitations to interview," identifying that "[t]here is good evidence that many people harbour attitudes that conflict with those they endorse").

[67] *See* Madva, *supra* note 66, at 53.

[68] *See* Ward Farnsworth, *The Legal Regulation of Self-Serving Bias*, 37 U.C. DAVIS L. REV. 567, 568–69 (2003) (describing self-serving bias as "[p]eople tend to believe what they want to believe"). Self-serving bias can be manifested as a tendency for people to see themselves as having a greater than average share in their level of skill at common tasks like driving or ability to get along with others, as well as skewed predictions—"that which is desired is thought more likely to occur than that which is undesired." *Id.* at 569; *see also* Daniel S. Nagin & Greg Pogarsky, *An Experimental Investigation of Deterrence: Cheating, Self-Serving Bias, and Impulsivity*, 41 CRIMINOLOGY 167, 171 (2003) (calling self-serving bias "the tendency for individuals to shade judgments in a manner favorable to themselves," and noting that "most individuals believe they are above average" in the areas of driving, managing, productivity, and ethics).

[69] *See also* Levy, *supra* note 66, at 535 (pointing out that "[i]mplicit attitudes . . . are not able to be deployed at the personal level [but] rather they influence cognition in ways that escape conscious control"). Levy notes that implicit biases have a disproportionate effect on behavior when humans cannot or do not exercise personal level control—when they lack the cognitive resources because they are "tired, stressed, or under cognitive load," or because they are "required to respond too quickly for effortful processing." *Id.*

[70] *See* Jordan Etkin & Anastasiya Pocheptsova Ghosh, *When Being in a Positive Mood Increases Choice Deferral*, 45 J. CONSUMER RES. 208, 215 (2018) (finding that being in a positive mood can make unrelated decisions more difficult, and as a result, choice deferral increases); *see also* Ellen C. Garbarino & Julie A. Edell, *Cognitive Effort, Affect, and Choice*, 24 J. CONSUMER RES. 147, 148 (1997) (noting that "[a] consistent finding is that humans have limited cognitive resources and allocate them judiciously" and "[a]s environments require more cognitive effort to process information fully, decision makers often switch to decision strategies or heuristics that are easier to implement," though these heuristics "frequently result in less accurate decisions, biased responses, and preference reversals").

the choices and attitudes they produce. When it comes to punishing, most of the conversation assumes that we know why we punish—we argue about which *why* is most justified, but not whether we even understand our own attitudes. A descriptive project explains much of the unconscious influences that animate our conscious choices around punishment. Exploring how human attitudes towards punishment are created is vital to understanding the motivations behind punishment and the usefulness or destructiveness of our punishment impulses.

Describing the psychology of punishment has been the life work of Paul Robinson, John Darley, Kevin Carlsmith, and their colleagues.[71] Darley, Robinson, and Carlsmith have created various scenarios designed to implicate certain considerations and have asked people to decide whether and how much to punish.[72] This method circumvents the need to solicit self-reports, and is therefore impervious to the problem of self-reporting bias created by asking people to explain their psychological processes. Instead, the research uses behavioral outputs to determine the rationale that drives human preference.[73]

Recent research on punishment has specifically focused on contrasting deontological and utilitarian motivations.[74] The majority of this work is designed to determine which of these two justifications best describes ordinary citizens' motivation when they choose to sanction another person. The studies focus on whether most people adopt a just deserts perspective, in which the focus is on an eye for an eye, or a utilitarian deterrence perspective, in which the focus is on preventing future harms against society.[75]

---

[71] *See, e.g.*, Kevin M. Carlsmith et al., *Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment*, 83 J. PERSONALITY & SOC. PSYCHOL. 284 (2002).

[72] Carlsmith & Darley, *supra* note 36, at 203–11.

[73] *Id.* at 198. Carlsmith and Darley, who have generated a great deal of the data in this area, describe the method this way:

> Our prototypical experimental design pits two perspectives against each other in a 2 x 2 arrangement. We create four versions of a vignette that describes an offense, and ask participants to recommend an appropriate punishment for one of those cases. . . . [W]e modify the vignette so that it is either relevant or irrelevant to each perspective. For example, the theory of retribution is concerned with the moral culpability of the perpetrator, and thus the vignette is adjusted such that the moral culpability is either high or low. A retributivist will adjust his or her punishment accordingly. By contrast, a committed deterrence theorist would be relatively unmoved by this variable, and far more sensitive to a deterrence-related variable such as the frequency of the particular type of crime.

Id. at 198–99.

[74] *See, e.g.*, Aharoni & Fridlund, *supra* note 23, at 599–600.

[75] *See* Carlsmith et al., *supra* note 71, at 285 (noting that just-deserts theory relies on several factors that help people determine how much punishment an offender deserves; one is the magnitude of harm and another is extenuating circumstances suggesting that leniency is appropriate).

The challenge with this type of work is substantial, because it is difficult to know whether people who were motivated by one goal were also motivated by the other goal. For example, someone who appears to have decided punishment based upon a just deserts motivation may also have been influenced by a desire to deter future crime.[76] Moreover, assuming that the perfect experimental design could completely parse the two goals, there would remain the question of whether individuals *in real life* operate strictly upon one dimension.[77] In other words, it may be that even if we can determine that most individuals are more strongly motivated by a just deserts rationale, that doesn't rule out the possibility that these individuals also intend to satisfy a utilitarian goal.

D. The Emotional Basis for Retribution

The foundational human experience that manifests itself in retributivist reasoning is primarily an emotional one. Psychologists have long theorized about the role of emotions in the evolution of humankind. In particular, emotions are signals that guide individuals to make choices that give the individual the best chance to survive and thrive.[78] Fear—and the fight-or-flight response—is one of the most basic examples of adaptive emotive responses.[79] The automaticity of the fear reaction was essential to survival. When our ancestors were confronted with a charging

---

[76] *See* Aharoni & Fridlund, *supra* note 23, at 601–02 (explaining that within psychological research on punishment, there is a difficulty disentangling retributive and consequentialist motives for punishment).

[77] *See generally* Carlsmith & Darley, *supra* note 36.

[78] Evolutionary biologists have theorized about the adaptive function of "cuteness" in baby animals. A positive and protective instinct that is aroused when humans see a kitten is an emotional reaction. This emotion is a cue that tells us not to harm vulnerable young creatures. Young animals of many species have developed features that cause older animals to be less likely to harm them when they are most vulnerable. Features that we associate with "cuteness" are beneficial to the young animal, of course, but may also have prevented early humans from killing prey too early, before the prey could develop into a worthwhile food source. *See* Gary D. Sherman et al., *Individual Differences in the Physical Embodiment of Care: Prosocially Oriented Women Respond to Cuteness by Becoming More Physically Careful*, 13 EMOTION 151, 155–57 (2012) (finding that since emotional states are associated with subtle shifts in the motor system, emotions "may be associated with motor preparation for emotion-specific behaviors, such as providing care" when presented with cuteness. (citation omitted)); *see also* Canton, *supra* note 36, at 59 (noting that "moral emotions are typically articulated *persuasively* . . . . [N]ot only [are] they . . . linked to the interests of others, but . . . we feel that other people ought to share them").

[79] For an explanation about how the emotions of punishment are distinctively moral emotions—"emotions of judgement . . . righteousness, and reprobation"—see Canton, *supra* note 36, at 59 (pointing out that "[o]ne defining criterion of a moral emotion is that it is 'linked to the interests or welfare either of society as a whole or at least of persons other than the judge or agent'") (citing Jonathan Haidt, *The Moral Emotions*, *in* HANDBOOK OF AFFECTIVE SCIENCES 852, 853 (Richard J. Davidson et al. eds., 2003)).

bear, hesitating could mean the difference between living and being mauled to death.[80] Several different kinds of negative emotions have been shown to be powerful motivators for human behavior. Our society and relationships with one another are a rich source of an extensive range of human emotions.[81] In fact, these emotions are what allow us to function in society.

The emotion dubbed "moral outrage" by modern psychologists has also been called an "ought force," and it is the basis for retributivist punishment. Research conducted by psychologist Phil Tetlock and his colleagues on moral outrage found it to be a nearly universal precursor to punishment determinations.[82] At a more basic level, anger is an important component of moral outrage, which in turn, drives retributivist decision-making.[83] When individuals make decisions about punishment, their feelings of moral outrage influence their choices.[84] When exploring these con-

---

[80] *See* Peter LaFreniere, *Human Emotions as Multipurpose Adaptations*: *An Evolutionary Perspective on the Development of Fear*, *in* EVOLUTIONARY PERSPECTIVES ON HUMAN DEVELOPMENT 189, 190–94 (Robert L. Burgess & Kevin MacDonald eds., 2nd ed. 2005) ("Basic human emotions such as joy, love, anger, fear, and sadness . . . [are] core, species-specific motivational systems that organize behavior and development across [a human's] life span."). Studies on infant attachment provide an example of how fear can contribute to regulate behavior. *See id*. at 192. The presence of an infant's caregiver provides a secure base for the infant to confidently explore the environment—if something threatening should occur, the infant's attention will be directed to the caregiver's face, and if the caregiver expresses fear, a fear response will be classically conditioned in the infant, to that event, after a single trial. *Id.* But, classical conditioning alone is insufficient to account for all fear stimuli and according to some scholars, at least five general principles are required to explain fear: intensity, novelty, evolutionary dangers, social stimuli, and conditioned stimuli. *Id.* at 197. Examples of evolutionary dangers include combinations of "biology and experience" such as "pain, being left alone, sudden changes in stimulation, and rapid approach." *Id*. Additional potential stimuli that have an evolutionary basis include "fear of strangers; separation anxiety and fear of being alone; fear of open places, heights, falling, or loss of support; fear of the dark; and fear of snakes or spiders." *Id.*

[81] For an explanation of how the way individuals interact with others has been shown directly related to a previously experienced emotion, see Rebecca Hollander-Blumoff, *Crime, Punishment, and the Psychology of Self-Control*, 61 EMORY L.J. 501, 546 (2012) ("[I]t may be that the very behavior that society would like individuals to control, the natural consequence of their violent impulse, is directly connected to the negative emotion they are experiencing and that the action they believe will ameliorate their negative emotional state is the act of violence itself. If affect regulation does trump self-control, some violent crime may be explained as an effort to improve mood.").

[82] *See* Carlsmith et al., *supra* note 71, at 286 (referring to Philip E. Tetlock et al., *Revising the Value Pluralism Model: Incorporating Social Content and Context Postulates*, 8 ONT. SYMP. ON PERSONALITY & SOC. PSYCHOL. 25 (1996)).

[83] *See id.* at 286–87; *see also* Alan Page Fiske & Philip E. Tetlock, *Taboo Trade-Offs: Reactions to Transactions that Transgress the Spheres of Justice*, 18 POL. PSYCHOL. 255, 285–86 (1997).

[84] *See* Zachary K. Rothschild & Lucas A. Keefer, *A Cleansing Fire: Moral Outrage Alleviates Guilt and Buffers Threats to One's Moral Identity*, 41 MOTIVATION & EMOTION 209, 215 (2017)

nections, Lerner, Goldberg, and Tetlock found a relationship between factors associated with just deserts, such as culpability of a transgressor, and later expressions of moral outrage in third-party observers.[85] The outrage was also positively related to subsequent punitive behavior.[86] "People respond to moral transgressions with gut-level emotional responses and these emotional responses play a central role in how people react to, and reason about, morally relevant behavior."[87] Given that moral outrage is often expressed on behalf of the victim of the moral violation, moral outrage is described as a "prosocial emotion reflecting a desire to restore justice by fighting [for] the victimized."[88] Scholars find that moral outrage promotes positive social outcomes and is associated with behaviors such as protesting, supporting political action, and desiring to punish transgressors on behalf of innocent victims.[89] Further, moral outrage arises in response to morally reprehensible behavior and influences punishment severity.[90]

This ought feeling—the automatic emotional reaction and feelings of satisfaction after punishing transgressors—has been defended as a legitimate basis for action on the ground that acting upon it feels good.[91] For example, according to legal scholar Toni Massaro, the attraction of retributivism is that it "satisfies deep emotional, intuitive instincts."[92] Moreover, Massaro points out the parsimonious nature of the retributivist objective, saying, "its ends are simply stated and fairly easy to

("Perceiving oneself to be the victim of illegitimate harm or insult . . . elicit[s] feelings of *personal anger*, [which] is similar, yet conceptually distinct from moral outrage.").

[85] *See* Jennifer S. Lerner et al., *Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility*, 24 PERSONALITY & SOC. PSYCHOL. BULL. 563, 564 (1998).

[86] *See id.* at 570.

[87] Brock Bastian et al., *The Roles of Dehumanization and Moral Outrage in Retributive Justice*, PLOS ONE, Apr. 23, 2013, at 1; *see also* Aarti Iyer et al., *Why Individuals Protest the Perceived Transgressions of Their Country: The Role of Anger, Shame, and Guilt*, 33 PERSONALITY & SOC. PSYCHOL. BULL. 572, 574 (2007) (noting that "emotions motivate individuals to take action [and that] [d]iscrete emotions orient individuals toward a specific mode of responding to a situation because specific emotions are linked to distinct goals and action intentions").

[88] Rothschild & Keefer, *supra* note 84, at 209.

[89] *Id.*; *see also* GERD GIGERENZER, GUT FEELINGS: THE INTELLIGENCE OF THE UNCONSCIOUS 211–12 (2007) (pointing out that "[m]oral outrage can result when powerful people . . . place their family interests over their loyalty to their country, such as when the news spread that out of all U.S. senators and congressmen only one had a son fighting in Iraq"); Iyer et al., *supra* note 87, at 584 (noting that in the context of a country's transgression, emotions can help explain why people participate in specific opposition strategies, and that "shame is a distinct emotional response to the perceived transgressions of one's country").

[90] *See* Bastian et al., *supra* note 87, at 2 (arguing that moral outrage and dehumanization are associated and that morally outraged individuals view offenders as unlikely to be rehabilitated).

[91] *See* Toni M. Massaro, *Shame, Culture, and American Criminal Law*, 89 MICH. L. REV. 1880, 1892 (1991); *see also* Markel, *supra* note 40, at 2181.

[92] Massaro, *supra* note 91, at 1892; *see also* Markel, *supra* note 40, at 2181.

secure."[93] In other words, when motivated by this force, "[w]e punish in order to avenge the harm, not to deter, rehabilitate, or contain. Revenge is easier to accomplish than these other objectives."[94] Interestingly, implicit in this assertion is the judgment that simplicity and satisfaction of emotional drives are both desirable and sufficient objectives to justify choosing one particular punishment scheme over another.

Although moral outrage is the emotion most closely associated with retributivism, directly connected to a just deserts punishment rationale, other emotions become relevant, depending upon the violation and context.

E. The Emotion-Action Connection

Negative emotions, such as those associated with transgressions, are powerful motivators. Research has revealed that disgust, anger, and fear serve as a basis for decision-making.[95] Although traditionally, the law has been somewhat ambivalent about the role of emotion in penal theory and practice, recent trends suggest the widespread acceptance of a larger role for emotion in criminal justice, along with an explicit acceptance of affect-based responses. As Suzanne Karstedt has pointed out,

> During the last decade, the secular process of restricting the space of emotions in the penal realm seems to have taken a turn towards bringing emotions back in. A process of 're-emotionalization of law' or the 'reassertion of emotionality

---

[93] Massaro, *supra* note 91, at 1892; *cf.* Kenneth Einar Himma, *Luck, Culpability, and the Retributivist Justification of Punishment*, 22 LEWIS & CLARK L. REV. 709, 723 (2018).

[94] Massaro, *supra* note 91, at 1892; *see* Markel, *supra* note 40, at 2180–81 (explaining that Massaro argues that retribution possesses the following characteristics: "First, retributivists argue that punishment is justified to counteract the harm inflicted by the wrongdoer because the wrongdoer deserves it. Second, an eye for an eye is proper redress for a crime, in order to set right the moral balance. Third, retributive justice is nonconsequentialist in that it is uninterested in influencing the offender's future behavior or the behavior of other community members. Fourth, retributivism presupposes free will by the criminal actor.") (citations omitted).

[95] *See* Krebs, *supra* note 14, at 165 (noting that scholars have suggested that "emotional reactions may be strategically superior to those based on rational calculation [and that] [a]ffective reactions such as sympathy, disgust, and righteous indignation should exert an immediate effect on moral decision-making processes [though] [p]eople should have difficulty justifying decisions derived in these ways, and if called on to justify them, offer plausible post hoc rationalizations.") (citations omitted); *see also* Cynthia E. Cryder et al., *Guilty Feelings, Targeted Actions*, 38 PERSONALITY & SOC. PSYCHOL. BULL. 607, 615 (2012) (finding that "[g]uilt does not lead to increased generosity toward all . . . it increases generosity toward those whom one has wronged and only when [they] can notice the gesture . . ." therefore not prompting the "general moral 'cleansing' in response to transgressions;" "guilt specifically and strategically prompts people to repair specific social transgressions"); Iyer et al., *supra* note 87, at 572 (noting that "[a]lthough citizens often agree with their government's actions . . ." when a policy is judged to be immoral or harmful, "[i]ndividuals express opposition to such perceived transgressions by participating in political activities").

> in law' spread around the globe, and has changed the criminal justice system in many ways. The 'return of emotions' to criminal justice and penal policies has occurred in two arenas: the emotionalization of public discourse about crime and criminal justice, and the implementation of sanctions in the criminal justice system that are explicitly based on—or designed to arouse—emotions.[96]

One of the debates lurking beneath the recent acceptance of emotional decisions about punishment is the balance between natural human impulses and the restriction of these impulses. Recently, the vital role of emotions in decision-making has come to light.[97] The former notion that rational, effortful decisions are exclusively the domain of emotion-free brain activity has been challenged by the important role of emotion in cognition, as recognized in the work of Antonio Damasio, whose "Somatic Marker Hypothesis" has been among the most influential theories of emotion in recent years.[98] Somatic markers are emotional reactions that support rational decision-making. A primary function of somatic markers is to permit quick and automatic preselection of the relevant choice alternatives. Somatic makers increase the efficiency of human decision-making, but the accuracy of these affect-based cues depends upon their appropriateness to the context and their accuracy in directing the decision-maker toward an appropriate choice. Similar to other fast and frugal cognitive shortcuts, they streamline decision-making. Like other automatic cognitive processes, they are not infallible.

Behavioral researchers have studied commonalities in human behavior, particularly in decision-making and attitude formation, and have discovered that many human patterns end up looking irrational.[99] For example, human cognitive and social reasoning rarely follow rules of logic. People are not terribly good at solving

---

[96] Susanne Karstedt, *Emotions and Criminal Justice*, 6 THEORETICAL CRIMINOLOGY 299, 301 (2002) (citation omitted); *see also* Kathy Laster & Pat O'Malley, *Sensitive New-Age Laws: The Reassertion of Emotionality in Law*, 24 INT'L J. SOC. L. 35 (1996).

[97] *See, e.g.*, Antoine Bechara, *The Role of Emotion in Decision-Making: Evidence from Neurological Patients with Orbitofrontal Damage*, 55 BRAIN & COGNITION 30 (2004) ("Most theories of choice assume that decisions derive from an assessment of the future outcomes of various options and alternatives through some type of cost-benefit analyses. The influence of emotions on decision-making is largely ignored. The studies of decision-making in neurological patients who can no longer process emotional information normally suggest that people make judgments not only by evaluating the consequences and their probability of occurring, but also and even sometimes primarily at a gut or emotional level."); *see also* Hillary Brown, *The Role of Emotion in Decision-Making*, 13 J. ADULT PROTECTION 194 (2011).

[98] *See generally* Antonio R. Damasio, *The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex*, 351 PHIL. TRANSACTIONS: BIOLOGICAL SCI. 1413 (1996).

[99] *See* Bonnie M. Perdue & Ella R. Brown, *Irrational Choice Behavior in Human and Nonhuman Primates*, 21 ANIMAL COGNITION 227, 227–28 (2018); *see also* Anna Sircova et al., *Simulating Irrational Human Behavior to Prevent Resource Depletion*, PLOS ONE, Mar. 11, 2015, at 1, 3, 14.

puzzles; they are unjustifiably optimistic and counterproductively self-centered.[100] Their memories are poor and they are often influenced by emotions that cause them to act against interest.[101] People routinely exhibit social biases of which they are unaware and which they whole-heartedly reject on principle.[102] In order to be satisfied that designing our penal system on a retributivist model reflects the best policy choice, we should have evidence that our moral instincts are "correct," or minimally, that having them entitles us to act on them. If, on the other hand, we have evidence that the instincts that lead us to be retributivists are irrational or fail to comport with agreed-upon goals and values, then we should reject this model.[103]

III. RETRIBUTION AS AN EVOLUTIONARY ARTIFACT

The adaptive function of moral outrage can be traced back to early humans, and the evolutionary need related to the requirement of cooperation as these early humans lived in increasingly large social groups.[104] Members of the human race, like all other animal and plant species, have evolved to exhibit a wide variety of

---

[100] *See* Nagin & Pogarsky, *supra* note 68, at 171 (calling self-serving bias "the tendency for individuals to shade judgments in a manner favorable to themselves," and noting that "most individuals believe they are above average [in the areas of] driving, managing, productivity, and ethics") (citations omitted). *See generally* Farnsworth, *supra* note 68.

