Citizens' Assignments of Punishments for Moral Transgressions: A Case Study in the Psychology of Punishment

John M. Darley*

I. INTRODUCTION

A first appearance of much of the material in this article was at a conference on cognitive and neurological contributions to issues arising in law. Attention to this question turns out to be timely. Within cognitive psychology, there has been a resurgence of interest in moral decision making, and this research has progressed far enough that some conclusions about the psychological processes involved have been reached. Further, brain-imaging methods have been used in some of these studies and are revealing in their results. So a discussion of the cognitive science theorizing and neural imaging research can illuminate questions concerning the determinants of moral judgments.

This article will focus on the psychological aspects of a certain subset of moral judgments, namely whether an intentional action is a morally wrong one, and, if it is judged to be a morally wrong one, the severity of punishment to be assigned to that action and the motivations the respondent has for assigning those punishments. This article seeks to contribute to a discussion on a specific criminal justice question, namely that of the psychological drivers of the "punishment impulse."

A second question will also be considered involving the degree to which the judgments made by the respondent are relatively fixed in character and unchangeable or are alterable by persuasion or other means. This question is often raised, and is raised here, because in the case of moral judgments, some scholars suggest that these judgments are strongly affected by evolutionarily given propensities.[1] This position is often thought to produce judgments that are in the main fixed, although this is not a necessary consequence of an evolutionary stance on morality.[2]

Within the justice community, a good deal rides on questions concerning whether punishment judgments are relatively unalterable, whether various wrong-

* Warren Professor of Psychology and Professor of Psychology and Public Affairs in the Woodrow Wilson School of Public and International Affairs, Princeton University.

[1] *See* John Mikhail, *Universal Moral Grammar: Theory, Evidence and the Future*, 11 TRENDS COGNITIVE SCI. 143, 149 (2007).

[2] Robinson and Kurzban have recently made the case for largely evolutionarily determined punishment and sentencing intuitions that are quite difficult to change. Paul H. Robinson & Robert Kurzban, *Concordance and Conflict in Intuitions of Justice*, 91 MINN. L. REV. 1829 (2007).

doing acts are ones that people implacably see as demanding punishment, and whether the kinds of treatments that people perceive as required when administering punishment are relatively fixed in character, centering around the administration of pain.[3] If this is the perspective from which citizens necessarily think of crime and its punishment, there would seem to be limits to the degree to which criminal justice practices can move in, what many would consider, a more humane direction. Is this the message of psychological research in this area?

II. COGNITION'S TASK: FOLK THEORIES OF PUNISHMENT

Various traditions of inquiry intersect to study punishment decisions of citizens. The task here is one of articulating "folk theories" of the causes of immoral behavior and the appropriate responses to that behavior. Folk theories of this sort are studied by psychologists and sociologists, and certainly by criminologists. Lately, in the rapidly emerging field of psychology, the study of human judgment and decision-making processes has made substantial contributions to thinking on punishment. Recent advances in the degree of resolution possible on brain-imaging processes create the possibility of contributions from neural-imaging research. What picture of moral reasoning in general, and punishment assignments in specific, is emerging from that research?

A. *Cognitive, Behavioral, and Neurological Research Methods*

How is this research conducted? For our analysis, it is useful to distinguish two paradigms in common use. The first is "the scenario study." Research in this paradigm generally gives a respondent the task of assigning punishment judgments to one or more short scenarios that specify that an actor has intentionally (or carelessly or recklessly) committed a moral wrong, one that will be perceived as sufficiently serious to the reader to warrant the sort of punishment that is available to the criminal justice system. The reader, now the "respondent," decides whether the action is allowable or not, and, if not allowable, assigns a punishment to the transgressor, generally on a scale provided by the researcher.[4]

The second paradigm can be called the "experimental game" paradigm, although the word "game" adds a playful connotation to these studies that is

---

[3] Robinson and Kurzban make the case for a general immutability of the punishment sentences that citizens assign. *Id.* Adam Kolber argues that sentencing policies need to take more seriously the task of equating the subjective pain that persons of different sensitivities to pain experience if they have committed crimes of equal wrongness. Adam J. Kolber, *The Subjective Experience of Punishment*, 109 COLUM. L. REV. 182 (2009).