[101] *See* Edmund T. Rolls, *Functions of Human Emotional Memory: The Brain and Emotion*, *in* THE MEMORY PROCESS: NEUROSCIENTIFIC AND HUMANISTIC PERSPECTIVES 173, 179 (Suzanne Nalbantian et al. eds., 2011) (noting that a human's current mood "can affect the cognitive evaluation of events or memories. Thus, we are more likely to recall happy memories when we are happy and depressing memories when we are depressed."). For more functions of memories, see *id.* at 179–80 (noting that because they linger "for minutes or longer after a reinforcing stimulus has occurred, emotion may continue to motivate behavior, to help achieve a goal or goals"). Some scholars suggest that there are two routes for making decisions: an emotional route that can operate unconsciously, and a rational, conscious route. Both of these produce different processing systems for different types of emotional memory. *See id.* at 185 (explaining that "[w]hen the decisions are made by the unconscious emotional system, the rational, conscious system may confabulate an explanation for the decision having little to do with the original causes of the behavior").

[102] Madva, *supra* note 66, at 53 (describing Dovidio's study on white college students with anti-racist explicit attitudes, but racially biased implicit attitudes).

[103] *See* Aharoni & Fridlund, *supra* note 23, at 618. Evidence that people's retributivist instincts fail to result in optimal outcomes is discussed *infra*, Part IV.

[104] "One illustration of ancestral prosociality from the fossil record relates to the existence of hominine skeletal remains dating back as far as 1.77 million years ago showing signs of both severe physical disabilities and years of survival with those disabilities." Michael Bang Petersen, *Evolutionary Political Psychology: On the Origin and Structure of Heuristics and Biases in Politics*, 36 POL. PSYCHOL. (SUPPLEMENT I) 45, 50 (2015) (citing Jean-Jacques Hublin, *The Prehistory of Compassion*, 106 PROC. NAT'L ACAD. SCI. U.S. 6429, 6429 (2009) to note the inference drawn by archaeologists that someone must have cared for these disabled individuals).

behavioral predispositions designed to increase the chance of survival of future generations.[105] In the case of human beings, behavioral adaptations are highly contingent—more so than for other species because of our ability to meta-cognate and engage in abstract reasoning. Evidence suggests that the predisposition to exact revenge for others' moral transgressions reflect tens of thousands of years of natural selection. As evolutionary biologist Evan Jones notes, the accumulated effects of selective favoring of certain traits operate "not only on the external form of our distant ancestors, but also on the brain's neural architecture and information processing pathways."[106] Behaviors that ensured the greatest chance of survival of our ancient ancestors became internalized as automatic intuitions that increase the probability that humans will respond to certain situations in predictable ways.[107] "In sum, the more directly and substantially a behavior affected the reproductive success of our ancestors, human and nonhuman, the greater the likely effect of evolutionary processes on the current patterns of its incidence."[108]

In order to understand how our retributivist instincts evolved, evolutionary behavioralists turn to fossil records revealing early development of social order in ancient humans and human ancestors. Comparing the social structure of related species provides valuable insight into which behaviors were advantageous, and which were not. For example, Neanderthal and modern humans both lived on the

---

[105] Owen D. Jones, *Sex, Culture, and the Biology of Rape: Toward Explanation and Prevention*, 87 CALIF. L. REV. 827, 833 (1999); *see* Dominic Johnson & Jesse Bering, *Hand of God, Mind of Man: Punishment and Cognition in the Evolution of Cooperation*, 4 EVOLUTIONARY PSYCHOL. 219, 220 (2006) (noting that "[c]ooperation is widespread among mammals, birds, insects, cells, microscopic organisms, and different organs of the body").

[106] Jones, *supra* note 105, at 833; *see also* Dennis Krebs, *An Evolutionary Reconceptualization of Kohlberg's Model of Moral Development*, *in* EVOLUTIONARY PERSPECTIVES ON HUMAN DEVELOPMENT 243, 250 (Robert L. Burgess & Kevin MacDonald eds., 2nd ed. 2005) (pointing out that "the ultimate goal of most (but not all) behavior is to enhance actors' inclusive fitness, defined in terms of the number of replicas of their genes they contribute to future generations"); *id.* at 251 (noting that some evolutionary psychologists and scholars believe that "mental structures that give rise to moral judgments and moral behaviors are activated by particular kinds of problems and their function is to process information in ways that give rise to particular kinds of decisions").

[107] Jones, *supra* note 105, at 833. For an example discussing the genetic costs of inbreeding arising from recessive genes as well as the results of outbreeding, see PATRICK BATESON, BEHAVIOUR, DEVELOPMENT, AND EVOLUTION 67 (2017) ("Finding a compatible partner is an important part of reproductive behaviour in many animals . . . [m]embers of different species do not make good mates [and] too much inbreeding can also reduce reproductive success.").

[108] Jones, *supra* note 105, at 833; *see* Joshua M. Ackerman et al., *The Behavioral Immune System: Current Concerns and Future Directions*, 12 SOC. & PERSONALITY PSYCHOL. COMPASS 1, 2–3 (2018) (finding that "reactive and proactive behavioral immune responses help protect individuals, their offspring, and slow the rate of horizontal disease transmission within groups").

Earth 450,000 years ago.[109] Modern humans gained a survival advantage over Neanderthal when the former began expanding its social groups to include greater numbers of members.[110] As early humans expanded their social groups to include a growing number of individuals, they gained an increasing advantage over other social species, eventually becoming a top predator.[111] Cooperative living allowed early humans to harness the hunting, gathering, protection, and childbearing skills of many members of the group in order to optimize the overall success of the species.[112] Interspecies "[c]ooperation occurs when an individual incurs a cost to provide a benefit for another person or people."[113] Because there are some associated costs, cooperative tendencies only evolved biologically because they conferred a "competitive advantage on the cooperators."[114] "[W]e developed the cognitive, linguistic,

---

[109] Terence A. Brown, *Stranger from Siberia*, 464 NATURE 838, 839 (2010).

[110] *See* Richard G. Klein, *Language and Human Evolution*, 43 J. NEUROLINGUISTICS 204, 207 (2017) (noting that "[o]n average, Neanderthal brains were as large or larger than those of living humans [and that] [t]he benefits of a larger brain are obvious," but come with costs such as difficulty giving birth and consumption of a larger percentage of the body's metabolic resources.) However, a larger brain comes with a "greater ability to process sensory input and to construct more effective mental models of the physical and social environment," which would contribute to more effective representations of social relationships as groups grew larger and more complex. *Id.*

[111] *See* BRIAN HARE & VANESSA WOODS, THE GENIUS OF DOGS 26–27 (2013) (pointing out that Neanderthal could never become top predator because their social groups were too small).

[112] *See* Krebs, *supra* note 106, at 252 (noting that homo sapiens are "among the most social" of the species in the animal kingdom). Further, "mechanisms that dispose animals to aggregate and interact with other members of their species, and . . . members of other species [interspecies cooperation] evolve when the mechanisms help animals foster their biological interests. Such mechanisms may help animals enhance their inclusive fitness in several ways. As examples, individuals who band together may be less susceptible to predators than more solitary individuals, and groups may be able to hunt larger game than individuals could kill on their own." *Id.*

[113] HENRICH & HENRICH, *supra* note 18, at 37–39 (noting that this cost could involve resources, time, or exposure to some threat); *see* Johnson & Bering, *supra* note 105, at 220 ("Sometimes cooperation results in mutual payoffs to all actors involved, and can therefore be easily understood as each pursuing their own selfish interest. However, other instances of cooperation are more surprising, because individuals help others despite incurring a cost in doing so."); *see also* Krebs, *supra* note 106, at 253 ("Selfish preferences pose a serious problem for the evolution of cooperative mechanisms . . . . Most people consider selfish behaviors, defined as fostering one's own interests at the expense of others, immoral; and most people consider cooperative behaviors, defined as fostering one's interests in ways that foster the interests of others, moral."). *But see* Johnson & Bering, *supra* note 105, at 220 (recognizing that "cooperation *among humans* is still not understood").

[114] JOSHUA GREENE, MORAL TRIBES: EMOTION, REASON, AND THE GAP BETWEEN US AND THEM 24 (2013) (noting that the phrase "altruistic punishment" reflects the fact that punishment is costly for the individual who punishes in addition to the one who is punished). The benefits of curtailing selfish activities by punishment are often shared by a larger group that includes but is not restricted to the punisher; when this condition is met, punishment becomes an act of altruism because it increases the fitness of others at the expense of one's own fitness. However, scholars

and other capacities to structure our social interactions in ways that allowed altruistic cooperators to proliferate."[115] Early human environments demonstrate that "members of groups that sustained cooperative strategies for provisioning, child-rearing, sanctioning non-cooperators, defending against hostile neighbors, and . . . sharing information had significant advantages over members of non-cooperative groups."[116] For example, cooperation in hunting, "whereby different individuals perform different, well-specified roles," allowed for an increased reliance on large game as a calorie source.[117] Over the course of human history, humans created social and physical environments exhibiting similar or greater benefits of cooperation, including irrigated agriculture, modern industry, warfare, and information systems.[118] The expansion of patterns of interaction within and between members of groups has been termed the "hyper-sociality" of humans.[119] Behavioral byproducts of hyper-sociality have been examined by anthropologists, who have compiled detailed records of universal traits that have been found in every human society studied. These behaviors include "territoriality, conflict, family, food sharing, group living, empathy, dominance/submission, cooperation, coalitions, collective decision making, etiquette, rituals, and weapons."[120]

A. Larger Group Size Required More Social Cooperation

As the groups became larger, individual relationships necessarily became more tenuous; less direct and repeated contact between any given two members of a group became less frequent. As a result, the potential for free-loading and non-cooperation increased. Opportunities to interact directly with all members in the group to test

---

maintain that punishment is motivated by very different psychological mechanisms—anger and moral outrage—than the helping behaviors typically associated with altruism: empathy and sympathy. *Id.*

[115] SAMUEL BOWLES & HERBERT GINTIS, A COOPERATIVE SPECIES: HUMAN RECIPROCITY AND ITS EVOLUTION 196 (2011); *see* HENRICH & HENRICH, *supra* note 18, at 40 (pointing out that human cooperation varies substantially from that of nonhuman primates in scale and nature of variability).

[116] BOWLES & GINTIS, *supra* note 115, at 3. For an example of this type of cooperation, see *id.* at 196 (noting that "[h]uman reliance on the meat of large hunted animals [developed along with] . . . [o]rganizing and sharing the returns to successful hunting" because of the high risk, which was mitigated by sharing information about hunting and other valued resources); *see also* Klein, *supra* note 110, at 213 (noting that "[l]iving humans can learn to produce stone tools entirely by imitation, but they learn more reliably and efficiently when they are instructed verbally").

[117] Petersen, *supra* note 104, at 49.

[118] *See* BOWLES & GINTIS, *supra* note 115, at 3.

[119] Petersen, *supra* note 104, at 48–49.

[120] *Id.* at 50.

their loyalty and adherence to the social pact diminished.[121] The growth of the size of cooperative groups diminished opportunities for direct reciprocity ("if you help me, I will help you").[122] "Direct reciprocity depends on direct, ongoing experience between interacting individuals."[123] Accordingly, the success of small groups could not translate directly to the success of larger groups. As Henrich explains, "the capacity for reciprocity to maintain cooperation decreases geometrically as the group size increases."[124] Therefore, "we should not expect direct reciprocity to be the primary factor in maintaining cooperation in large groups."[125] Given what scholars now know about the increase of early human social groups, direct reciprocity can no longer be accepted as the main model. Instead, *indirect* reciprocity, where cooperation is sustained not by individual enforcement but by community enforcement, evolved as a behavioral strategy that would allow for cooperation on a large scale.[126]

Prior to evolving to live cooperatively in increasingly larger groups, direct confrontation was the primary way one individual would hold another accountable.[127] Because ancient humans lacked technologies that could reduce the immediacy of the act—there was no nuclear button, no law enforcement agency, no organized

---

[121] *See id.* at 50–51; Krebs, *supra* note 14, at 155 ("Members of groups cannot interact with everyone to the same extent . . . [and] it is in group members' interest to fill their 'association niches' with partners who are most willing and best able to foster their fitness. As a result, members of groups tend to form mutually-beneficial relationships or friendships with those who possess matching or complementary abilities and resources.").

[122] HENRICH & HENRICH, *supra* note 18, at 48–51.

[123] *Id.* at 48; *see* Krebs, *supra* note 14, at 155 (noting that "[m]echanisms have evolved in many species that dispose them to form social bonds with and support those on whom their fitness is dependent, such as mates, offspring, parents, siblings, friends, exchange partners, or members of coalitions and groups").

[124] HENRICH & HENRICH, *supra* note 18, at 51.

[125] *Id.*

[126] *See* Krebs, *supra* note 14, at 155–56 (describing the "evolution of altruism," which induces animals to "behave in biologically altruistic ways" to enhance fitness, and noting that the evolution of altruism occurred in at least three ways: sexual selection, kin selection, and group selection); *see also* Francesco Guala, *Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate*, 35 BEHAV. & BRAIN SCI. 1, 3 (2012) (noting that reciprocal altruism is "the idea that what seems altruistic in the short run might actually be self-serving in the long term [and that] [o]rganisms that help others may be indirectly maximizing their own fitness, if their help is going to be reciprocated in the future").

[127] HOFFMAN, *supra* note 10, at 14–17 (pointing out that humans evolved in small groups of mostly related individuals, which gives humans "enormous survival advantages, and therefore . . . incentives to cooperate with one another. . . . we have deep emotional ties to our groups, and a powerful hunger for social belonging"). In Hoffman's view, this deep need for belonging creates a "deeply embedded tension between cooperation and cheating, between community and individuality, between selflessness and selfishness," and leads to the problem of whether and how to punish. *Id.* at 14–15.

criminal justice system—the punishment process was immediate and intimate. Survival depended on individuals understanding and accepting consequences to violating norms. A rough, automatic, instinctive impulse to physically reprimand a non-cooperative other was therefore not only appropriate, it was necessary.[128]

As an intensely social species, humans were so reliant upon cooperation that the instinct to enforce cooperation and to punish non-cooperation is a central theme of human evolution. Cooperative norms became enmeshed in the evolutionary fabric of human adaptation and are central to our success story.[129] According to Harvard biologist Joseph Henrich,

> By spreading group-beneficial cooperative norms involving the punishment of non-cooperative norm violators, cultural group selection may have altered the selective environment faced by genes. This altered environment may have favored genes that promote things like a readiness to acquire cooperative and punishing norms, a default bias toward helping (to avoid being punished), a preparedness to respond to punishment, and numerous other social faculties.[130]

According to anthropologists and evolutionary psychologists, adaptive goals have been to "(1) meet requests from reciprocators, (2) deny help to cheaters, and (3) potentially to educate cheaters in order to widen and strengthen the social exchange system for mutual insurance."[131]

In order to function together and to survive and thrive, a mechanism was needed to keep a check on all members of the group, and punish anyone who was

---

[128] *See* Krebs, *supra* note 106, at 254 ("Viewed biologically, moral judgments are a form of communication. . . . [A]nimals are evolved to send signals that induce recipients to behave in ways that foster the senders' interests or that manipulate them."). Some scholars argue that there is no necessary inconsistency between behaving morally and fostering one's biological interests—that is to say, it's not immoral to attempt to survive or to protect one's children; instead, some people argue that parents have a moral obligation to preserve their lives and the lives of their children. *Id.* at 258.

[129] *See id.* at 254–58. Third party punishment plays a key role in establishing and enforcing social norms in large organizations, and third-party punishment is a fundamental ingredient of social cohesion. *See* HOFFMAN, *supra* note 10, at 150–73 (discussing third-party punishment); *see also id.* at 16 (pointing out that human brains are the most complex in the animal kingdom and humans are "more intensely social than any genetically heterogeneous . . . species"—these facts are not unrelated, Hoffman argues, since humans need "massively networked brains just to be able to keep track of each other"); Marco Fabbri & Emanuela Carbonara, *Social Influence on Third-Party Punishment: An Experiment*, 62 J. ECON. PSYCHOL. 204, 205 (2017).

[130] HENRICH & HENRICH, *supra* note 18, at 134; *see also* HOFFMAN, *supra* note 10, at 16 (pointing out that humans are intensely social beings because being intensely social gave humans "significant survival advantages in areas such as mutual defense and hunting," which were crucial to survival in the warm, wet, and rich Southern African jungles where our predecessor primates lived).

[131] Petersen, *supra* note 104, at 60.

behaving in a way that would contravene the interests of the group as a whole.[132] In response to these pressures, early hominids developed emotive reactions that functioned to prompt compliance with a set of cooperative norms.[133] These gut emotive reactions served the dual goal of discouraging humans from harming each other while simultaneously encouraging individuals, and by extension society, to punish those who did.[134] The feeling of outrage had to be strong and automatic to function reliably across of range of situations. Scholars have suggested that a just deserts model assures that punishment occurs—the emotion of outrage, which is associated with retributivism, compels punishment, even in situations in which individuals have competing motives not to punish.[135] Angry blame is thus a product of evolution that helps maintain social order by encouraging individuals to administer sanctions, even when to do so comes at a cost to the punisher.[136] Scholars Victoria McGeer and Friederike Funk noted:

---

[132] *See* Matteo Mameli, *Evolution, Motivation, and Moral Beliefs*, *in* COOPERATION AND ITS EVOLUTION 525, 527 (Kim Sterelny et al. eds., 2013) ("[T]hose in our lineage who had moral beliefs in favor of (certain kinds of) cooperative behaviors had higher fitness (on average). Thinking of cooperation in moral terms generates a robust and reliable motivation to cooperate. Believing that cooperation is morally required—demanded in an inescapable and automatically authoritative way—makes one more likely to cooperate."); *see also* Johnson & Bering, *supra* note 105, at 221 (arguing that "[i]t would be incredible to suggest that religion has nothing to do with cooperation" in ancient or modern societies, and that religion is a key promoter of within-group cooperation during human evolution). Further, Johnson and Bering argue that "the expectation and fear of supernatural punishment . . . serves to promote cooperation" and that this mechanism evolved via individual selection and that any group selection effects, though not necessary, would help drive the system. *Id.*

[133] For a discussion on how these emotions are often viewed today as "moral beliefs," see Mameli, *supra* note 132, at 527 (noting that the view that "moral beliefs are a source of cooperative motivations that can work in cases where (i) cooperating is extremely important for genetic fitness and (ii) the other sources of motivation are unlikely to be effective. It is because of this that . . . the ability to form moral beliefs was genetically selected for").

[134] *See* Neil Vidmar, *Retributive Justice: Its Social Context*, *in* THE JUSTICE MOTIVE IN EVERYDAY LIFE 291, 293 (Michael Ross & Dale T. Miller eds., 2002). *See generally* HEIDER, *supra* note 2.

[135] *See* Caryl E. Rusbult et al., *Forgiveness and Relational Repair*, *in* HANDBOOK OF FORGIVENESS 185, 189–93 (Everett L. Worthington, Jr. ed., 2005) (providing physiological evidence in support of the idea that retaliation is an instinctual response to being transgressed against).

[136] *See* Ernst Fehr & Armin Falk, *Psychological Foundations of Incentives*, 46 EUR. ECON. REV. 687, 706 (2002); Ernst Fehr & Urs Fischbacher, *Social Norms and Human Cooperation*, 8 TRENDS COGNITIVE SCI. 185, 188 (2004); Ernst Fehr & Simon Gächter, *Altruistic Punishment in Humans*, 415 NATURE 137, 137–39 (2002); Gerald Gaus, *Retributive Justice and Social Cooperation*, *in* RETRIBUTIVISM: ESSAYS ON THEORY AND POLICY 73, 84–85 (Mark D. White ed., 2011); Joseph Henrich et al., *In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies*, 91 AM. ECON. REV. 73, 77 (2001).

> [O]ur human taste for punishment . . . is undoubtedly honed and shaped by cultural norms, a number of theorists suggest that natural selection may have solved the problem of large scale social cooperation in our human lineage by endowing us with a "punishment mechanism"—a hardwired piece of cognitive machinery that ensures our sensitivity to the transgression of social norms and triggers a punitive reaction.[137]

B. Moral Outrage as Communication Signal

In a system of indirect reciprocity, the population behaved according to an increasingly complex network of social norms.[138] Signals provided a way for the population to distinguish the defectors from the cooperators using reputation.[139] Under a system of indirect reciprocity, individuals interact with each other only occasionally but benefit by "receiv[ing] information about the past behavior of the individual with whom they are about to interact."[140] In order for this system to function reliably, however, violators had to be consistently sanctioned.[141] Hence, punishment served an important communication function as between group members, as well as between the group and the individual violator.[142] When viewed as

---

[137] Victoria McGeer & Friederike Funk, *Are 'Optimistic' Theories of Criminal Justice Psychologically Feasible? The Probative Case of Civic Republicanism*, 11 CRIM. L. & PHIL. 523, 534–35 (2017).

[138] *See* Rory Smead, *Indirect Reciprocity and the Evolution of "Moral Signals"*, 25 BIOLOGY & PHIL. 33, 35–36 (2010); *see also* HOFFMAN, *supra* note 10, at 17 (noting that the development of language contributed to humans' ability to communicate rules for behavior, and "not only memorialized a compact for behavior, they also memorialized a compact to punish"). Every member in a group became able to enforce the rules. *Id.* ("[P]unishment itself had been socialized."); *see also id.* at 22 ("Punishment deterred enough cheating so that living in groups was possible. Having brains that punished allowed us to have brains that cooperated.").

[139] Smead, *supra* note 138, at 36.

[140] HENRICH & HENRICH, *supra* note 18, at 58.

[141] *See* Smead, *supra* note 138, at 35–36; *see also* HOFFMAN, *supra* note 10, at 33 (noting that the rule of law serves as a way for humans to bring together cooperation and punishment as groups get bigger. "Laws automatically redefined the expanding social whole, and entitled, at least in theory, all members of the new groupings to the protections of the old ones."); *id*. at 34 (asserting that law is one of the few remaining modern reminders of our deeply embedded "evolutionary schizophrenia" over cooperating and cheating, over right and wrong). Hoffman also points out that under English common law, acts that were considered inherent wrongs—as opposed to wrongs that were merely prohibited wrongs—were dealt with by way of private revenge. *Id.* This meant that except for crimes of treason and regicide, most ancient and medieval states did not get involved in crimes, even those like homicide, and instead, left the punishment to the victims and their family, clan, or tribe. *See id.* This is because it is evolutionarily significant to punish. *See id.*

[142] *See* Smead, *supra* note 138, at 33–36; *see also* Helen Y. Weng et al., *The Role of Compassion in Altruistic Helping and Punishment Behavior*, PLOS ONE, Dec. 10, 2015, at 2 ("Societal benefits may also occur . . . where personal resources are used to negatively impact those who have

motivation to communicate and enforce norms, emotional reactions to violations were themselves signals of a commitment to act, following a third-party transgression.[143] As humans evolved, the species' capacity to build institutions and to use cultural transmission of learned behavior grew, allowing early humans to transmit information to create distinctive social environments that reduced the costs borne by altruistic cooperators and raised the costs of free riding.[144]

Humans' cognitive, linguistic, and physical capacities made them particularly good at forming cooperative groups.[145] Economists Bowles and Gintis have argued that "[t]hese capacities allow [humans] to formulate general norms of social conduct, to erect social institutions regulating this conduct, [and] to communicate these rules and what they entail in particular situations."[146] Additional psychological capacities include the ability to internalize norms, to experience social emotions like shame and moral outrage, and to base group membership on non-kin characteristics like ethnicity and language.[147] Further, "[p]unishment reduces the gain to free-rid-

---

transgressed against others"—what scholars call "altruistic punishment"; also, "[s]ocial norms aim to foster social peace, stabilize cooperation and enhance prosperity.").

[143] McGeer & Funk, *supra* note 137, at 535 ("In Frank's terms, emotions like angry blame do not simply prime us to act in ways that serve our broader rational interests—*e.g.*, punishing to stabilize cooperative social norms. They also operate as 'commitment devices,' making credible to others the overwhelming likelihood of our so acting. Hence, our blaming sensitivity to transgression acquires, in addition, an important signalling function: we signal to others that bad behaviour *will* be punished, and so our blaming anger comes to serve as a powerful deterrent to such behaviour." (discussing ROBERT H. FRANK, PASSIONS WITHIN REASON: THE STRATEGIC ROLE OF THE EMOTIONS (1988))); *see* Jordan & Rand, *supra* note 8, at 57 (pointing out that "research demonstrates that punishment can serve to promote and maintain prosocial behavior by deterring selfishness . . . [t]hus, moralistic punishment plays a critical role in shaping human morality and supporting prosocial behavior").

[144] BOWLES & GINTIS, *supra* note 115, at 197; *see* Johnson & Bering, *supra* note 105, at 222 (arguing that there is ethnographic evidence to suggest that "the threat of supernatural punishment for norm transgressions exerts a powerful effect" on human behavior; not only is fear of supernatural punishment common across cultures, but it is commonly linked to taboos concerning life, death, scarce resources, sexual access, food sharing, hunting, divisions of labor, defense, and warfare); *see also* Krebs, *supra* note 14, at 151 (noting that Charles Darwin "recognized that animals may obtain benefits from group living by exchanging goods and services and by coordinating . . . efforts to obtain food, defend . . . against predators, remove parasites, and build shelters").

[145] BOWLES & GINTIS, *supra* note 115, at 5; *see* HOFFMAN, *supra* note 10, at 16 (humans' "powerful social brains, built for cooperation," are also good at cheating, meaning that humans have "sophisticated and sensitive systems for detecting opportunities to cheat—so that we can decide whether other members will catch us if we steal that food the group worked so hard to gather, and, if so, whether and how they will punish us," for example).

[146] BOWLES & GINTIS, *supra* note 115, at 5; *see* HOFFMAN, *supra* note 10, at 29–30 ("[S]ocially cooperative behaviors" evolved in our species.).

[147] BOWLES & GINTIS, *supra* note 115, at 5.

ing" individuals in a group, and may induce even those who are entirely self-interested to cooperate—thus, "groups with more punishers can sustain more cooperation" and ultimately be more successful.[148]

According to Bowles and Gintis, altruistic social preferences supporting cooperation outcompeted unmitigated and amoral self-interest for three key reasons: (1) "human groups have devised ways to protect their altruistic members from exploitation by the self-interested" via behaviors such as shunning, ostracism, and even execution of those who violate cooperative norms; (2) "humans adopted prolonged and elaborate systems of socialization" to cause individuals to "internalize the norms that induce cooperation, so that contributing to common projects and punishing defectors became objectives in their own right"; and (3) "between-group competition for resources and survival . . . remains a decisive force in human evolutionary dynamics," so that "[g]roups with many cooperative members tend[] to survive challenges and to encroach upon the territory of the less cooperative groups."[149]

Heider's "ought force" derives from the individual's commitment to the values of that individual's moral community.[150] The violation of these forces represents an insult to the integrity of the community and provokes in its members both moralistic anger and the urge to punish the offender. The automaticity of the experience of outrage and the desire to punish has been explained this way:

> If Charles believes that Tim has violated a moral norm—that is, Charles has a moral belief *M* about a practical demand *m* and also believes that Tim has violated *m*—then Charles will feel moral disapproval toward Tim and may be motivated to express this moral disapproval and punish Tim. Punishment is of course unpleasant, and so is being morally disapproved of, which can be seen as a particular kind of punishment. Insofar as Tim is motivated to avoid being punished by Charles and to avoid Charles's moral disapproval, Tim will also be motivated to behave in accordance with Charles's moral belief. . . . When Tim considers acting in a way that goes against a moral belief that Tim knows Charles to have, Tim will experience a negative emotion at the thought that Charles will feel moral disapproval toward him. The potential action will become negatively emotionally marked, and Tim will thereby be motivated not to perform the action. So, for example, if Charles believes

---

[148] *See id.* at 148; Joan B. Silk & Bailey R. House, *Evolutionary Foundations of Human Prosocial Sentiments*, 108 PROC. NAT'L ACAD. SCI. U.S. 10,910, 10,910 (Supp. 2, 2011) (noting that "humans are remarkably altruistic primates. . . . and [that] cooperation extends beyond the bounds of close kinship and networks of reciprocating partners. In humans, altruism is motivated at least in part by empathy and concern for the welfare of others").

[149] BOWLES & GINTIS, *supra* note 115, at 4; *cf.* HENRICH & HENRICH, *supra* note 18, at 69; Silk & House, *supra* note 148, at 10,910.

[150] *See* HEIDER, *supra* note 2, at 219.

> that keeping promises is morally required, and Tim knows this, Tim's fear of Charles's moral disapproval will motivate Tim to keep promises.[151]

"Viewed from this perspective, disinterested justice reactions are not disinterested at all," because every member of the community "has a stake in seeing that the rules and values of the authority structure under which they live are respected."[152] A relatively cohesive social response kept would-be violators in check. As one anthropologist notes, "the evolutionary dynamics of this relationship between punishment and prosociality make it likely that punishment will operate via a specialized behavioral adaptation . . . ."[153] The moral outrage impulse evolved in response to the need for retribution-based punishment, based upon the clearest and most basic marker: the deservingness of the punishment, or the culpability of the wrongdoer.[154] The characteristics of the behavioral response—the inflexibility of the retributivist impulse that has evolved in humans—therefore ironically had a utilitarian function in the evolutionary history of humankind.[155] One way to think about the retributivist impulse is that it serves as a proxy for a more nuanced and complex case-by-case determination of the appropriate response to transgressions to serve a utilitarian need. One evolutionary psychologist proposed this analogy: "Having an innate taste for sugar or fat circumvents the problem of learning by brute association which properties of potential foodstuffs are correlated with which future energetic states. Similarly, having an innate taste for punishment would circumvent the problem of learning associatively how to elicit future prosociality from social partners."[156] As a result of evolving at a time when brute, reflexive responses were critical to survival, the impulse has become so fixed and inflexible, that it defies reason, and operates even when the circumstances would call for a different response.[157]

C. Retribution and Social Differentiation

Additional support for the notion that retributive impulses evolved in order to keep group members in check derives from studies of the relative punitiveness of

---

[151] Mameli, *supra* note 132, at 531.

[152] Miller, *supra* note 20, at 535.

[153] Fiery Cushman, *The Role of Learning in Punishment, Prosociality, and Human Uniqueness*, *in* COOPERATION AND ITS EVOLUTION 333, 334 (Kim Sterelny et al. eds., 2013).

[154] John M. Darley & Thane S. Pittman, *The Psychology of Compensatory and Retributive Justice*, 7 PERSONALITY & SOC. PSYCHOL. REV. 324, 326, 332–33 (2003).

[155] Cushman, *supra* note 153, at 347 ("Of course, retributive motivations might reliably produce deterrent or incapacitative effects. In fact, I have taken pains to argue that the best way to understand the functional value of punishment is precisely in terms of deterrence—that is, eliciting contingent prosociality in future interactions.").

[156] *Id.* at 348.

[157] *Id.*

members of a group to in-group and out-group members.[158] Vidmar has described a variety of incidents in which people responded more punitively to in-group than out-group offenders.[159] He describes an incident in Newfoundland, Canada, where Catholics and Protestants are both salient religious groups. Several years ago, it was discovered that members of a Catholic religious order had sexually abused young boys who were in their care. Following discovery of the abuse, Catholics expressed a much stronger desire for severe punishment than did Protestants.[160] The more intense reaction of Catholics illustrates the fact that moral outrage is stronger when one shares an affiliation with the offender.[161] From an evolutionary standpoint, cooperation among those individuals with whom one shared a group would be *particularly* important. In-group members are far more likely to be dependent upon one another. Moreover, the community would have a particularly acute interest in encouraging prosocial behavior from powerful individuals within their own social group. High-status individuals would likely have greater access to resources, and bad behavior on the part of high-status individuals would be most likely to result in disaster for the group. One study found that the status of a transgressor was a factor in determining how punitive an in-group member was toward that transgressor. When the transgressor was high in status—signaling more power and responsibility in the group—participants were more punitive than they were toward an out-group transgressor. But when the in-group transgressor was low status—and therefore less important to the group's survival—respondents did not express a desire to punish the in-group transgressor more than the out-group transgressor.[162]

---

[158] An in-group member is someone who is part of one's own group. An out-group member is a member of a different group. It is important to note that individuals belong to a variety of different groups. A law professor is a member of the legal academy and shares membership with her colleagues at her own institution and others. She may also be a member of a neighborhood organization, professional groups, a book club, church or synagogue, co-op, and a family. Moreover, she shares membership in the broader community, her country, her gender, and when relevant, her role as a parent, adult child, sibling. Any significant aspect of an individual's identity can be grounds for in-group affiliation. *See generally* Michael E. McCullough et al., *Cognitive Systems for Revenge and Forgiveness*, 36 BEHAV. & BRAIN SCI. 1, 35 (2013) ("[E]volutionarily-relevant characteristics of the defendants (such as their sex and attractiveness), and shared characteristics between defendants and jurors (such as race or sexuality, triggering in-group/out-group prejudice), influence punitive sentiment and leniency or harshness in sentencing.").

[159] Vidmar, *supra* note 134, at 294–300.

[160] *Id.* at 295–96.

[161] *Id.* at 300 ("In short, social harm to the community is far worse when the deviant acts are committed by those who are in-group members than when they are committed by outsiders. The acts are viewed not only as a violation of rules, but also as an explicit rejection of the norms and values by one who is required by group membership to adhere to them.").

[162] Jan-Willem van Prooijen & Jerôme Lam, *Retributive Justice and Social Categorizations: The Perceived Fairness of Punishment Depends on Intergroup Status*, 37 EUR. J. SOC. PSYCHOL. 1244, 1246 (2007); *see also* HOFFMAN, *supra* note 10, at 32 (noting that scholars have found that

D. Altruistic Punishment and Group Maintenance

In the case of punishment of third-party violators, researchers have proposed that the feeling of anger induces retribution regardless of the cost to the punisher, because the punishment serves to deter future transgressions against all group members, including kin of the punisher and the punisher him- or herself.[163]

Vidmar asserts that retributive reactions reinforce group cohesion and solidify support for the legitimacy of group norms among not only the offenders, but also the non-offending members of the social group.[164] Vidmar writes:

> We have only to notice what happens, particularly in a small town, when some moral scandal has just been committed. [Community members] stop each other on the street, they visit each other, they seem to come together to talk of the event and to wax indignant in common. From all the similar impressions [which] are exchanged, for all the temper that gets expressed, there emerges a unique [public] temper.[165]

Similarly, anthropologists Thomas and Znaniecki concluded that in Polish peasant communities, punishment served the important purpose of obtaining consensus about the rules that were violated.[166] Other examples of the notion that moral offenses garner public outrage and cohesion around a shared desire for revenge include the 1995 bombing of the Murrah Federal Building in Oklahoma City and the 9/11 attacks on the World Trade Center in New York City.[167] Importantly, in the case of the Oklahoma City bombing, citizens of that city demonstrated stronger reactions and calls for justice than did members of other Oklahoma communities,[168] suggesting that violations within one's own community garner special attention and

---

human-on-human violence has been in steady decline as human groups have become larger and the number of outsiders correspondingly smaller). According to Hoffman, forensic archeologists estimate that "before agriculture, when [humans] were still living in relatively small nomadic tribes and regularly clashing with other tribes, 15 percent of [humans] died violent deaths;" however, after agriculture, there was "a 3 percent violent death rate" among humans. *Id.*

[163] Rob M.A. Nelissen & Marcel Zeelenberg, *Moral Emotions as Determinants of Third-Party Punishment: Anger, Guilt, and the Functions of Altruistic Sanctions*, 4 JUDGMENT & DECISION MAKING 543, 544 (2009).

[164] Vidmar, *supra* note 134, at 293.

[165] *Id.* at 293–94 (quoting EMILE DURKHEIM, THE DIVISION OF LABOR IN SOCIETY 102 (George Simpson trans., Free Press ed. 1960)).

[166] 2 WILLIAM I. THOMAS & FLORIAN ZNANIECKI, THE POLISH PEASANT IN EUROPE AND AMERICA 1250–55 (1958).

[167] *See, e.g.*, Christina A. Studebaker et al., *Assessing Pretrial Publicity Effects: Integrating Content Analytic Results*, 24 L. & HUM. BEHAV. 317, 323, 330–31, 333 (2000).

[168] *Id.* at 323, 331.

the highest degree of moral outrage. This seems adaptively correct. It is deviant members of one's own community who pose the biggest existential threat.[169]

IV. SCIENTIFIC EVIDENCE OF RETRIBUTION EVOLUTION

A number of areas of inquiry dovetail nicely with the evolutionary psychology explanation of retribution. Evidence suggests that cooperative punishment instincts are hardwired, resistant to change, and not explained by logic or contemporary goals.

A. Brain Imaging

Brain imaging reinforces the primitive nature of the retribution instinct.[170] The brain reacts differently to situations that have moral implications versus situations that do not implicate a moral choice. "Compared to moral scenarios involving only unintentional harm, moral scenarios involving intentional harm elicit more activity in areas associated with emotion (orbitofrontal cortex and temporal pole) and less activity in areas associated with cognition (including the angular gyrus and superior frontal gyrus)."[171] Imaging work has shown that "administering punishment to a transgressor" is a pleasurable act that "activates reward centers in the brain."[172] Research suggests that the positive feeling associated with punishing exists regardless of whether the victim of the transgression administers the punishment or a third-

[169] "Judge Matsch of the U.S. District Court concluded that 'the entire state had become a unified community, sharing the emotional trauma of those who had become directly victimized' . . . . Denver, Colorado, further removed from the death, destruction, and personal knowledge of the victims, had the lowest level of reaction of the cities studied." Vidmar, *supra* note 134, at 298. In deciding to move the trial from Oklahoma, the court concluded that "Oklahomans were 'united as a family,' that there was 'extraordinary provocation of their emotions of anger and vengeance,' that there was 'a prevailing belief that some action must be taken to make things right again,' and that the common reference in articulating these feelings was 'seeing that justice is done.'" *Id.* at 298 n.3; *see* United States v. McVeigh, 923 F. Supp. 1310 (D. Colo. 1996).

[170] *See* HOFFMAN, *supra* note 10, at 126–28 (noting that retaliation and revenge are common human behaviors that have cellular roots that go back to the first life forms). The immunological response of retaliation and revenge is the most widespread of all systems of retaliation. *Id.* at 127. There is an important difference in humans between revenge and retaliation: revenge may connote a more sophisticated, cognitive, kind of planning, while retaliation might signal a more automatic, emotional, reaction. *Id.* at 128. In social species, the most common kind of retaliation is a simple refusal to reciprocate (altruistic punishment)—meaning, "[i]f you don't cooperate with me, I won't cooperate with you." *Id.*

[171] Jana Schaich Borg et al., *Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation*, 18 J. COGNITIVE NEUROSCIENCE 803, 803 (2006).

[172] Darley, *supra* note 17, at 14.

party witness of the transgression administers it.[173] The experience of feeling outrage and the instinctive desire for revenge can be understood as a byproduct of a basic need to take action when a transgressor threatens the individual or community.[174] The question is whether this remains a legitimate basis for state-sanctioned punishment. Like our ancient ancestors, emotion continues to be an important and powerful part of the human experience, but in certain contexts it no longer serves the need it originally served. The feeling of moral outrage may still prompt us to take protective action, and so may be adaptive on the individual level, but the host of unintended consequences stemming from overly punitive practices might counsel against using retribution as a basis for punishment.

One theory for why human societies have not abandoned retribution as a legitimate basis for punishment is because, as a species, we have created post-hoc justifications for imposing pain. Although retributivist goals come from an automatically triggered, intuitive, and emotional response, they are often not described that way. Instead, the human desire for just deserts is treated as if it originates in effortful reasoning. Without question, human reasoning ability is unique in the animal kingdom, but we share instinctive automatic reactions with other animal species.[175] Yet our post-hoc rationalizations have deep roots in human history. As mentioned above, complex justifications for the appeal of a just-deserts model has existed since the time of the Ancient Greek philosophers.[176] Today, renowned thinkers continue

---

[173] *Id.* at 14.

[174] Miller, *supra* note 20, at 535.