[4] A number of scholars have conducted scenario rating studies. *See, e.g.*, Mark Warr, *What Is the Perceived Seriousness of Crimes?*, 27 CRIMINOLOGY 795 (1989) (reviewing many of the landmark studies using this paradigm).

importantly false.[5] In this paradigm, two or more respondents are brought together to take part in elaborate, often computerized "games." Generally the respondents are kept unaware of the outside-the-game identities of the other respondents so that concerns for the consequences of their actions persisting into "the real world" are eliminated. One reason for this concern is that in those games, the structure creates the possibility that subjects can commit actions that are perceived by other subjects as serious moral violations. If participants are known to one another, past relationships or concern for future sanctions would alter behavior in the game. It is the subjects' retaliatory responses to the violations inflicted on them that are of interest, and the games are carefully engineered to create the possibility of these retaliations.

Research of either the scenario or game kind is sometimes conducted in ways that allow imaging of the brain of the subject judging the scenario or of the brain of one or more of the game-playing subjects while in the course of making their decisions. Thus in some of these studies, neuroscience contributions to discoveries of the underlying brain processes that correlate with the cognitive decisions can be explored.[6]

B. *The Findings of Scenario Studies*

Scenario studies are done within an experimental context, and this generally involves giving groups of respondents varying versions of the core scenario or scenarios. Researchers are also sometimes interested in how subjects' political ideologies may lead them to react to scenarios differently, so they administer scales that are known to measure these ideologies which are then correlated with scenario responses. In morality and punishment research, it is common to measure the liberal or conservative ideologies of the respondents to see if they contribute to patterns of judgment.

The variations made in the text of the scenario generally create the differences that are of theoretical interest to the researchers. For instance, the core of the

---

[5] See John M. Darley, *Morality in the Law: The Psychological Foundations of Citizens' Desires to Punish Transgressions*, 5 ANN. REV. L. & SOC. SCI. 1, 9–14 (2009) for a review of recent experimental game studies and a further discussion of the reasons that the participants remain anonymous to one another.

[6] The three studies done by Joshua Green's research group did fMRI imaging while respondents were judging the moral acceptability of actions proposed in various short scenarios: Joshua D. Greene et al., *An fMRI Investigation of Emotional Engagement in Moral Judgment*, 293 SCIENCE 2105 (2001) [hereinafter Greene, *An fMRI Investigation*]; Joshua D. Greene et al., *Pushing Moral Buttons: The Interaction Between Personal Force and Intention in Moral Judgment*, 111 COGNITION 364 (2009) [hereinafter Greene, *Pushing Moral Buttons*]; Joshua D. Greene et al., *The Neural Bases of Cognitive Conflict and Control in Moral Judgment*, 44 NEURON 389 (2004) [hereinafter Greene, *Neural Bases*]. De Quervain and Sanfey's studies discuss image while subjects were engaged in an experimental game. Dominique J.-F. de Quervain et al., *The Neural Basis of Altruistic Punishment*, 305 SCIENCE 1254 (2004); Alan G. Sanfey et al., *The Neural Basis of Economic Decision-Making in the Ultimatum Game*, 300 SCIENCE 1755 (2003).

scenario could detail a harm that one person caused to another, but in one variation the description of the harm would make clear that it had been caused negligently while in another variation it would have been caused deliberately.[7] In the studies to be discussed later, the information we altered in the scenarios concerned the various possible goals for sentencing offenders, such as retribution, just deserts, and incapacitation.

Generally, the dependent variable of the study is the magnitude of the punishment the respondent assigns to the actor in the scenario he or she examines. This is usually measured via the assignment of prison terms of variable durations. If the option of a death penalty is available, some but not all respondents will assign it, generally when the scenario involves an incident of murder.[8]

C. *Dual Process Theories*

To understand the findings of the scenario experiments, it is necessary to understand a distinction that is commonly made in psychology concerning the thought processes by which decisions are reached. Psychologists today often differentiate between reasoned decisions and intuitive decisions. Reasoned decisions are ones that follow the rules for good decisions laid down in books that advise us all about decision making. Intuitive decisions are the result of the heuristic decision-making processes that psychologists have discovered lead to many of our decisions in practice. These are the decisions that often "pop into" our minds. They are triggered into action automatically and take place rapidly. Importantly, they are implicit processes, meaning that their workings are not accessible to conscious scrutiny. Therefore, we are not aware that they have taken place. In this sense they are like our perceptual processes, which are similarly constructive in character but do not seem that way to the perceiver. The intuiting decider, like the perceiver, often emerges certain of the correctness of his intuitions.[9]

An important point to make here is that many decisions may be made in either an intuitive or reasoning mode. Thus this class of theories has been called dual process theories, a good many of which have been discovered.[10]

---

[7] For a number of these studies asking subjects to rate transgressions done with varying levels of intentionality, contribution, or culpability, see PAUL H. ROBINSON & JOHN M. DARLEY, JUSTICE, LIABILITY, AND BLAME: COMMUNITY VIEWS AND THE CRIMINAL LAW (1995). Appendix A discusses the research methodology of scenario studies, and Appendix B gives the text of the scenarios used. *Id.* at 217–28, 229–81.