[175] Margolis divides decision-making into two processes, the "seeing-that" (unconscious pattern matching—intuition, lower animals) and the "reasoning-why" (effortful, conscious, higher animals). *See* Jonathan Haidt, *Reasons Matter (When Intuitions Don't Object)*, N.Y. TIMES (Oct. 7, 2012, 5:00PM), https://opinionator.blogs.nytimes.com/2012/10/07/reasons-matter-when-intuitions-dont-object/; *see also* de Waal, *supra* note 11, at 4 ("[T]here never was a point at which we became social: descended from highly social ancestors, the monkeys and apes, we have been group-living forever."). Further, de Waal points out that humans "come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy" and that "[h]aving companions offers advantages in locating food and avoiding predators." *Id.* at 5. De Waal argues that humans have a social nature and illustrates this by pointing out that "second to the death penalty, solitary confinement is the most extreme punishment we can think of." *Id. See also generally* HOWARD MARGOLIS, PATTERNS, THINKING, AND COGNITION: A THEORY OF JUDGMENT (1987).

[176] *See* de Waal, *supra* note 11, at 13–14 (referencing a long tradition, reaching "back to Aristotle and Thomas Aquinas, which firmly anchors morality in the natural inclinations" of our species; rather than being considered "the antithesis of rationality, emotions aid human reasoning."); *see also* Hopt v. Utah, 110 U.S. 574, 579 (1884) ("The great end of punishment is not the expiation or atonement of the offense committed, but the prevention of future offences of the same kind."); Gregg v. Georgia, 428 U.S. 153, 183 (1976) (plurality opinion) ("'Retribution is no longer the dominant objective of the criminal law,' but neither is it a forbidden objective nor one inconsistent with our respect for the dignity of men.") (citing Williams v. New York, 337 U.S. 241, 248 (1949)).

to justify the retributivist model.[177] Moreover, the proposal that humans are guided by ancient instincts is often met with skepticism, and even anger. For example, sociobiologist Edward O. Wilson created significant waves when he published *Sociobiology: The New Synthesis* in 1975.[178] In this book, he proposed that all animal behavior—including human—should be understood as having evolved to benefit the group.[179] The work gained a great deal of attention, both from supporters and from detractors. A fundamental objection to Wilson's theory was that it characterized some of human explanation for our own behavior as made up to explain behavior that was, in essence, evolutionarily determined.[180] Wilson suggested that humans experience a feeling, behave in a particular way as a result, and then create a reason for the behavior that does not account for the instinct that actually created the behavior.[181] Wilson agreed with Hume, and claimed that moral philosophers were just creating justifications after consulting with their own brains.[182] Supporters pointed to evidence that many features of human behavior can be predicted by considering what behaviors increase cooperation and success of the social group. Critics were unhappy with what they saw as biological determinism.[183] However, Wilson himself never expressly advocated for biological determinism, instead opining that human behavior was profoundly shaped by evolutionary forces, but that neverthe-

---

[177] Some scholars assert that retributivism is no longer looked down upon as a method of punishment, but seems to be ascending and has replaced rehabilitation as the conventional justification for the amount of punishment a person should receive. Dolinko, *supra* note 39, at 537–38 (also asserting that America's unique affection for the death penalty is driven by retribution).

[178] *See* EDWARD O. WILSON, SOCIOBIOLOGY: THE NEW SYNTHESIS (2000).

[179] *Id.* at 117; *see also* 1 CHARLES DARWIN, THE DESCENT OF MAN, AND SELECTION IN RELATION TO SEX 166 (1871) ("[A]lthough a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe . . . an advancement in the standard of morality . . . will certainly give an immense advantage to one tribe over another. . . . [A] tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection. At all times throughout the world tribes have supplanted other tribes; and as morality is one element in their success, the standard of morality and the number of well-endowed men will thus everywhere tend to rise and increase.").

[180] *See generally Why You Do What You Do*, TIME (Aug. 1, 1977), http:/www.time.com/time/printout/0,8816,915181,00.html.

[181] WILSON, *supra* note 178, at 22–23.

[182] JONATHAN HAIDT, THE RIGHTEOUS MIND: WHY GOOD PEOPLE ARE DIVIDED BY POLITICS AND RELIGION 38 (Vintage Books 2013).

[183] *E.g.*, R.C. Lewontin, *The Fallacy of Biological Determinism*, SCIENCES, Mar./Apr. 1976, at 6; Sociobiology Study Grp. of Sci. for the People, *Sociobiology—Another Biological Determinism*, 26 BIOSCIENCE 182, 184 (1976).

less, human beings continue to have free choice. Supporters from a variety of disciplines—including sociology, biology, and psychology—concurred with Wilson's theory, even if they occasionally quibbled with the evidence.[184]

Since the publication of Wilson's controversial book, behavioral researchers have found support for his theory. Psychologist Jonathan Haidt and colleagues, for example, have gathered data suggesting that emotion-intuition comes first, and reasoning comes second.[185] Haidt's work involves something he calls "moral dumbfounding," in which he creates stories involving people engaging in behaviors that are traditionally taboo, but where there is no cost or resulting harm. In one story, for example, he told the story of a family who cooked and ate their dog after it had been killed by a car. It was made clear to the participants that no harm came to any of the family members and that nobody saw them engage in this behavior.[186] Haidt and colleagues found that respondents' affective reactions to the stories (statements that it would bother them to witness the action) were better predictors of their moral judgments than were their claims about harmful consequences.[187] Haidt found that while affective reactions were good predictors of how an individual would judge an action, perceptions of harmfulness were not. Although participants consistently said that the actions were "wrong," they also reported that the act had not caused harm. When Haidt and colleagues asked participants why the acts described were wrong, participants often made up harms that the experimenters had expressly precluded.[188]

---

[184] Psychologist David Barash welcomed Wilson's theory, and said publicly that psychology should take more seriously the role of evolution and selection in the study of human behavior. *See* David P. Barash, *The New Synthesis*, 1 WILSON Q. 108, 109–19 (1977).

[185] Jonathan Haidt et al., *Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?*, 65 J. PERSONALITY & SOC. PSYCHOL. 613, 613 (1993). Haidt has discussed Wilson's work with approval in his own publications. *See* HAIDT, *supra* note 182, at 38 (noting that Wilson "deserves to be called a prophet of moral psychology"); *see also* de Waal, *supra* note 11, at 6 (noting that psychological research suggests that human behavior derives above all from fast, automated emotional judgments and only secondarily from slower conscious processes). Humans, de Waal points out, seem about as emotional in their dealing with each other as any other social animal. *Id.*

[186] Haidt et al., *supra* note 185, at 617.

[187] *Id.* at 613.

[188] *Id.* at 625–26. Even when Haidt removed all signs of a victim, the respondents try to create victims. *Id.* Haidt concludes that the reasons people gave for why an action was wrong was a post hoc rationalization. *Id.* In other words, they had an emotional reaction *first* that did not make sense or could not be rationally explained, and then they created a reason second. *Id.*; *see also* Jordan & Rand, *supra* note 8, at 84 (concluding that while moral outrage is a private and genuine response to wrongdoing, the experience of and drive to report moral outrage also tracks the reputational benefits a person gains from punishing; thus supporting the theory that emotions are "adaptive motivators of action" and "moral outrage specifically is a motivator of punishment"). Moral outrage "appears to track the potential reputation value of punishment even when reputation is not at stake"—meaning that moral emotions and judgments can "misfire" in contexts where they are not adaptive. *Id.*

When the experimenters reminded the participants that no harm resulted, the participants generally appeared to be morally dumbfounded—they continued to maintain that the actions were wrong, but they were unable to offer reasons for why.[189] Findings from these studies suggest that moral judgment is caused by immediate intuitive emotional reactions, and is followed by effortful, ex post facto moral reasoning.[190]

B. Game Research and Altruistic Punishing Behavior

More evidence that moral outrage has evolved to assure cooperation comes from research involving cooperative behavior and games. In an experimental setting, when respondents receive a scenario in which some person commits a moral transgression, they experience a feeling of moral outrage, which is a substantial predictor of the punishments that will be assigned to the transgressor. One suggestion is that this "feeling" of moral outrage is the conscious registration of the intuitive reaction to instances of moral wrongdoing.[191] Participants will inflict punishment, even when it is costly for them to do so, and even when there is no conceivable pay off to them in terms of eliciting future cooperation from a current offender. The game is arranged so that the participants will never play against each other again and par-

---

[189] HAIDT, *supra* note 182, at 58–59.

[190] Other work on brain imaging has also supported the idea that the instinct, intuition, or emotion comes first, and the awareness and explaining comes second. For example, in one study, researchers imaged participants' brains to study the timing of brain activity involved with motor movement. The researchers asked participants to sit in a chair and simply to move their arm when they felt like it. The participants were asked to signal when they had decided to move. Researchers discovered that participants' brains "decided" to move before the participants experienced the conscious will to move, indicating that there was physiological change prior to conscious awareness. Benjamin Libet et al., *Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act*, 106 BRAIN 623, 623, 625 (1983). Nonconscious priming is another example of human decision-making affected by factors outside of an individual's awareness. *See, e.g.*, Sheila T. Murphy et al., *Additivity of Nonconscious Affect: Combined Effects of Priming and Exposure*, 69 J. PERSONALITY & SOC. PSYCHOL. 589, 589 (1995). Other examples of unconscious influences on decision-making are discussed *infra* in Part V.

[191] *See* Daniel Kahneman, *A Perspective on Judgment and Choice: Mapping Bounded Rationality*, 58 AMERICAN PSYCHOLOGIST 697, 710–12 (2003) (noting that the initial punitive decisions of jurors are likely to reflect an outrage heuristic, according to his research, but that jurors can also be instructed to consider other factors to influence their decision making); *see also* Petersen, *supra* note 104, at 63–64 (finding, "[i]n sum, these analyses support the motivational outputs of the deservingness heuristic being narrowly focused on (1) investing in reciprocators through compassion and (2) recalibrating the behavior of cheaters through anger").

ticipants are assured that all parties in the experiment will be kept in a state of anonymity from each other.[192] Researchers refer to this phenomenon as "altruistic punishment."[193] In 2002, Fehr and Gächter published a study on altruistic punishment in humans.[194] They showed that, in an experimental game, participants punish defectors in a cooperative activity although the punishment was costly for them and yielded no personal benefit. The adaptive function of punishment in one-shot interactions is to deter future harms perpetrated against third parties, and thus assure group cohesion and cooperation beyond one's own immediate experience.[195]

The typical "ultimatum" game paradigm is simple: one member of a dyad receives some amount of money and is told that she may allocate it however she chooses between herself and her partner. Her partner, in turn, may choose either to accept what is offered, or to reject it. If the partner rejects what is offered, then neither member of the dyad is allowed to keep any of the money. From a rational choice perspective, the partner should accept any amount offered, since she will end with more than she would get otherwise. However, research shows that this is not what happens.[196] One team of researchers administered the ultimatum game in many different cultures.[197] Participants in all cultures were increasingly less likely to accept the offer as it diminished from an even split of the resources.[198] In another study involving the ultimatum game conducted in the United States, responders' brains were imaged.[199] When the responder received an offer that allocated 80% to the decider and 20% or less to the responder, increased activity in a brain area that registers negative emotions was observed, and the degree of activity in that area predicted an increased tendency for the responder to punish the offending decider by rejecting the offer, thus depriving the decider of any gains.[200] Researchers studying the ultimatum game explained that "[t]aken together, our findings suggest a prominent role of the caudate nucleus, with possible contributions of the thalamus, in

---

[192] Fehr & Fischbacher, *supra* note 136, at 185–88.

[193] Fehr & Gächter, *supra* note 136, at 137.

[194] *Id.*

[195] *See id.* at 139; Herbert Gintis et al., *Explaining Altruistic Behavior in Humans*, 24 EVOLUTION & HUM. BEHAV. 153, 153–54, 161–62 (2003).

[196] John M. Darley, *Citizens' Assignments of Punishments for Moral Transgressions: A Case Study in the Psychology of Punishment*, 8 OHIO ST. J. CRIM. L. 101, 109 (2010).

[197] *See* Joseph Henrich et al., *Costly Punishment Across Human Societies*, 312 SCIENCE 1767, 1768 (2006).

[198] *Id.* at 1769.

[199] Dominique J.-F. de Quervain et al., *The Neural Basis of Altruistic Punishment*, 305 SCIENCE 1254, 1255 (2004) (discussing image while subjects were engaged in an experimental game); *see also* Alan G. Sanfey et al., *The Neural Basis of Economic Decision-Making in the Ultimatum Game*, 300 SCIENCE 1755, 1755 (2003) (discussing the same).

[200] *See* Sanfey et al., *supra* note 199, at 1755–57.

processing rewards associated with the satisfaction of the desire to punish the intentional abuse of trust."[201]

Not only will participants forgo gains to punish an uncooperative game-player, they will *pay* to punish. This has been found even where there is no expectation that a future encounter will occur, so there is no "teaching" value to the punishment as far as the participant is concerned. So, in one study, a participant observed another participant behave in an untrustworthy way toward a third participant who could not inflict any punishment on the bad actor.[202] The observing participant frequently took the opportunity that the game rules afforded him to punish the untrustworthy participant, expending his points to purchase a fine that was inflicted on the defector.[203] This latest result, demonstrating that a third party is sometimes willing to expend resources to punish a person, is an example of "altruistic punishment."[204] In these situations, there are no material gains to the punisher and often costs incurred. These facts make the actions of the punisher "irrational."

The studies on cooperation, defection, and punishment in game experiments reveal that a drive to punish exists even when the punisher has no relationship with the victim and no expectation that she will ever interact with the transgressor.[205] Moral outrage causes individuals to punish when there is no plausible utilitarian rationale for the punishment. These scenarios are ones in which there is no utilitarian rationale for the punishment *today*, but there is an excellent evolutionary reason for the drive to punish. A similarly intuitive and automatic desire has been shown to exist to reward: psychological research supports the idea that when deciding on

---

[201] de Quervain et al., *supra* note 199, at 1256; *see also* Darley, *supra* note 17, at 11 ("The conclusion here is that humans find punishing norm violations in these experimental games to be a rewarding activity and are willing to spend resources to do so. No similar brain activity pattern was found when the punishment administered was only symbolic. Rewarding punishment needs to inflict actual pain.").

[202] Ernst Fehr & Urs Fischbacher, *Third-Party Punishment and Social Norms*, 25 EVOLUTION & HUM. BEHAV. 63, 65 (2004); *see also* Darley, *supra* note 17, at 11 (discussing Fehr & Fischbacher's study).

[203] Fehr & Fischbacher *supra* note 202, at 85; *see also* Daniel Kahneman et al., *Fairness and the Assumptions of Economics*, 59 J. BUS. S285, S290–91 (1986) (noting that third-party witnesses who saw a split of $18.00 kept by the decider and only $2.00 given to the receiver could punish the decider who inflicted that unfair outcome with a fine of $5.00, by paying a cost of $1.00—in such cases, 74% of third-party witnesses chose to do so).

[204] Weng et al., *supra* note 142, at 2.

[205] *See id.* at 3 (discussing how punishment could be motivated by compassion to help the victim, the perpetrator, or both: "By deciding to punish, third parties may help protect future potential victims and also provide valuable feedback to the transgressors regarding the social acceptability of their behavior."). *But see* Peter Duersch & Julia Müller, *Taking Punishment into Your Own Hands: An Experiment*, 46 J. ECON. PSYCHOL. 1 (2015) (finding that following an unfair decision, experiment subjects bid positive amounts for the right to personally punish the decider, and are happier if they get to punish personally).

whether to donate resources to an individual or a group, people first determine whether the recipients would be good cooperation partners.[206]

C. Retribution as a Specialized Behavioral Response

As mentioned earlier, retribution elevates punishment to a primary objective. In contrast to punishment with a goal of achieving deterrence or making a victim feel more secure, retributive punishment is characterized as a necessary or deserved response.[207] According to evolutionary psychologists, this motivational structure is more compatible with punishment as a specialized behavioral response.[208] A behavioral response model of punishment and prosociality suggests that this response is supported by specific behavioral adaptations (rather than by general learning processes). In other words, this behavioral response is a function of evolutionary adaptation. The behavioral response paradigm can also provide us with clues about how the psychological profile of retributive punishment affects the functional design. If human beings needed punishment to be uniform, inflexible, and consistent because behavior modification is critical, then an inflexible approach to punishment is desirable. A built-in, automatic affective response that is ubiquitous, or nearly so, is best able to achieve the necessary behavior control—this is precisely how moral outrage functions.[209]

The built-in response to moral infractions is so powerful that hearing about a single norm violation can prime an individual to look for future opportunities to exact revenge. When people perceive that violations have gone unpunished, the instinct to impose a punishment becomes stronger. The strengthening of the punishment instinct makes sense from the perspective of needing to enforce reciprocity and group norms.[210] Phil Tetlock and colleagues have found that if people are in-

---

[206] Petersen, *supra* note 104, at 60–61.

[207] *See* Dolinko, *supra* note 39, at 539 (noting that retributivists believe that "the crime must be nullified, that the criminal must pay his debt to society (or . . . that society must pay him back), that the wrongdoer has in some sense willed his own punishment").

[208] Cushman, *supra* note 153, at 337. Importantly, the ubiquitous nature of retributive impulses—and the consistency with which human beings feel moral outrage and react uniformly—relates to the desired outcome, and how we evolved to meet these goals. *See id.* at 342, 348.

[209] *See id.* at 342 ("To put it another way, punishment is a mechanism that exploits general learning processes; it gets social partners to adopt prosocial behavior roughly by operant conditioning. This sets up a clear prediction about the functional design of punishment. Punishment should be designed to match the constraints of general learning processes, obtaining the maximum response from social partners at the minimum cost.").

[210] *See* Michael J. Sargent, *Less Thought, More Punishment: Need for Cognition Predicts Support for Punitive Responses to Crime*, 30 PERSONALITY & SOC. PSYCHOL. BULL. 1485, 1485 (2004).

formed of prior wrongdoing that went unpunished, they behave as "intuitive prosecutors" when responding to subsequent wrongdoing by other perpetrators.[211] Interestingly, they become primed to punish, more sensitive to factors that could increase culpability, and less sympathetic to mitigating factors.[212]

D. Punishment, Forgiveness, & Reconciliation

Early humans needed to develop an automatic, universal mechanism to assure reinforcement of norms of cooperation, but they also needed a mechanism for reintegrating the wayward society member.[213] Importantly, research has found this connection independent of explicit utilitarian goals of curbing bad behavior in the future.[214] Even the simple act of priming participants with the idea of an opportunity to punish (as opposed to priming an inability to punish) resulted in a greater rate of forgiving.[215] It appears that the acceptance of members who have transgressed is, at least in part, dependent upon victims being able to get their just deserts. The authors conclude that "punishment plays a crucial role in regulating human behaviours and coexistence—yet the individual, interpersonal and group benefits of a contrary response, forgiveness, are also well established."[216] The authors also focus on the importance of the affective aspect of punishment, concluding that "[s]eeing offenders suffer for their actions helps victims feel better."[217] Some researchers argue that a more nuanced view of the just deserts instinct would involve consideration of the important goal of reforming the transgressor—of not only communicating disapproval of the offensive behavior, and not only receiving confirmation of the transgressor's understanding of the moral indignation, but also knowledge that the transgressor is repentant, has learned her lesson, and is reformed for the future.[218]

---

[211] Philip E. Tetlock, *Social Functionalist Frameworks for Judgment and Choice: Intuitive Politicians, Theologians, and Prosecutors*, 109 PSYCHOL. REV. 451, 454 (2002); *see also* Julie H. Goldberg et al., *Rage and Reason: The Psychology of the Intuitive Prosecutor*, 29 EURO. J. SOC. PSYCHOL. 781, 782 (1999).

[212] Tetlock, *supra* note 211, at 463.

[213] Maintaining a community of a certain size requires forgiveness, lest elimination of transgressors results in the dwindling of the group to maladaptively small numbers.

[214] *See* Peter Strelan & Jan-Willem van Prooijen, *Retribution and Forgiveness: The Healing Effects of Punishing for Just Deserts*, 43 EUR. J. SOC. PSYCHOL. 544, 544 (2013).

[215] *Id.*

[216] *Id.* at 545 (providing, as an example, MICHAEL E. MCCULLOUGH, BEYOND REVENGE: THE EVOLUTION OF THE FORGIVENESS INSTINCT (2008)).

[217] *Id.*; *see also* Arlene M. Stillwell et al., *We're All Victims Here: Toward a Psychology of Revenge*, 30 BASIC & APPLIED SOC. PSYCHOL. 253, 254 (2008).

[218] *See* McGeer & Funk, *supra* note 137, at 535, 538 (stating that "many theorists make the stronger case that a retributive [view of human psychology] has genuine adaptive value;" also, there is a premium on blaming reactions that "communicate a normatively loaded message to offenders and thereby initiate a process in them of reflective self-transformation").

Psychologists Peter Strelan and Jan-Willem van Prooijen hypothesized that victims seeking just deserts were able to forgive their offender after taking the opportunity to punish him or her. The authors concluded that when victims achieve justice by punishing the person who wronged them, they feel satiated, and this satisfaction in turn facilitates forgiveness.[219] Studies on the relationship between punishment and forgiveness tell us that when someone hears that a perpetrator—who has victimized them or a third party—received a consequence, they are then able to forgive.[220] Forgiveness is essential because it allows groups to maintain a steady population, in spite of the fact that the individuals in the population are not perfect, and may err. Once the instinct to correct the transgressor has been fulfilled, that individual can be restored to the group.[221] This would have been important as early humans were reliant on each other for survival. If humans did not develop the ability to restore the transgressor, and instead killed or ousted the individual, their groups may have dwindled to the point where the size of the group was no longer sufficiently large to survive. "In short, punishment plays a crucial role in regulating human behaviours and coexistence—yet the individual, interpersonal and group benefits of a contrary response, forgiveness, are also well established."[222]

"Apologies serve the dual purpose of helping to integrate the offender back into the group and helping to reaffirm the moral basis of the rule that was violated."[223] Kleinke, Wallis, and Stalder conducted an experiment that found that a rapist's expression of remorse lessened recommended prison sentences.[224] In other words, a signal from the transgressor that indicates recognition of the rules created by the social pact and acknowledgment of his violation lessens the need for punishment, and therefore lessens the sense of moral outrage. After all, if the primary purpose of the feeling of outrage is to serve as a check on deviant behavior, then acceptance of fault and communication that the transgressor has learned not to repeat the behavior makes the punishment less necessary. Individuals are influenced not only by formal, spoken apologies, but also by signals that individuals feel regret for their actions, suggesting that they are less likely to behave similarly in the future. For example, in a review of studies of defendants convicted of homicide, Sundby found that jurors

---

[219] Strelan & van Prooijen, *supra* note 214, at 545, 550.

[220] *Id.*

[221] *See id.* at 551 (finding that the idea of restoration is an important one that played a critical role in the preservation of early human communities).

[222] *Id.* at 545; *see also* Jordan & Rand, *supra* note 8, at 85 ("Third party punishment is central to human morality, and plays a key role in promoting cooperation.").

[223] Vidmar, *supra* note 134, at 304.

[224] Chris L. Kleinke et al., *Evaluation of a Rapist as a Function of Expressed Intent and Remorse*, 132 J. SOC. PSYCHOL. 525, 525 (1992).

reported that in almost all cases resulting in a death verdict or a life sentence, defendants exhibited a lack of emotion during the trial.[225]

E. Cultural Differences in Moral Reactions

The processes individuals undergo to determine whether others deserve punishment or help has the character of being automatic and based upon universal kinds of factors.[226] However, specific features of the situation or context vary, depending upon specific cultural norms. An act that is seen as a serious violation in one culture may be viewed positively or with tolerance in another. For example, violent efforts to control certain members of society, such as married women and children, are sanctioned in some societies, even while they are criminalized in others.[227] Less extreme examples of differing norms exist within a single society, among members of

---

[225] Scott E. Sundby, *The Capital Jury and Absolution: The Intersection of Trial Strategy, Remorse, and the Death Penalty*, 83 CORNELL L. REV. 1557, 1564–65 (1998).

[226] Janice Nadler & Mary R. Rose, *Victim Impact Testimony and the Psychology of Punishment*, 88 CORNELL L. REV. 419, 423–25 (2003). Empirical research on the psychology of justice supports the idea that people's punishment judgments are guided to a large degree by harm-based, retributive psychology, including the ideas that punishment judgments track the harm caused by the crime, the severity of the crime, and perceived deservingness of the criminal for the crime they committed. *Id.* at 423–24. "[H]arm is a critical factor in people's views about just punishments, and harm is not rendered irrelevant simply because it is adventitious or unforeseen." *Id.* at 425. Human judgments of the moral wrongness of a behavior rely on the agent's mental state, while human judgments of deserved punishment show greater sensitivity to the harm actually caused by the agent. Fiery Cushman, *Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment*, 108 COGNITION 353, 355 (2008). Harmful intentions alone are found to be sufficient to warrant moral punishment, even in the absence of any harmful consequence. *Id.* at 360. The effects of belief, desire, and consequences on judgments of blame are generally additive. *Id.* at 360–61.

[227] For instance, "honor crimes" are "acts of violence committed against female family members who are perceived to have brought shame to the family by engaging in dishonorable acts," such as "premarital sex, adultery, pregnancy out of wedlock, and even mere contact with a man who is not a relative." *See* Manuel Eisner & Lana Ghuneim, *Honor Killing Attitudes Amongst Adolescents in Amman, Jordan*, 39 AGGRESSIVE BEHAV. 405, 405–06 (2013). Honor crimes receive normative support in their culture—they are "often deliberate, committed collectively by members of a family, condoned by the community, and treated leniently by the criminal justice system." *Id.* at 405. Controversial international examples include an Italian penal code which permitted reduced sentences for men who killed their adulterous wives, daughters, or sisters until 1981, and provisions of the Jordanian Code which until 2001, sanctioned a man to a range of three months to two years in prison for killing his wife or female relative after catching her in an "un-lawful bed." *Id.* at 406. For a discussion of instances where women participate in honor killings, see Phyllis Chesler, *When Women Commit Honor Killings*, MIDDLE EAST Q., Fall 2015, 1, 1–11.

sub-cultures. Certain sub-populations have adopted vastly different moral intuitions based upon different codes.[228]

These moral intuitions are just as ingrained as the majority of moral codes.[229] For example, in certain parts of the United States, the flying of a Confederate flag is viewed as a symbol of intolerance and bigotry, while in other parts of the country, the flying of the same flag may be tolerated or even commended as a symbol of pride in southern heritage.[230] Behavior that is viewed as socially undesirable by certain sub-cultures in the United States may afford others "street credibility," allowing for safety and status in inner-city neighborhoods, where appearing vulnerable can exact a variety of costs.[231] Similarly, cultural norms can shape how one views the violation of a law, and whether a particular violation is taboo or accepted as business as usual.[232] Cultural differences can also shape how life events are interpreted. To the extent that judgments about intent, consequences, and behavior—or behavioral responses to triggers—are relevant to a legal excuse or justification, cultural norms

---

[228] *E.g.*, Donald B. Kraybill, *Why the Amish Forgive So Quickly*, CHRISTIAN SCI. MONITOR (Oct. 2, 2007), https://www.csmonitor.com/2007/1002/p09s02-coop.html.

[229] John Darley, *Realism on Change in Moral Intuitions*, 77 U. CHI. L. REV. 1643, 1650 (2010).

[230] *See* Scott H. Huffmon et al., *Down with the Southern Cross: Opinions on the Confederate Battle Flag in South Carolina*, 132 POL. SCI. Q. 719, 721–24 (2017). For a discussion of the role of the Confederate flag in American culture, see Frances Stead Sellers, *The Confederate Flag: A 150 Year Battle*, WASH. POST (Oct. 23, 2018, 10:58 AM), https://www.washingtonpost.com/politics/the-confederate-flag-a-150-year-battle/2018/10/23/622ae7e2-d179-11e8-83d6-291fcead2ab1_story.html?utm_term=.8af653c0b8cb; *see also* Jessica Owley et al., *Private Confederate Monuments*, 25 LEWIS & CLARK L. REV. (forthcoming Feb. 2021) (discussing removal and relocation of Confederate monuments).

[231] *See* Joseph B. Richardson et al., *Pathways to Early Violent Death: The Voices of Serious Violent Youth Offenders*, AM. J. PUB. HEALTH e1, e2–e3 (2013) (arguing that a high rate of "poverty, joblessness, violence, alienation, lack of faith in the police and the judicial system, and hopelessness have produced a neighborhood street culture 'code' that influences how individuals . . . negotiate interpersonal violence.") The code of the street is a "set of informal rules that govern interpersonal public behavior, including violence;" street-oriented people establish and enforce the rules, but their norms oppose mainstream values and everyone must "know the rules or suffer the consequences." *Id.*; *see also* Desmond Upton Patton et al., *Sticks, Stones, and Facebook Accounts: What Violence Outreach Workers Know About Social Media and Urban-Based Gang Violence in Chicago*, 65 COMPUTERS HUM. BEHAV. 591, 598 (2016) (discussing examples of gang members using Facebook accounts and posts to develop or boost their street credibility). *See generally* ELIJAH ANDERSON, CODE OF THE STREET: DECENCY, VIOLENCE, AND THE MORAL LIFE OF THE INNER CITY (2000).

[232] *See* Derek D. Rucker et al., *On the Assignment of Punishment: The Impact of General-Societal Threat and the Moderating Role of Severity*, 30 PERSONALITY & SOC. PSYCHOL. BULL. 673, 673 (2004) (showing that when participants feel that the social order is threatened, they behave more punitively toward a perpetrator of a moderate-level crime).

may explain varying interpretations.[233] Research has shown that when citizens do not know what the law is, they infer what it is by referencing the majoritarian norm.[234] The importance of social agreement on norms and the adherence of individuals to these norms in determining whether to inflict pain on another person in the social group is, in and of itself, evidence supporting the idea that the function of punishment is group coherence and cooperation.

Regardless of the particular manifestation of the retributivist instinct, it appears in all cultures. The very proliferation of a range of manifestations of Heider's "ought force" suggests the innate humanness of it. However, as mentioned earlier in this piece and discussed in detail below, it also shares commonalities with identified biases that similarly pervade every human culture.

V. RETRIBUTION AS MENTAL HEURISTIC

> If human punishment is driven largely by retribution, and retribution is arrived at by way of heuristic judgments, we are forced to ask whether we can trust such judgments. Sometimes heuristic decision making serves us well. . . . At other times, however, even the most common heuristic inclinations can be plainly wrong, as research in psychology has long warned. For example, research on infanticide behavior shows how impulses that might have benefitted our ancestors' individual fitness may not be useful to societal groups. Likewise, it remains a topic of continued debate as to whether and when retribution serves prosocial goals.[235]

The pervasiveness of the instinct, along with evidence from evolutionary biology, psychology, and anthropology, reveals the ancient adaptive nature of the retribution drive. At the same time, modern approaches to social problem-solving and the rise of rationalism and utilitarianism provide reason to question the legitimacy of this ancient heuristic.

Retributivist impulses have less in common with utilitarian methodologies than they do with other "fast and frugal" methods of evaluation and decision-making. Just as evolutionary pressures have resulted in adaptive emotional reactions, these pressures have also led to the development of social and cognitive heuristics and biases. A heuristic is a mental shortcut, or rule of thumb, that helps human

---

[233] Cultural differences have been used as a basis for a criminal defense. *See* James J. Sing, Note, *Culture as Sameness: Toward a Synthetic View of Provocation and Culture in the Criminal Law*, 108 YALE L.J. 1845, 1846 (1999); Sharon M. Tomao, Note, *The Cultural Defense: Traditional or Formal?*, 10 GEO. IMMIGR. L.J. 241, 241 (1996).

[234] John M. Darley et al., *The Ex Ante Function of the Criminal Law*, 35 L. & SOC. REV. 165, 165 (2001) (finding that residents of a state with minority positions on criminal codes indicated that their state followed the rule of the majority of jurisdictions).

[235] Aharoni & Fridlund, *supra* note 23, at 618 (internal citations omitted).

beings make quick decisions with limited information.[236] A bias is a predisposition to act or think in a particular way in response to a stimulus.[237] Heuristics and biases evolved to provide humans automatic rules allowing for quick decision-making when action was necessary for survival. As one article notes, cognition is "shaped by a host of demonstrable and perhaps nearly universal cognitive biases and heuristics, many or all of which are the product of evolutionary pressures or accidents."[238] An example of a heuristic is the tendency to run when an animal is charging (note the relationship between this heuristic for survival and the automaticity of the fight-or-flight response). The decision to run will, on balance, create the greatest likelihood for survival, although it is not always the best strategy. An example of a bias is the tendency for individuals to distrust and avoid people and situations that are unfamiliar or dissimilar to them. Although this distaste for unfamiliar people and situations was protective at some point, today, it can result in what we characterize as unwarranted prejudice against people on the basis of their race, religion, ethnicity, sexual orientation, or other characteristics that make them different from the individual evaluator.

The quick and automatic nature of much of human decision-making was first studied by Herbert Simon, a psychologist, sociologist, and political scientist who coined the term "bounded rationality."[239] The notion that people are boundedly rational (or "satisficers") refers to human beings' need to make decisions quickly, often with limited information, and with cognitive constraints. Simon and other proponents of bounded rationality, such as Nobel Prize-winning psychologist and economist Daniel Kahneman and his longtime collaborator, Amos Tversky, ques-

---

[236] Gerd Gigerenzer & Wolfgang Gaissmaier, *Heuristic Decision Making*, 62 ANN. REV. PSYCHOL. 451, 454 (2011).

[237] *See* Martie G. Haselton et al., *The Evolution of Cognitive Bias*, *in* 1 THE HANDBOOK OF EVOLUTIONARY PSYCHOLOGY 968, 968 (David M. Buss ed., 2d ed. 2016).

[238] Donald Braman et al., *Some Realism About Punishment Naturalism*, 77 U. CHI. L. REV. 1531, 1567 (2010) [hereinafter Braman et al., *Some Realism About Punishment Naturalism*]. *See generally* Donald Braman et al., *A Core of Agreement*, 77 U. CHI. L. REV. 1655 (2010) [hereinafter Braman et al., *A Core of Agreement*].

[239] Michael Mintrom, *Herbert A. Simon, Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, *in* THE OXFORD HANDBOOK OF CLASSICS IN PUBLIC POLICY AND ADMINISTRATION, at 1 (Martin Lodge et al. eds., 2016). Simon was a prolific scholar. *See generally* HERBERT A. SIMON, MODELS OF MAN, SOCIAL AND RATIONAL: MATHEMATICAL ESSAYS ON RATIONAL HUMAN BEHAVIOR IN A SOCIAL SETTING (1957); Herbert A. Simon, *A Behavioral Model of Rational Choice*, 69 Q.J. ECON. 99 (1955) [hereinafter Simon, *A Behavioral Model*]; Herbert A. Simon, *Human Nature in Politics: The Dialogue of Psychology with Political Science*, 79 AM. POL. SCI. REV. 293 (1985) (introducing the notion of "bounded rationality" to account for the fact that human beings have finite computational resources available for making choices).

tioned rational choice theory (RCT), which claims people behave rationally to maximize their own gains.[240] Whereas RCT assumes that individuals have perfect memories, no biases, and limitless capacity to process information, the empirical work that forms the basis for bounded rationality contradicts these assumptions. Studies in social and cognitive psychology reveal that people make broad generalizations, use cognitive shortcuts to arrive at educated guesses about the state of the world, and exhibit predictable biases in their interpretation of events, outcomes, and social information.[241] The body of empirical work demonstrating irrationality in human choice had reached a critical mass by the turn of the last century, and today, the conclusion that people are influenced by a wide range of unconscious processes is irrefutable. In short, social science research has revealed an extensive network of interrelated heuristics and biases that serve as the basis for much of human decision-making.[242] Moreover, while many of our automatic mental processes can serve a useful function,[243] many others have largely outlived their usefulness.[244] An important objective of our modern-day psychological study is to discover how and

[240] Thomas S. Ulen, *Rational Choice Theory in Law and Economics*, *in* 1 ENCYCLOPEDIA OF LAW AND ECONOMICS 790, 791–92 (Boudewijn Bouckaert & Gerrit De Geest, eds., 2000).

[241] *See generally* Simon, *A Behavioral Model*, *supra* note 239; BEHAVIORAL LAW AND ECONOMICS (Cass R. Sunstein ed., 2000); Daniel Kahneman & Amos Tversky, *Choices, Values, and Frames*, *in* CHOICES, VALUES, AND FRAMES (Daniel Kahneman & Amos Tversky eds., 2000) (discussing empirical investigations of how human beings process information and make choices).

[242] BEHAVIORAL LAW AND ECONOMICS, *supra* note 241, at 170. "Behavioral decision theory" or "behavioral law and economics" comprise findings on a range of heuristics and biases, including anchoring and adjustment, optimism bias, representativeness heuristic, hindsight bias, conjunction fallacy, endowment effect and related status quo bias, risk aversion, and the availability heuristic, in addition to others. Some of these features of human decision-making are discussed in Part IV, *supra*.

[243] Gerd Gigerenzer, a German psychologist, has spent most of his career examining the adaptive nature of mental shortcuts. Gigerenzer's claim is that the natural tendencies that manifest when human beings make decisions have evolved to allow human beings to function in a complex world in which complete consideration of all features of the decision task is impossible. Gigerenzer has devoted much of his work to defending human decision making, calling it "fast-and-frugal," and identifying conditions under which fast and automatic processes can lead to optimal decision-making in certain situations. *See* Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 ECONOMETRICA 263, 263–291 (1979); Daniel Kahneman & Amos Tversky, *Subjective Probability: A Judgment of Representativeness*, 3 COGNITIVE PSYCHOL. 430, 430–54 (1972). For some early law review pieces discussing heuristical processing and responses in legal frameworks, see Mark Kelman, *Moral Realism and the Heuristics Debate*, 5 J. LEGAL ANALYSIS 339, 347 (2013) (discussing the availability of representative heuristics); *see also* Barbara D. Underwood, *Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment*, 88 YALE L.J. 1408, 1428 (1979) ("[S]tudies show that in making individualized judgments people rely primarily on information about the case at hand, paying relatively little attention to background information about other cases.").

[244] *See* Haselton et al., *supra* note 237, at 979 (discussing biases related to disease transmission, sexuality, and other evolutionary processes, and noting that humans possess a bias

when these ancient cognitive and social intuitions prevent optimal choice formation so that we can thoughtfully address the predispositions that will lead to bad decisions and outcomes. The brief overview below describes some of the heuristics and biases that can lead to irrational decisions. These biases have, in the past, served an evolutionary function. Understanding their past usefulness provides glimpses into how the retributivist instinct was once adaptive, but now is of limited usefulness, and may even cause harm in predictable and important ways.

A. Cognitive Biases

To the extent that humans are intuitive retributivists, they must reconcile this instinct with their rational, sentient selves who tend to value utilitarian goals. This is nothing new. Social psychological literature is rife with examples in which people are inaccurate about the basis of their attitudes and behavior.[245] The large and growing literature on heuristics and biases (motivated reasoning) provides the ideal window into this type of process.

1. Framing

When people are making decisions, the way in which choices are presented impacts a decision-maker's preference.[246] This effect, called framing, leads individuals to pick one option over another specifically because of how the various options are ordered or described.[247] For example, physicians and patients prefer a particular course of treatment when they are told that 90% live through the postoperative

---

toward inferring that members of outgroups are less generous and kind, and more dangerous and ill-tempered.) This bias might have been adaptive for ancestral humans, because "the costs of falsely assuming peacefulness on the part of an aggressor were likely to outweigh the . . . low[er] cost of elevated vigilance." *Id.* Implicit bias is an example of a bias that may have kept ancient humans safe by encouraging interaction with in-group members who were more likely to cooperate and less likely to pose an existential threat. Today, implicit bias is condemned as manifesting in unconscious predispositions against out-group members. *Id.*; *see also* Ackerman et al., *supra* note 108, at 2–3 (discussing the behavioral immune system's role in preserving societies via unconscious bias against group members with heuristics of infection risks).

[245] Phoebe C. Ellsworth & Lee Ross, *Public Opinion and Capital Punishment: A Close Examination of the Views of Abolitionists and Retentionists*, 29 CRIME & DELINQ. 116, 140, 144 (1983); Richard E. Nisbett & Timothy DeCamp Wilson, *Telling More than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231, 231 (1977); Margaret Wilson, *Six Views of Embodied Cognition*, 9 PSYCHONOMIC BULL. & REV. 625, 625 (2002).

[246] *See* Eldar Shafir, *Prospect Theory and Political Analysis: A Psychological Perspective*, 13 POL. PSYCHOL. 311, 313–14 (1992) ("Framing refers to the tendency of normatively inconsequential changes in the formulation of a choice problem to affect the ways people represent the problem and, consequently, their preferences."). *See generally* Amos Tversky & Daniel Kahneman, *Rational Choice and the Framing of Decisions*, 59 J. BUS. S251 (1986). For early discussions of framing, see generally ERVING GOFFMAN, FRAME ANALYSIS: AN ESSAY ON THE ORGANIZATION OF EXPERIENCE (Ne. Univ. Press 1986) (1974).

[247] *See* Matthew Rabin, *Psychology and Economics*, 36 J. ECON. LITERATURE 11, 36 (1998).

period[248] than when they are told that 10% die during the postoperative period.[249] The fact that these two statements convey identical information makes the change in choice irrational.[250] Researchers Hanson and Kysar have noted that "framing effects are somewhat different from the other cognitive anomalies that have been identified by behavioral researchers. They are perhaps the most obviously exploitable of the biases, capable, for instance, of causing dramatic preference reversals based on an entirely nonsubstantive shift in terminology."[251] Human responses to framing are so profound because human beings are hardwired to be influenced by rhetorical shifts. Researchers have examined how the human brain has evolved to respond to frames by using brain imaging; some researchers have looked at the physiological basis for these responses and have discovered that the framing effect is associated with activity in the amygdala, an area of the brain responsible for processing emotions.[252] Their effects are so powerful that some conclude that framing strategies "can become freewheeling exercises in pure manipulation."[253]

2. Priming

Priming describes the situation in which early exposure to a stimulus sensitizes the subject to a later presentation of the same or a similar target.[254] Stimuli that have been primed will influence cognitive and emotional reactions to subsequent targets or events.[255] Put simply, priming increases cognitive accessibility, so concepts that

---

[248] *Id.* at 36–37.

[249] *Id.*; *see also* Donald A. Redelmeier et al., *Understanding Patients' Decisions: Cognitive and Emotional Perspectives*, 270 J. AM. MED. ASS'N 72, 73 (1993).

[250] *See* Kahneman & Tversky, *supra* note 241, at 2–4. Framing is described by prospect theory, which articulates two claims: first, individuals assign more significance to a loss than they do to an equivalent gain; second, people overweigh low probabilities and underweigh moderate and high probabilities. *Id.*

[251] *See* Jon D. Hanson & Douglas A. Kysar, *Taking Behavioralism Seriously: The Problem of Market Manipulation*, 74 N.Y.U. L. REV. 630, 684–85 (1999).

[252] Benedetto De Martino et al., *Frames, Biases, and Rational Decision-Making in the Human Brain*, 313 SCIENCE 684, 686 (2006).

[253] *See* Donald R. Kinder & Don Herzog, *Democratic Discussion*, *in* RECONSIDERING THE DEMOCRATIC PUBLIC 347, 363 (George E. Marcus & Russell L. Hanson eds., 1993); *see also* Jonathan Remy Nash, *Framing Effects and Regulatory Choice*, 82 NOTRE DAME L. REV. 313, 317 (2006) (calling framing "the ability of someone who is propounding an option to present the option . . . in such a way as to . . . make the option seem more or less desirable").

[254] *See* Robert B. Cialdini et al., *A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places*, 58 J. PERSONALITY & SOC. PSYCHOL. 1015, 1023 (1990) ("Most, although not all, explanations of priming effects incorporate the notion of spreading activation, which posits that similar concepts are linked together in memory within a network of nodes and that activation of one concept results in the spreading of the activation along the network to other related concepts." (citations omitted)).

[255] Sheila T. Murphy & R.B. Zajonc, *Affect, Cognition, and Awareness: Affective Priming with Optimal and Suboptimal Stimulus Exposures*, 64 J. PERSONALITY & SOC. PSYCH. 723, 735 (1993)

are most easily retrieved from the brain will have the greatest impact at the moment of choice.[256]

Priming is not purely concept based. Emotions can also be primed so that decision-makers who have been exposed to information that has caused a particular affective state will respond to subsequent information in a way that is consistent with the emotional state.[257] This response is particularly likely when the subsequent information is relevant to the earlier, emotion-triggering information. For example, viewing images of the aftermath of a violent crime makes people more likely to react with anger and condemnation to a criminal defendant, regardless of the relevance of the photos to the defendant's culpability.[258]

3. Cognitive Availability

Cognitive availability is closely related to priming. Cognitive availability relates to the tendency of individuals to overestimate the frequency of a situation based upon how easy it is to bring that situation or outcome to mind.[259] A commonly cited example is a plane crash. Plane crashes tend to be cognitively available because they are dramatic, catastrophic events.[260] When asked about the frequency of plane crashes, individuals search their memories and are readily able to come up with examples. The ease with which individuals can bring past examples to mind leads people to overestimate the probability of the event.[261]

4. Anchoring

Research has shown that when people are provided an initial value, they unconsciously "anchor" on that value and adjust away from it.[262] As a result, their estimation of a correct value is unduly influenced by whatever value they saw initially. This tendency to anchor is true even when they are *aware* that the initial value

---

(finding that millisecond-long encounters with negative or positive stimuli can produce nonspecific emotional reactions to unrelated stimuli).

[256] *See generally* Robert S. Wyer, Jr. & Thomas K. Srull, *Category Accessibility: Some Theoretical and Empirical Issues Concerning the Processing of Social Stimulus Information*, *in* 1 SOCIAL COGNITION: THE ONTARIO SYMPOSIUM 161 (E. Tory Higgins et al. eds., 1981).

[257] Murphy & Zajonc, *supra* note 255, at 735–36.

[258] *Cf.* Kathryn M. Stanchi, *The Power of Priming in Legal Advocacy: Using the Science of First Impressions to Persuade the Reader*, 89 OR. L. REV. 305, 323 (2010).

[259] Amos Tversky & Daniel Kahnerman, *Availability: A Heuristic for Judging Frequency and Probability*, 5 COGNITIVE PSYCHOL. 207, 207–08 (1973).

[260] Emma Hughes et al., *The Media and Risk*, *in* RISK IN SOCIAL SCIENCE 250, 255 (Peter Taylor-Gooby & Jens Zinn eds., 2006).

[261] *Id.* at 250, 255.

[262] Adam D. Galinsky & Thomas Mussweiler, *First Offers as Anchors: The Role of Perspective-Taking and Negotiator Focus*, 81 J. PERSONALITY & SOC. PSYCHOL. 657, 660 (2001); *see also* Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124, 1128 (1974) (explaining the anchoring effect).

is random or even far too high or low. For example, in a study in which business school students were asked to negotiate to purchase a company, researchers found that the purchase price was vastly different, depending upon whether the buyer or seller proposed the initial price.[263] Remarkably, decisions about what to pay or accept were influenced by anchoring, in spite of the fact that all students knew that the party who made the initial offer was working against the other sides' interests.[264]

5. Belief Perseverance

The tendency for beliefs to be "sticky"—resistant to change—is called "belief perseverance."[265] People who are provided with information persist in using the information to form judgments, even when expressly told that the information was incorrect. This finding has been repeatedly demonstrated in empirical studies.[266] One study on perceptions of personality traits and professions showed that people who were led to believe that there was a negative or positive association between risk preferences and firefighting ability adhered to this belief, even when the information was discredited.[267] For example, jurors are significantly more likely to convict a defendant who has confessed, even when there is good reason to suspect the validity of the confession.[268] Empirical evidence has revealed that experts are particularly confident in the veracity of their own judgments and are resistant to change.[269]

---

[263] Galinsky & Mussweiler, *supra* note 262, at 660–61.

[264] *Id.* at 661. When seller-students made the opening offer, they first offered to sell the plant for an average of $26.6 million, and the average final purchase price was $24.8 million. *Id.* When buyer-students made the initial offer, they first offered to buy the plant for an average of $16.5 million, and the average final purchase price was $19.7 million. *Id.* The difference between final purchase price (depending upon who set the initial offer) was $5.1 million dollars. *Id.*

[265] Craig A. Anderson & Kathryn L. Kellam, *Belief Perseverance, Biased Assimilation, and Covariation Detection: The Effects of Hypothetical Social Theories and New Data*, 18 PERSONALITY & SOC. PSYCHOL. BULL. 555, 555–57 (1992); *see also* Lee Ross et al., *Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm*, 32 J. PERSONALITY & SOC. PSYCHOL. 880, 880, 891 (1975).

[266] *See, e.g.*, Ross et al., *supra* note 265, at 882–84 (noting that, even after having been shown the experiment materials that randomly assigned participants to various feedback conditions, people continued to exhibit beliefs consistent with the original, false feedback).

[267] Craig A. Anderson et al., *Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information*, 39 J. PERSONALITY & SOC. PSYCHOL. 1037, 1039–41 (1980).

[268] Saul M. Kassin & Holly Sukel, *Coerced Confessions and the Jury: An Experimental Test of the "Harmless Error" Rule*, 21 LAW & HUM. BEHAV. 27, 38–40 (1997).

[269] Andrea O. Baumann et al., *Overconfidence Among Physicians and Nurses: The 'Micro-Certainty, Macro-Uncertainty' Phenomenon*, 32 SOC. SCI. & MED. 167, 168 (1991). For a discussion on over-confidence bias, see William A. Edmundson, *Contextualist Answers to Skepticism, and What a Lawyer Cannot Know*, 30 FLA. ST. U. L. REV. 1, 14 (2002) ("I do not for a moment deny that criminal-defense lawyers routinely form the belief that their clients are guilty.").

The automatic cognitive predispositions discussed above can be characterized as the tendency for the human brain to come preprogrammed to see consistencies and patterns when processing data in the world. Rather than starting fresh in *tabula rasa*[270] fashion, the brain makes sense of what is already known, or interpretations that are most readily available. The automatic tendency to see patterns and draw on existing data and experiences is evolutionarily adaptive in that it allows for efficient decision-making in cases of uncertainty.[271] The list presented here is merely a sampling; many more biases and mental shortcuts that could be characterized as "fast and frugal" have been identified by behavioral researchers.

In addition to the pattern-based biases, social scientists have discovered a host of biases that are classified as motivational biases. Note that the biases above are either motivationally neutral or operate against the interest of the decision-maker. Motivational biases, in contrast, serve some conscious—or more often unconscious—need on the part of the decision-maker. Often this need is ego-related; people have a psychic drive to perceive themselves as competent, moral, and consistent. When certain logical interpretations of information would portray an individual in a negative light, that individual will often discount, discredit, or ignore that information. Conversely, when information is ambiguous, the individual will interpret it in a light that is most consistent with the preservation and reinforcement of self-serving goals.

B. Motivational Biases

Motivated reasoning is a process by which human beings reach conclusions that satisfy some desire or goal.[272] There is ample evidence that people rationalize the instinct for revenge by engaging in post-hoc rationale.[273]

1. Confirmation Bias

The confirmation bias has been the subject of a great deal of discussion in the criminal justice literature, largely due to the prevalence of wrongful convictions.[274] The confirmation bias relates to the tendency for decision-makers to erroneously confirm preexisting beliefs by selectively seeking out information or interpreting

---

[270] "Tabula rasa" is the ancient Greek term for "blank slate"—the notion that one proceeds without the influence of any preexisting tendencies. *See* Janine Ungvarsky, *Tabula rasa*, SALEM PRESS ENCYCLOPEDIA (2017), https://biblioteca.sagrado.edu/eds/detail?db=ers&an=87325096.

[271] *See* Anderson & Kellam, *supra* note 265, at 556–57.

[272] *See* Ziva Kunda, *The Case for Motivated Reasoning*, 108 PSYCHOL. BULL. 480, 480–81 (1990).

[273] *See* HAIDT, *supra* note 182, at 47.

[274] D. Kim Rossmo & Joycelyn M. Pollock, *Confirmation Bias and Other Systemic Causes of Wrongful Convictions: A Sentinel Events Perspective*, 11 NE. U. L. REV. 790, 792–93 (2019).

ambiguous information in a way that is consistent with that existing belief.[275] There is a two-part explanation for the confirmation bias. The first is an efficiency explanation, related to the biases discussed earlier; the human mind attempts to "solve the puzzle" in the least effortful way. Second, confirmation bias can also be fueled by motivational goals.[276] Researcher Ziva Kunda has identified two types of goals: accuracy goals and directional goals.[277] Accuracy simply relates to the need human beings have to be correct in their judgments.[278] Directional goals become relevant anytime an individual desires a particular conclusion or outcome.[279] In the case of directional goals, Kunda explains that in order to avoid psychic discomfort, decision-makers maintain an "illusion of objectivity" to avoid recognizing that a preexisting preference has influenced their interpretation.[280]

2. Overconfidence Bias

The vast majority of human beings are overly confident in their own choices and skills.[281] One explanation for overconfidence is that people fail to account for the uncertainty of situational variables when making judgments, and hence fail to appropriately adjust their confidence downward.[282] This bias has been

---

[275] *See* DAVID A. HARRIS, FAILED EVIDENCE: WHY LAW ENFORCEMENT RESISTS SCIENCE 3 (2012) (exploring the question of why investigators and prosecutors resist the application of social science findings to their work); *see also* Rachel E. Barkow, *Prosecutorial Administration: Prosecutor Bias and the Department of Justice*, 99 VA. L. REV. 271, 313 (2013); Alafair S. Burke, *Prosecutorial Passion, Cognitive Bias, and Plea Bargaining*, 91 MARQ. L. REV. 183, 197 (2007); Alafair Burke, *Neutralizing Cognitive Bias: An Invitation to Prosecutors*, 2 N.Y.U. J.L. & LIBERTY 512, 516–18 (2007); Barbara O'Brien, *A Recipe for Bias: An Empirical Look at the Interplay Between Institutional Incentives and Bounded Rationality in Prosecutorial Decision Making*, 74 MO. L. REV. 999, 1011 (2009).

[276] *See supra* notes 64–70 and accompanying text.

[277] Kunda, *supra* note 272, at 481–83.

[278] *See id.* at 481.

[279] *See id.* at 482–83.

[280] *Id.* at 483.

[281] Stephen V. Burks et al., *Overconfidence and Social Signalling*, 80 REV. ECON. STUD. 949, 950 (2013); Don A. Moore & Paul J. Healy, *The Trouble with Overconfidence*, 115 PSYCHOL. REV. 502, 502 (2008) (describing the multidimensionality of overconfidence as "(a) overestimation of one's actual performance, (b) overplacement of one's performance relative to others, and (c) excessive precision in one's beliefs").

[282] David Dunning et al., *The Overconfidence Effect in Social Prediction*, 58 J. PERSONALITY & SOC. PSYCHOL. 568, 576 (1990).

demonstrated in business ventures,[283] stock purchases,[284] consumer behavior,[285] lawyers' predictions about trial outcomes,[286] and mental health professionals' judgments about patients,[287] among other contexts. Other biases, such as priming, anchoring, and belief perseverance, often work hand-in-hand with overconfidence. Cognitive biases make certain judgments particularly likely, and the confirmation bias serves to cement the beliefs, foreclosing on the possibility of reexamining the judgment.

3. Self-Serving Bias

The self-serving bias is a tendency for an individual to interpret events in a way that benefits oneself.[288] When forming attitudes or making judgments, people tend to do so with reference to that which is most personally advantageous.[289] The self-serving bias becomes dangerous in situations in which a decision-maker's interests lead her to unfairly or irrationally cause harm to another. Of course, history is replete with atrocities perpetuated by human beings against one another—what makes the self-serving bias insidious is its potential to cause well-meaning actors who have been entrusted to make fair and just decisions to do otherwise.[290] A similar bias, called

---

[283] Daniel P. Forbes, *Are Some Entrepreneurs More Overconfident than Others?*, 20 J. BUS. VENTURING 623, 636–37 (2005).

[284] James Scott et al., *Overconfidence Bias in International Stock Prices: Consistent Across Countries and Trading Environments*, 29 J. PORTFOLIO MGMT. 80, 82 (2003).

[285] Wee-Kek Tan et al., *Consumer-Based Decision Aid that Explains Which to Buy: Decision Confirmation or Overconfidence Bias?*, 53 DECISION SUPPORT SYSTEMS 127, 128 (2012).

[286] Jane Goodman-Delahunty et al., *Insightful or Wishful: Lawyers' Ability to Predict Case Outcomes*, 16 PSYCHOL. PUB. POL'Y & L. 133, 133 (2010) (finding that "lawyers were overconfident in their predictions . . . . Female lawyers were slightly better calibrated than their male counterparts and showed evidence of less overconfidence").

[287] Hillel J. Einhorn & Robin M. Hogarth, *Confidence in Judgment: Persistence of the Illusion of Validity*, 85 PSYCHOL. REV. 395, 396 (1978).

[288] Russell Korobkin & Chris Guthrie, *Heuristics and Biases at the Bargaining Table*, 87 MARQ. L. REV. 795, 800–01 (2004).

[289] Farnsworth, *supra* note 68, at 572 ("A claim that a judgment about fairness is self-serving typically is a counterfactual about a value judgment: [you] would not be arguing that outcome X is fair if it were not advantageous to [you], or if [you] did not have a stake in the resolution of the dispute.").

[290] As Stephan Bibas has noted, in the case of well-meaning decision-makers, "the more information people have, the more room there is for bias." Stephanos Bibas, *Plea Bargaining Outside the Shadow of Trial*, 117 HARV. L. REV. 2463, 2498 (2004). When people receive new information about a policy that they are either for or against, regardless of which side they are on, they interpret the same ambiguous evidence as supportive of their own preferred view. *See* Charles G. Lord et al., *Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence*, 37 J. PERSONALITY & SOC. PSYCHOL. 2098, 2105, 2107–08 (1979); *see also* Constance R. Campbell & Cathy Owens Swift, *Attributional Comparisons Across Biases and Leader-Member Exchange Status*, 18 J. MANAGERIAL ISSUES 393, 403–04 (2006) (discussing how the self-serving bias can cause managers to favor subordinates with whom they

the actor-observer bias or fundamental attribution bias, causes people to view their own actions as relatively more understandable, sympathetic, and praise-worthy, while simultaneously devaluing or unfairly condemning the same actions in others. Importantly, when it comes to our moral intuitions, "it appears that . . . our moral intuitions are also influenced by an actor-observer bias as well—a bias whereby we hold other people to different moral standards than we would hold ourselves even if we were in the same situation."[291]

4. Bias Blindspot

Given how unaware we tend to be about so many of our own cognitive tendencies, it is unsurprising that we have trouble perceiving our own biases.[292] Psychologist Emily Pronin coined the term "bias blind spot" after conducting studies examining how individuals evaluated themselves and others with respect to biased attitudes.[293] Pronin and her colleagues found that, while participants had little difficulty perceiving biases in third parties, they rated themselves to be less biased.[294] Even after receiving information about research on the unconscious nature of biases, Pronin's participants continued to rate themselves as low with respect to bias.[295]

C. Retribution as Bias

Like other automatic impulses that influence decision-making, the retributivist heuristic leads humans to seek just deserts automatically, even when this instinct is irrational or counterproductive. Behavioral law scholars Braman, Kahan, and Hoffman have identified "innate cognitive traits," that function by "interacting with and generating a variety of social meanings that ultimately determine [human] understanding of and reaction to wrongdoing."[296] In other words, as Haidt has argued, human society constructs norms and values around the instinct and often largely

---

have positive relationships or similar in-group membership status); Gregory H. Dobbins & Jeanne M. Russell, *Self-Serving Biases in Leadership: A Laboratory Experiment*, 12 J. MGMT. 475, 476 (1986) (describing how management attributes failures to subordinates); Edwin P. Hollander, *Leadership, Followership, Self, and Others*, 3 LEADERSHIP Q. 43, 50 (1992) (identifying situations in which a "follower" might over-identify with a trusted leader and hence fail to question the leader's actions because of the self-serving bias).

[291] Thomas Nadelhoffer & Adam Feltz, *The Actor-Observer Bias and Moral Intuitions: Adding Fuel to Sinnott-Armstrong's Fire*, 1 NEUROETHICS 133, 133 (2008).

[292] *See generally* Emily Pronin & Matthew B. Kugler, *Valuing Thoughts, Ignoring Behavior: The Introspection Illusion as a Source of the Bias Blind Spot*, 43 J. EXPERIMENTAL SOC. PSYCHOL. 565 (2007).

[293] *Id.* at 565; Emily Pronin et al., *The Bias Blind Spot: Perceptions of Bias in Self Versus Others*, 28 PERSONALITY & SOC. PSYCHOL. BULL. 369, 369 (2002).

[294] Pronin, *supra* note 293, at 369–70.

[295] *Id.* at 374–76.

[296] Braman et al., *Some Realism About Punishment Naturalism*, *supra* note 238, at 1567; *see also* Braman et al., *A Core of Agreement, supra* note 238, at 1655.

treats these collective meanings as indisputable. Retribution *feels* like a natural guiding principle for punishment because it results from an emotional response that has deep roots in human evolutionary history.

Human societies have developed complex criminal justice systems designed to reinforce behavioral norms so that the usefulness of the automatic revenge response is in question. Given how different modern society looks from early human society, there exists a serious question of whether the retribution heuristic is still an appropriate guiding principle for deciding punishment. One test for "irrationality" used by legal scholars, ethicists, economists, and others, is to ask whether a decision-maker would make the same choice produced by the heuristic that she would make if she were engaging in slow, effortful reasoning.[297] Some evidence suggests that punishment decisions resulting from gut intuition are *not* the same as those produced by slow, effortful decision-making. When people are asked to think carefully about the optimal punishment scheme, they tend to avoid those based exclusively or primarily on a retributivist model. Instead, they often favor utilitarian, harm-minimizing solutions.[298] So, although individuals are unconsciously motivated by retaliation, their intent is to fashion consequences in a way that maximizes utility.[299]

Judgments about punishment are typically a product of outrage regarding the norm violation.[300] When legislators craft criminal laws, they are typically responding to the indignation of their constituents,[301] and when juries punish criminal acts, they are often motivated by outrage.[302] Unlike punishment based upon some agreed-upon objective, punishment based upon an emotional reaction is variable, and the appropriateness of the punishment is virtually impossible to verify.[303] Cass

---

[297] Neal Feigenson & Jaihyun Park, *Emotions and Attributions of Legal Responsibility and Blame: A Research Review*, 30 L. & HUM. BEHAV. 143, 145 (2006).

[298] *See* Robert M. McFatter, *Purposes of Punishment: Effects of Utilities of Criminal Sanctions on Perceived Appropriateness*, 67 J. APPLIED PSYCHOL. 255, 266 (1982) [hereinafter McFatter, *Purposes of Punishment*].

[299] *See* Carlsmith et al., *supra* note 60, at 1323–24; Robert M. McFatter, *Sentencing Strategies and Justice: Effects of Punishment Philosophy on Sentencing Decisions*, 36 J. PERSONALITY & SOC. PSYCHOL. 1490, 1499–1500 (1978) [hereinafter McFatter, *Sentencing Strategies and Justice*]; McFatter, *Purposes of Punishment*, *supra* note 298, at 260; Uli Orth, *Punishment Goals of Crime Victims*, 27 L. & HUM. BEHAV. 173, 181–82 (2003).

[300] Fiske & Tetlock, *supra* note 83, at 286; *see also* Carlsmith et al., *supra* note 71, at 295.

[301] Molly J. Walker Wilson, *The Expansion of Criminal Registries and the Illusion of Control*, 73 LA. L. REV. 509, 511 (2013).

[302] Cass R. Sunstein, *On the Psychology of Punishment*, 11 SUP. CT. ECON. REV. 171, 171 (2004).

[303] This distinction becomes clear when comparing two rationales for punishing a child. If a child throws a toy at another child and the basis for a parent's subsequent action is satisfaction of a desire to give the child "what he deserves" it is difficult to say whether the consequence is appropriate. Is it sufficient to put the child in a timeout? Does the child "deserve" a spanking? Perhaps any reaction that satisfied the parent's subjective desire for retribution is appropriate.

Sunstein has pointed out that "it is extremely difficult to translate outrage into the terms that the legal system makes relevant [and so] the legal system risks incoherence in the sense of erratic and unpredictable patterns [which] show a kind of irrationality."[304] The arbitrariness of these patterns of punishment has concrete consequences within the criminal justice context. Sunstein argues that members of juries fail to recognize the irrational outputs of the product of their retribution-based punishment decisions. According to Sunstein, when jurors make individual judgments based upon their feeling of outrage, "they are likely to produce patterns that they themselves would repudiate. The result is another kind of incoherence—incoherence not in the sense of unpredictability, but in the sense of patterns that are extremely hard to justify."[305] In other words, these outrage-based punishment outcomes are irrational and often harmful.

D. Intent-Oriented Teaching Versus Outcome-Based Teaching

Like other irrational tendencies that can result in suboptimal outcomes, basing punishment decisions on retributivist instincts can be a costly misstep. Research on learning and punishment suggests that a results-based, utilitarian model is better for getting compliance with desirable behavioral norms. The "just deserts" model focuses on punishing *intent*, while the deterrence model targets behavior or outcomes. Empirical research suggests that outcome-based learning, rather than intent-based learning, is most effective. Researchers trained a dart thrower to learn which of two color choices would benefit the teacher.[306] Before she threw the dart, the thrower expressed her target color choice to be measured against the result. Teaching was regulated so that the dart thrower was either punished for the "wrong" intent (when she picked the no-reward color) or the "wrong" outcome.[307] The dart-thrower learned faster when the teacher targeted the outcome, rather than the intent. In

However, if the punishment goal is the utilitarian objective of deterrence, then the minimal amount of misery imposed upon the child that will accomplish the goal of preventing the child from throwing the toy again is sufficient. Achieving this goal is imperfectly testable through experimentation. If, following the timeout, the child refrains from throwing the toy, the goal is accomplished. Moreover, other creative methods of preventing the behavior can be attempted. For example, the child can be educated, or made to "pay" restitution to the victim, by giving the victim a toy. These methods can also be used conjunctively, for maximum effect. The restitution allows for additional positive effects because the child who receives the toy experiences a benefit, so that two positive goals that improve the state of the world are accomplished simultaneously.

[304] Sunstein, *supra* note 302, at 172.

[305] *Id.*

[306] Cushman, *supra* note 153, at 350–51.

[307] *Id.* at 351.

other words, the clearest way to encourage the desired behavior and deter bad behavior is to focus on the behavior and not the intent behind the behavior.[308] These findings support the notion that a utilitarian, rather than a retributivist approach, achieves the best outcome.

E. The Malleability of Emotion-Based Choice

Retributivist models base punishment decisions on impulse, emotion, and desire to inflict pain. As a method, it is responsive to an emotional "need" and thus is responsive to public sentiment. In the United States, this instinct is exacerbated by factors like the portrayal of violence in media and entertainment, and political rhetoric.[309] For many years, researchers have been puzzled by the persistence of broad-based fear of crime on the part of the American populace.[310]

Safety and crime deterrence would seem to be utilitarian goals that should be responsive to a change in the level of crime. Yet the response to crime is influenced by perceptions rather than reality. During the 1990s and the first decade of the twenty-first century, in the face of plummeting crime rates, the United States became increasingly punitive.[311] Long after the crime rates in the United States peaked and began to drop, polls revealed that the public continued to feel unsafe.[312] One explanation for the disconnect between crime rates and punitiveness is a lag in the public's knowledge about risks, and a disincentive for lawmakers to facilitate the education of the public. Fear is a powerful motivator, and the rhetoric of fear is often used for political advantage. This sentiment, as Darley has remarked, "creates fear among politicians of appearing 'soft on crime,' lest some other politician gain advantage when running against them."[313]

---

[308] Note that to find these results to be a compelling argument for a utilitarian model of punishment, one must prioritize less crime over punishing bad intent. Researchers have found that when people are asked why offenders *should* be punished, they appear to endorse both retributive and consequentialist justifications. *See* Carlsmith & Darley, *supra* note 36, at 204–05; Carlsmith et al., *supra* note 71, at 295; McFatter, *Sentencing: Strategies and Justice*, *supra* note 297, at 1491; McFatter, *Purposes of Punishment*, *supra* note 298, at 267; Orth, *supra* note 299, at 182.

[309] Molly J. Walker Wilson, *The Rhetoric of Fear and Partisan Entrenchment*, 39 L. & PSYCHOL. REV. 117, 142, 144 (2015).

[310] *Id*. at 142; Molly J. Walker Wilson, *The Expansion of Criminal Registries and the Illusion of Control*, 73 LA. L. REV. 509, 542–42 (2013).

[311] Darley, *supra* note 30, at 190, 207.

[312] THE PEW CHARITABLE TRUSTS, THE PUNISHMENT RATE: NEW METRIC EVALUATES PRISON USE RELATIVE TO REPORTED CRIME 2–3 (2016).

[313] Darley, *supra* note 196, at 116.

VI. HARMS CREATED BY RETRIBUTIVIST PUNISHMENT

> How vainly shall we endeavor to repress crime by our barbarous punishment of the poorer class of criminals so long as children are reared in the brutalizing influences of poverty, so long as the bite of want drives men to crime![314]

Not only are members of the public and lawmakers ill-informed about crime rates, but they are mistaken in their faith that "tough on crime" measures work. As crime rates have been dropping, jails and prison populations have been growing.[315] According to the most recent statistics available from the U.S. Department of Justice, there were an estimated 2,162,400 adults in prison at year-end 2016; another 4.5 million adults in the United States were on probation or parole.[316] This figure represents a ten-fold increase since 1971, a year which had already seen a four-fold increase since 1950s.[317] At year-end 2017, the U.S. prison population was 1.5 million, and the population of jail inmates in the United States was 745,000 at midyear 2017.[318] The percentage of adults in American prisons makes the United States a clear global leader in incarceration.[319] In the last several decades, a "historically unprecedented policy shift" has emphasized "tougher, more punitive sanctioning of offenders, including greater use of incarceration and other types of correctional system punishments."[320]

Scholars identify at least four factors accounting for this increase in the U.S. prison population. These include population growth, a three-fold increase in the length of prison sentences, the expansion of the definition of many crimes (particularly concerning drug sales), and the criminalization of many activities that were

---

[314] HENRY GEORGE, SOCIAL PROBLEMS 82 (1898).

[315] Darley, *supra* note 30, at 190–91.

[316] DANIELLE KAEBLE & MARY COWHIG, BUREAU JUST. STAT., CORRECTIONAL POPULATIONS IN THE UNITED STATES, 2016, at 1–2 (2018).

[317] Darley, *supra* note 30, at 190–91.

[318] OFF. JUST. PROGRAMS, PRISON AND JAIL INCARCERATION RATES DECREASED BY MORE THAN 10% FROM 2007 TO 2017, at 1 (2019), https://www.bjs.gov/content/pub/press/p17ji17pr.pdf. For 2016 statistics, see KAEBLE & COWHIG, *supra* note 319, at 1. As of 2012, nearly 1 in 100 adults were in prison or jail nationwide. *Id.* at 4. Over the past 40 years, the United States has seen a dramatic increase in the use of prisons to combat crime, and as a result, incarcerations rates have skyrocketed, with the country's state prison population having grown by more than 700% since the 1970s. Christian Henrichson & Ruth Delaney, *The Price of Prisons: What Incarceration Costs Taxpayer*s, 25 FED. SENTENCING REP. 68, 68 (2012). States' corrections spending, including prisons as well as probation and parole, nearly quadrupled in the period from 1992 to 2012. *Id.*

[319] Alfred Blumstein, *Bringing Down the U.S. Prison Population*, 91 PRISON J. 12S, 14S (2011).

[320] Mears et al., *supra* note 28, at 87.

previously only considered regulatory offenses.[321] In the meantime, sentences have become more structured and increasingly severe.[322] The purported logic behind the increase in incarceration rates in the United States has been both retributivist and deterrence-based.[323] However, repeated studies have concluded that this attempt to produce deterrence is failing.[324] Moreover, although deterring crime has continued to be advanced as a rationale for keeping people in prison, experts, policy advisors, and lawmakers have known for years that incarceration actually *increases* recidivism and does not make the public safer.[325]

A. Prison Does Not Deter and Creates Additional Problems

Spending time in prison or jail impacts an individual significantly, and usually not in a way that benefits society.[326] Prison is costly, both in terms of the price to taxpayers and to the incarcerated individual, who often loses employment, educational opportunities, family, friends, and status in society.[327] From the perspective of a utilitarian, these significant burdens argue in favor of incarceration *only* when there are significant benefits to offset the losses.[328] As a practical matter, research has shown no beneficial effects of incarceration on recidivism rates, and it may be the case that prison *increases* recidivism.[329] Offenders who spent time in prison may

---

[321] Darley, *supra* note 30, at 190–91; *see also* Blumstein, *supra* note 319, at 18S (finding that more effective policing and a growth in the crime rate "contributed hardly at all" to the growth in the prison population; however "[v]irtually the entire growth was attributable to increases in the commitment rate and in the time served, especially in time served more recently"). An additional factor Blumstein identified was the recommitment of drug offenders. *Id.* at 17S.

[322] *See also* Hessick & Berman, *supra* note 44, at 169.

[323] Mears et al., *supra* note 28, at 84–85. Additional factors include the belief that more time in prison would or could reduce recidivism and that the benefits of increased incarceration would outweigh social or economic costs. *Id.* at 85. "[R]etribution and public safety constitute the avowed goals expressed by legislatures." *Id.* at 88; *see also* Markel, *supra* note 40, at 2159 (pointing out that "recent scholarly and policymaking interest in retributivism stems in part from negative reactions to problems associated with recidivism," as well as the fact that "the theory has a stronger rationale than it once seemed to have").

[324] Darley, *supra* note 30, at 194–95.

[325] DON STEMEN, VERA INST. JUST., THE PRISON PARADOX: MORE INCARCERATION WILL NOT MAKE US SAFER 2 (2017); *see also* Cassia Spohn & David Holleran, *The Effect of Imprisonment on Recidivism Rates of Felony Offenders: A Focus on Drug Offenders*, 40 CRIMINOLOGY 329, 334–35 (2002) (finding that individuals sentenced to prison had higher recidivism rates and recidivated more quickly than individuals sentenced to probation).

[326] Darley, *supra* note 30, at 193.

[327] *Id.*; Henrichson & Delaney, *supra* note 318, at 69.

[328] *See* Cullen et al., *supra* note 27, at 59S–60S.

[329] Mitchell et al., *supra* note 28, at 19; *see also* Cullen et al., *supra* note 27, at 60S (finding that prisons do not reduce recidivism more than noncustodial sanctions). *But see* Rhodes et al., *supra* note 28, at 733 (finding that lengthening a prison term does not increase recidivism, but

exhibit increased levels of reoffending following release in comparison to offenders who received non-prison sanctions.[330] In one study, researchers found that the majority of re-offenders were non-violent property and drug offenders, and determined that for other offenders, imprisonment generally yields no effects or substantively small adverse effects on the likelihood of reconviction compared to alternative sanctions.[331]

Any deterrent effect created by the threat of incarceration would seem to be offset by significant collateral consequences of this form of punishment.[332] Formerly incarcerated individuals experience difficulty reintegrating into mainstream society.[333] Many ex-convicts have difficulty finding housing and employment, and the disintegration of social networks leaves them with few options other than criminal activity to sustain themselves.[334] Research reveals the importance of employment and related income in determining the life course of the ex-incarcerated.[335] On average, inmates have significantly lower levels of literacy and educational attainment than adults in the general population, further aggravating attempts to find stable employment after release.[336]

---

rather reduces recidivism by a small amount). *Compare with* Mears et al., *supra* note 28, at 118–19 (concluding that there is an "inverted U-shaped relationship between time served and recidivism," at least for inmates serving up to five to six years in prison). This means that there may be no single effect of time served on recidivism and the effect of time served may vary depending on the specific amount of time served. *Id.*

330 Mitchell et al., *supra* note 28, at 12–13 (finding that their studies show a non-statistically significant difference between recidivism among offenders who served prison terms and those who received non-prison sanctions).

331 *Id*. at 17–18; *see also* Mears et al., *supra* note 28, at 122 (finding that lengthier terms of incarceration, beyond a few months, do not "readily appear to reduce recidivism" and may actually increase it).

332 *See* Mitchell et al., *supra* note 28, at 4–5.

333 Darley, *supra* note 30, at 193.

334 *Id.*; *see also* Donald T. Hutcherson II, *Crime Pays: The Connection Between Time in Prison and Future Criminal Earnings*, 92 PRISON J. 315, 316–17 (2012) (noting that formerly incarcerated offenders are stigmatized by their past and that employers are less likely to hire the ex-incarcerated compared to those without criminal records); Rhodes et al., *supra* note 28, at 761–62 (pointing out that the longer someone "is separated from his or her family, the more likely it harms spouses and children," including loss of income and loss of social support, resulting in "increased reliance on welfare and problems with children's prosocial development"); Mears et al., *supra* note 28, at 88 (pointing out adverse effects on "ties to family and friends, mental and physical health, employment prospects, and the ability to access public housing"); Matthew Makarios et al., *Examining the Predictors of Recidivism Among Men and Women Released from Prison in Ohio*, 37 CRIM. JUST. & BEHAV. 1377, 1378–79 (2010) (pointing out the pressures inmates face when finding housing on their own).

335 Hutcherson, *supra* note 334, at 316–17.

336 Makarios et al., *supra* note 334, at 1378.

Further, "spending significant time incarcerated can erode the social networks necessary for stable conventional employment opportunities" after release.[337] Imprisonment imposes a cost in "social influence[s] that shapes inmates' attitudes toward crime and violence, peer networks, ties to the conventional order, and identity."[338] Whereas individuals with an incarceration history experience difficulty finding legitimate forms of employment, prison often provides them with new avenues of obtaining illegal income. Legitimate and healthy relationships that existed outside of prison are often replaced by relationships with other inmates who may have continuing involvement with illegal enterprises outside of prison. Prisoners are very likely to join gangs, as this is often a method of obtaining protection inside prisons.[339] These relationships forged during a period of imprisonment make reoffending particularly likely after release.[340] Some scholars argue that "prisons increase offending by serving as 'schools of crime' or by stigmatizing offenders in ways that increase their propensity for future criminal behavior."[341] Membership in a prison gang is likely to increase recidivism by signaling a commitment to a criminal lifestyle, altering social and human capital, and invoking an institutional response.[342] A prison gang member is perceived to be more trustworthy by a fellow

[337] Hutcherson, *supra* note 334, at 317; *see also* Cullen et al., *supra* note 27, at 53S–54S (discussing the costs of imprisonment on the inmate, including the risk of physical victimization, being cut off from family and prosocial contacts on the outside, facing stigmatization upon release, association with other offenders, and the general harms of imprisonment); Rhodes et al., *supra* note 28, at 762 (discussing the relationship between incarceration levels, crime, and disadvantage—the communities with the highest levels of incarceration also have the highest levels of crime and disadvantage).

[338] Cullen et al., *supra* note 27, at 53S.

[339] John L. Worrall & Robert G. Morris, *Prison Gang Integration and Inmate Violence*, 40 J. CRIM. JUST. 425, 426–27 (2012). There has been a significant increase in the number of prison gangs in recent decades, and prison gangs have become a principal form of inmate organization in many prisons because prison gangs protect the extensive in-prison contraband market. *Id.* at 427. Additionally, since the typical prison inmate is from a community that already has a gang problem, gang ties and culture are naturally imported into the correctional setting. *Id.* Violent inmate misconduct may be explained as "calculated risk taking"; it is a means of "self-help, self-defense, or social control by which inmates seek to establish social status, maintain public identities, or gain economic benefits." Marie L. Griffin & John R. Hepburn, *The Effect of Gang Affiliation on Violent Misconduct Among Inmates During the Early Years of Confinement*, 33 CRIM. JUST. & BEHAV. 419, 421 (2006). "Many inmates import the values, norms, and experiences acquired as members of street gangs" prior to their imprisonment, while "still other inmates become affiliated with gangs only after incarceration. Gang affiliation is rewarded with social support, social status, personal security, and access to contraband." *Id*. at 423.

[340] *Cf.* Hutcherson, *supra* note 334, at 331–32.

[341] Mitchell et al., *supra* note 28, at 1; *see also* Cullen et al., *supra* note 27, at 53S.

[342] Brendan D. Dooley et al., *The Effect of Prison Gang Membership on Recidivism*, 42 J. CRIM. JUST. 267, 268 (2014). Dooley et al. concluded that membership in a prison gang increased recidivism by about six percentage points, which was a quantitatively large effect. *Id.* at 272.

inmate because of his affiliation, "and this trust facilitates cooperation in criminal activities."[343] Collectively, these factors suggest that offenders become more, rather than less, criminally oriented due to their prison experience.[344]

Evidence from behavioral science suggests that in terms of deterring crime, increasing prison sentences is less effective than other methods of crime control.[345] In contrast to increasing a prison sentence for a crime, "campaigns that make salient in the mind of the public the possibility of being caught for committing [an] offense are often successful."[346] For many crimes, "a prevention strategy that relies on potential perpetrators to mentally weigh the consequences of conviction and punishment simply does not comport with the evidence of the actual 'thought' process of convicted criminals," whether that is due to mental representations incongruent with reality, participation in groups of individuals aimed at committing a crime, or the use of mind-altering substances.[347] In sum, despite the reality that prisons impose high costs on offenders for their behavior, they appear to be a "weak change agent" in terms of specific deterrence, and many offenders are not "moved by imprisonment to stay out of trouble."[348] Custodial sentences do not reduce recidivism more than non-custodial sanctions or rehabilitation and prevention efforts.[349]

Finally, the costs of high levels of incarceration to the public are significant.[350] The cost of holding a person in prison is only part of the overall financial burden.[351] Spending time in prison takes a toll physically, mentally, and behaviorally. Many released prisoners reenter society with problems they did not have prior to incarceration that impose additional costs on the public. For example, former prisoners are more likely than the general public to return to society having contracted AIDS or drug-resistant tuberculosis, which requires treatment and can also then be transmitted to the public.[352] Parolees are less likely than the average person to have gainful

---

[343] *Id.* at 269.

[344] Cullen et al., *supra* note 27, at 53S.

[345] Darley, *supra* note 30, at 200; *see also* Mears et al., *supra* note 28, at 90 (noting that a potential theory underlying the use of incarceration as punishment includes perceptions about other aspects of incarceration, such as "the experience of incarceration, prison conditions, or the extent to which actual time served in prison accords with sentence length").

[346] Darley, *supra* note 30, at 204.

[347] *See id*. at 197–98 (discussing criminal acts committed while drunk or using other mind-altering drugs); *see also* Anthony N. Doob & Cheryl Marie Webster, *Sentence Severity and Crime: Accepting the Null Hypothesis*, 30 CRIME & JUST. 143, 190–91 (2003) (concluding that variation in the severity of sanctions is unrelated to levels of crime).

[348] Cullen et al., *supra* note 27, at 54S.

[349] *Id*. at 53S–54S.

[350] Darley, *supra* note 30, at 193.

[351] *Id.*

[352] *Id.*

employment, and so are more likely to be reliant on public services and entitlements.[353] Because incarceration often disrupts family ties, released prisoners are less likely to support a dependent child following incarceration.[354] For those who do attempt to support a child, there are significant challenges in doing so. As noted by criminologists and sociologists, entire communities suffer because so many of their young adults—mostly male—spend so many years behind bars.[355] For these communities, many of which already face significant challenges, the absence of a disproportionate percentage of their young male population poses an additional burden. Society at large ends up suffering because of the existence of these blighted communities.[356]

The reality of the costs of imprisoning so many people leads to two inescapable conclusions. The first is that there is a low utility in our current criminal justice system. While the costs vary, they are calculable and significant. Even if the public were willing to pay a heavy price for deterrence, data on recidivism and prison-related factors that can exacerbate crime rates reveal a pattern of unintended consequences that refute claims that the threat of incarceration is an effective general or specific deterrent.[357] In short, it is difficult to impossible to defend current incarceration practices on utility grounds. The second conclusion is related to the first; society pays a steep price to keep so many people behind bars.

Finally, there is the question of whether the prisoners who end up losing months, years, decades of their lives to imprisonment are really deserving of this

---

[353] *Id.*; *see also* ADAM LOONEY & NICHOLAS TURNER, WORK AND OPPORTUNITY BEFORE AND AFTER INCARCERATION 7 (2018) (finding that almost half of all ex-prisoners earn less than $500 in their first full year after release from incarceration).

[354] LOONEY & TURNER, *supra* note 353, at 10.

[355] Dorothy E. Roberts, *The Social and Moral Cost of Mass Incarceration in African American Communities*, 56 STAN. L. REV. 1271, 1281, 1293–94 (2004).

[356] *See* Melissa Li, *From Prisons to Communities: Confronting Re-Entry Challenges and Social Inequality*, AM. PSYCHOL. ASS'N (Mar. 2018), https://www.apa.org/pi/ses/resources/indicator/2018/03/prisons-to-communities. People face significant challenges with re-entry after incarceration. *Id.* A criminal record limits employment prospects, public housing assistance, and social services. *Id.* For example, "individuals with past drug or felony convictions are ineligible for public housing," and many private rental housing associations "have policies against renting to people with criminal records." *Id.* Most states "ban individuals with drug felony convictions from being eligible for federally funded public assistance and food stamps." *Id.* When previously incarcerated individuals face barriers to housing, employment, and public assistance, they are more likely to become homeless or to reoffend. *Id.*

[357] It is also important to acknowledge that prison can reduce the criminal participation of offenders "simply by caging them so that they cannot break the law in the community." Cullen et al., *supra* note 27, at 51S. While this incapacitation effect is undoubtedly present, scholars find that it is often "inadvertently rig[ged]" because estimates compare how many crimes are prevented if offenders are locked up as compared to doing nothing to them, instead of considering the effect of imposing a non-custodial penalty. *Id.*

punishment. Some scholars characterize prison as a form of retaliation that fits into the utilitarian scheme, arguing that validating a community's collective desire for revenge increases happiness and creates respect for a community's norms.[358] A discussion of the causes of criminal activity is well beyond the scope of this Article. Suffice it to say, there are systematic inequities in education, health care, housing, drug and mental health treatment, economic opportunities, and other resources that affect the likelihood that an individual will be convicted of a crime. The very argument often used to bolster a retributivist approach—namely that punishment must be legitimate in the eyes of the members of society to encourage public buy-in—would seem to cut against a just-deserts model that systematically inflicts pain disproportionately on a particularly disadvantaged segment of society.

B. Satisfaction of a Drive is a Poor Reason to Punish

Many behaviors that satisfy an instinctive drive are condemned on moral grounds and heavily sanctioned through our criminal justice system. Humans are sometimes driven to cheat, steal, squander, litter, commit acts of violence against one another, free-ride, and engage in acts of vigilantism. These behaviors satisfy an instinct to put oneself first, above the good of the collective. Society punishes these behaviors with fines and incarceration, and even death. Criminal codes are designed to maximize good, create parity, and promote peace efficiently. Society values these goals above the goal of allowing behavior that satisfies an intuitive drive. Equally important, the retribution instinct looks a great deal like other instincts that we reject as bases for behavior.

Lately, there has been debate and push-back to this notion that human moral intuitions are somehow universally "correct" and worthwhile. Specifically, moral intuitionism has been undermined by research and theory around how our intuitions operate, and it has been shown that they are subject to contamination by factors, such as context framing, long viewed as distortive.[359] Distortion in our "natural" instincts suggests that what we imagine to be a kind of "natural law"[360] is an emotional response to a stimulus in the environment, and the emotional response is

[358] Kalstein et al., *supra* note 43, at 576. Additional benefits of this type of retribution include satisfying a community's blood lust and deterring a community from "tak[ing] the law into their own hands." *Id.* at 582.

[359] *See* Nadelhoffer & Feltz, *supra* note 291, at 137.

[360] The part of the Declaration of Independence that reads, "We hold these truths to be self-evident, that all men are created equal" is a profoundly ironic example of a "natural law" given the status of slaves in some of the colonies at the time. The equality extended to white men, and the naturalness of this sentiment was a product of the time, as well as historical and political factors. THE DECLARATOIN OF INDEPENDENCE para. 2 (U.S. 1776).

profoundly dependent upon our attitudes, values, memories, and other inputs that shape our choices.[361]

C. Casting Doubt that People are Pure Retributivists

Also undermining the retributivist rationale for punishment is uncertainty about whether individuals are as revenge-oriented as portrayed. Civic republican political theory[362] has advanced an optimistic view of punishment behavior, namely that people are animated in their judgments about punishment by something other than pure retribution. There is reason to think this may be so. A careful look at some of the assumptions underlying the empirical work on punishment motivation reveals potential ambiguity. For example, one assumption psychologists have made is that *intentionality* of the wrongdoer implicates purely retributivist impulses.[363] However, a punisher's attention to the intent of the wrongdoer could implicate mixed motives. A punisher could care about intentionality because wrongdoers who intentionally cause harm are more likely to harm in the future and may be more likely to escalate, eventually committing more serious infractions. In contrast, an offender who causes harm unintentionally may not establish a pattern of systematic, repeated harm in the future. Hence, intent could serve as a proxy for future dangerousness, a measure that would animate a *utilitarian* approach to punishment.

Conversely, some factors that researchers have assumed were utilitarian could implicate retributivist notions. Research by Carlsmith manipulated information about wrongful acts, intending to influence either utilitarian or just-deserts motives selectively.[364] However, the singular way in which Carlsmith categorized some of the information is debatable. For example, when Carlsmith asked participants to select "frequency of crime in society," he explicitly assumed that this was a deterrence-related category.[365] However, the frequency of *any* behavior can serve a signal that it is normatively more accepted. The reason for this is that social-informational referencing constantly occurs during a social exchange. An informational cascade is a well-recognized psycho-social phenomenon that leads people who are not already

---

[361] There is a robust and important debate among scholars in this area. This Article in no way does justice to this important conversation. For more on this topic, see Henry Mather, *Natural Law and Right Answers*, 38 AM. J. JURIS. 297, 306 (1993).

[362] Civic republican theory is most easily understood as a form of government that rejects autocratic forms of government in favor of a government by and for the people. *See* PHILIP PETTIT, ON THE PEOPLE'S TERMS: A REPUBLICAN THEORY AND MODEL OF DEMOCRACY 5–6 (2012). An important premise is the engagement in a political project aimed at securing the common good of all its citizens. *Id.* at 5; *see also* PHILIP PETTIT, REPUBLICANISM: A THEORY OF FREEDOM AND GOVERNMENT 5–9 (1997).

[363] *See* Michael S. Moore, *Justifying Retributivism*, 27 ISR. L. REV. 15, 32–33 (1993).

[364] Carlsmith et al., *supra* note 71, at 288–90, 293.

[365] *See* Carlsmith & Darley, *supra* note 36, at 203.

committed to a particular belief to base their own beliefs on the apparent beliefs of others.[366] Certain infractions (such as exceeding the speed limit while driving) are so common that moral blameworthiness for this violation is relatively low.[367]

Efforts to disentangle instrumental from retributive motives have proven tricky because as crimes become more "worthy" of punishment, the value in deterring them also increases.[368] Darley described the results of one study as evincing a desire for just deserts, saying, "even when participants were instructed to ignore the retributive factors, the moral severity of the crime intruded on their sentencing and remained a significant predictor of the sentence."[369] However, an equally plausible explanation for Darley's finding is that respondents used crime severity to gauge the need for deterrence. This explanation finds support in the experimental design Darley used; participants were admonished *only* to use a deterrence rationale. The seriousness-deterrence explanation is as plausible, if not more plausible, than the recalcitrant respondent explanation.

In another study, Darley varied the moral seriousness of the crime by telling participants that stolen funds were either used to benefit underpaid factory workers at the company's overseas plant or used to finance a lavish lifestyle and extensive gambling debts.[370] Darley concluded that, because the moral seriousness of the scenario influenced respondents, people are retributionists. Another explanation is that people see some good in theft for beneficent reasons, whereas they see no good in theft for wasteful and selfish reasons.[371] If this is correct, then weighting these two outcomes is a utilitarian exercise. Adopting a myopic view of the utilitarian perspec-

---

[366] Timur Kuran & Cass R. Sunstein, *Availability Cascades and Risk Regulation*, 51 STAN. L. REV. 683, 685–86 (1999).

[367] *See generally* H. Laurence Ross, *Folk Crime Revisited*, 11 CRIMINOLOGY 71, 76 (1973). Another example is drinking alcohol under the age of 21 on college campuses. This is a frequent activity, and most consider the blameworthiness of this activity to be low. Of course, there is a question of cause and effect. It is difficult to know whether the activity is considered acceptable because it is so common or if it is common because it is acceptable. Either way, the commonness-blameworthiness bear the same relationship; commonality is a marker for acceptability.

[368] Some studies have circumvented this problem by showing that participants desire punishment even when the offender is paralyzed, ruling out the consequentialist motive of incapacitation. Nonetheless, such a design cannot rule out other consequentialist concerns such as that of general deterrence. *See* Carlsmith et al., *supra* note 71, at 286–87.

[369] Carlsmith & Darley, *supra* note 36, at 205.

[370] Carlsmith et al., *supra* note 71, at 289.

[371] The fable of Robin Hood provides an interesting parallel: stealing from the rich to give to the poor is likely judged to be less morally wrong than other types of crime. But, in fact, it may be viewed as a net gain or benefit to society in an entirely utilitarian fashion. Society suffers when the rich get too rich, particularly at the expense of the poor. Where is this more evident in modern society than in the big company versus blue-collar worker paradigm?

tive could lead to hyperfocus on crime prevention as the *only* utilitarian goal considered by members of society. Taking a broader view, findings from punishment studies suggest a wide range of utilitarian motives.

One understudied utilitarian motivation for punishment is its communication value. In one study, Carlsmith found that punishers did not feel the anticipated satisfaction following a just-deserts based punishment.[372] However, in another study by Mario Gollwitzer et al., when punishers received feedback indicating that the offender understood the punisher's intent to communicate disapproval of the offensive behavior, the punisher *did* feel better.[373] These findings suggest that punishers value the ability to communicate disapproval through the imposition of sanctions. The emphasis on communicating society's expectations is most consistent with a learning model, which stresses the utilitarian goal of future crime prevention and rehabilitation. Consistent with this interpretation are the results of a study by Funk et al., who found that punishers were *most* satisfied when the offender received the message intended by the punishment and the offender appeared to change as a result.[374] According to Funk and her colleagues, punishing is only satisfying if it accomplishes the forward-looking goal of reproving and thereby transforming the "offender's moral attitude."[375]

D. Why the Infliction of Pain for Just Deserts is Morally Wrong

Philosopher and ethicist Braithwaite, an important scholar in the restorative justice movement, held that no legitimate system of punishment could include a retributivist aim.[376] Many who hold similar views point out that inflicting pain for the sake of causing pain does not comport with well-established normative tenets. One test for determining whether an intuition is "correct" and should be used to make substantive policy is whether it is consistent with other important, widely held principles.[377] This view is an excellent argument against reliance on "intuition" for punishment determinations. In response to the "why punish?" question, "brute" retributivists (as one may call them) simply insist that wrongdoers deserve to suffer

---

[372] Carlsmith et al., *supra* note 60, at 1323.

[373] Mario Gollwitzer et al., *What Gives Victims Satisfaction when They Seek Revenge?*, 41 EUR. J. SOC. PSYCHOL. 364, 369 (2011).

[374] Friederike Funk et al., *Get the Message: Punishment Is Satisfying if the Transgressor Responds to Its Communicative Intent*, 40 PERSONALITY & SOC. PSYCHOL. BULL. 986, 993 (2014).

[375] *Id.*

[376] *See* Carlsmith & Darley, *supra* note 36, at 208.

[377] Adam J. Kolber, *How to Improve Empirical Desert*, 75 BROOK. L. REV. 433, 436 (2009) ("On the other hand, widely-shared intuitions may be relevant to justification when they fit together with other intuitions and with deeper moral principles. For example, most people share the intuition that, absent unusual circumstances, it is wrong to intentionally kick a sleeping dog. This intuition is consistent with many other widely-shared intuitions about the impermissibility of causing unnecessary harm.").

for their wrongdoing on intuitive grounds. Braithwaite and colleague Pettit hold that a satisfying normative theory is "compelled to give reasons for an intuition that it is right intentionally to inflict suffering on the criminal because it must be reconciled with another intuition that it is generally wrong intentionally to inflict suffering on another human being."[378]

Laws in the United States reflect this moral norm. Criminal codes only allow one person to physically harm another when it is reasonably necessary in order to repel an attack that is *at least as* serious *and* imminent.[379] In this way, the laws dealing with self-defense reflect our collective abhorrence for the imposition of physical harm, absent compelling need. One famous moral dilemma illustrates a similar prohibition against harming another. The Trolley Problem asks respondents whether they would pull a trolley switch to kill one person in order to save the lives of five others. The very fact that this is a moral quandary illustrates the truism that causing harm to another person is wrong (even where there is a compelling rationale).[380] It is easy to see how the altruistic impulse also arises from the evolutionary need to behave cooperatively to preserve the social pact and thrive as a species. This Article has argued that the retribution impulse is an unnecessary and harmful artifact of early evolution, leading to excessive and inefficient punitive measures. This view is premised on a utilitarian argument for a utilitarian scheme. However, even assuming this impulse remains a legitimate basis for punishment, it nevertheless competes with the prohibition against harming another.

VII. CONCLUSION

Rather than justifying retribution with post hoc rationalizations that subvert utilitarian goals, we should recognize retribution as what it is: a costly, counterproductive, and morally archaic approach to punishment. Like other biases that we explicitly reject, motivated reasoning is widely recognized to lead people to behave irrationally, and at times, unethically. We should resist the urge to inflict pain for

---

[378] JOHN BRAITHWAITE & PHILIP PETTIT, NOT JUST DESERTS: A REPUBLICAN THEORY OF CRIMINAL JUSTICE 160–61 (1990).

[379] Waters v. Lockett, 896 F.3d 559, 569 (D.C. Cir. 2018) (citing Murphy-Bey v. United States, 982 A.2d 682, 690 (D.C. 2009)).

[380] There are a number of versions of the Trolley Problem, a moral dilemma. *See, e.g.*, Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 MONIST 204, 206 (1976). Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him.

the sake of inflicting pain and stop justifying practices arising from a just-deserts approach. Instead, we should move to a utilitarian focus based upon identifying desirable outcomes, gathering data on which practices increase those outcomes, and developing thoughtful practices from a maximizing-good standpoint. Doing so will not be a simple task. It will involve prioritizing various goals. For example, is it more important to focus on crime detection or prevention? How do we create the right level of deterrence without spending too many resources on punishment and not enough on prevention? How do we spend our limited resources in a way that puts them to the best use? Many questions will remain after we reject the retributivist approach.

A utilitarian approach also has the advantage of making space for creative methods of crime prevention that have not yet been attempted. By removing the focus on inflicting pain as a good in itself, policymakers become free to experiment with various proactive measures and responses to crime that are novel. Innovation in crime prevention, like innovation in other areas, will allow for cost-effective, creative new methods that may have unanticipated positive side effects.[381]

As vocal participants in the policy debate, we should be particularly concerned about using retribution as a guiding principle for punishment if it leads to a system that imposes high costs without sufficient justification. Our criminal law system teaches us that the most basic instincts and desires are not automatically a legitimate basis for action. Infanticide and rape satisfy primal instincts at times, but these acts are considered some of the most heinous transgressions, and they are severely sanctioned. The reason we prohibit and punish the satisfaction of these primal urges is that they impose an unacceptable cost on other individuals and on society generally. Deciding what behavior to prohibit always requires a balancing of the interests on either side, the productiveness of one activity versus the good generated.

Legal scholar Owen Jones points out that the evolutionary processes affecting human behaviors have resulted in combinations of genes that "predispose" humans to certain behaviors without rigidly "determining" the choices we make.[382] Despite the human impulse to satisfy a desire for revenge, we have sophisticated cognitive capabilities that make it possible to override this impulse. According to Jones, "[i]t is precisely because evolutionary processes favored behavioral plasticity that (with the exception of reflexes and the like) genes do not generally determine our behavior as if we were 'hard-wired' to respond inevitably to a certain stimulus with a single, corresponding act."[383]

---

[381] For example, funneling some of the hundreds of thousands of dollars currently spent to confine people into educational, drug treatment, and gang prevention programs may result in a higher quality of life with improvements to health and education, while reducing gang activity and the commission of crimes.

[382] Jones, *supra* note 105, at 851.

[383] *Id.* at 852.

The utilitarian model allows for more creative measures, such as restorative and compensatory practices like requiring the perpetrator to pay restitution to the victim as part of an effort to reverse a wrong and reinforce a sense of responsibility in the offender—therefore reaching a maximum beneficial impact through reeducation of the offender and restoration of the victim. A utilitarian model also is more efficient, recognizing that it is counterproductive to extract resources from society in order to support a punitive practice, such as lengthy incarceration or incarceration for petty offenses, for the primary purpose of inflicting pain on the offender. The counterproductivity of imposing pain through incarceration becomes particularly apparent when considering research that reveals that incarceration increases recidivism. In contrast, the utilitarian model allows for a clear-eyed examination of what methods decrease crime and benefit would-be victims. It allows for flexibility and creativity to minimize penalties where the net benefit would be zero or a loss. It allows for the refunneling of resources from support for record-breaking punitive measures to restorative practices that could return offenders to their communities, where they could support other members of those communities (a net gain) and possibly be agents for good in providing support to others who would otherwise (without their presence and support) become offenders.[384]

In rejecting a retributivist approach in favor of a utilitarian one, it is important not to fall into old patterns, using a novel justification. A true utilitarian approach should factor in *all* costs and weigh them against the benefits of incarceration, favoring incarceration only when it is the least costly, most effective method of deterrence. Existing data on the effectiveness of long-term incarceration and its associated costs suggests an overhaul in our current system and the development of new restorative approaches, focused on systemic solutions to the root problems that lead to criminal activity, such as lack of education, training, and job opportunities. While theoretically a just-deserts model could accommodate the current inefficient and overly punitive system, no system of punishment should weigh specific costs as more or less significant simply because they are borne by one segment of society instead of another.

[384] An example is a father, son, and brother who could provide financial and emotional assistance to his children, parents, and younger siblings who, without him, might turn to crime.