[8] For this study's measurement scale and a discussion of its use, see *id.* at 223–25.

[9] The distinction between intuitions and reasoning process is discussed in Daniel Kahneman, *A Perspective on Judgment and Choice: Mapping Bounded Rationality*, 58 AM. PSYCHOLOGIST 697 (2003). Also, note Kahneman's discussion of "natural assessments." *Id.* at 701. It is likely that the assessment of actions as morally good or bad is a natural assessment.

[10] *See* DUAL-PROCESS THEORIES IN SOCIAL PSYCHOLOGY (Shelly Chaiken & Yaacov Trope eds., 1999); *see also* Steven A. Sloman, *Two Systems of Reasoning*, *in* HEURISTICS AND BIASES: THE

D. *Punishment Scenario Study Results: Intuitively Made and Just Deserts Based*

One of the first points to notice is that punishment-duration judgments seem relatively unproblematic for respondents to make. If respondents are judging multiple scenarios, they read each scenario and respond quite quickly with a sentence duration. Watching experimental subjects work through the various scenarios they read and assign punishments to, we are also struck by the rapidity with which they do so, a clue that the process is an intuitive rather than reasoned one. Like all intuitive judgments, information is being processed, but which information is actually relevant to the punishment judgment and how it is combined to make the judgment cannot be reported by the subject.[11] This is because intuitive judgments are automatically made and fire rapidly, thus the decision processes are not available for conscious introspection. The scenario experiment is designed to enable the experimenters to determine which bits of information inserted in the scenario are being used in the subject's decision, information that the subject cannot provide.

In the last decade, scenario experiments have been done to determine which of the various proposed reasons for punishing wrongdoers are actually influencing the punishment assignments of citizens.[12] These studies use the classic justifications for punishment that are usually given in the literature, including general and specific deterrence, incapacitation of dangerous individuals, just deserts, and rehabilitation. Experimenters insert information relevant to one or more of these justifications into the scenarios that the subjects read and test to see which insertions alter the punishment decisions of different groups of subjects. The method works like this: A story element is created that can be manipulated to indicate that a criminal's punishment should be either relatively severe or relatively mild if a person is judging from a just deserts perspective. If the crime is embezzlement, the punishment should be severe if the actor embezzled to continue a life of debauchery but milder if the actor embezzled to purchase a medical treatment for an otherwise dying child. Each story is given to a different group of subjects. If the group reading about the debauched individual consistently assigns higher sentences to his embezzlement, then the subjects are assigning punishments, to some extent, according to what the actor justly deserves for his crime.

By slightly altering the study's design, it can be made to answer a comparative question by enabling comparison between the strength of two or more

---

PSYCHOLOGY OF INTUITIVE JUDGEMENT 379 (Thomas Gilovich et al. eds., 2002); Steven A. Sloman, *The Empirical Case for Two Systems of Reasoning*, 119 PSYCHOL. BULL. 3 (1996).

[11] *See* Richard E. Nisbett & Timothy DeCamp Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231 (1977).

[12] *See* John M. Darley, Kevin M. Carlsmith & Paul H. Robinson, *Incapacitation and Just Deserts as Motives for Punishment*, 24 LAW & HUM. BEHAV. 659 (2000); *see also* Kevin M. Carlsmith, John M. Darley & Paul H. Robinson, *Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment*, 83 J. PERSONALITY & SOC. PSYCHOL. 284 (2002).

justifications for punishment.[13] To demonstrate this by example, a scenario is made to contain information relevant to the just deserts perspective and the incapacitation perspective. The information on the incapacitation perspective could be that the person has not committed any crimes before or has committed several similar crimes in the past. The reader will (and does) believe that the latter individual is more likely to recidivate than the former. If four groups of subjects are run, the magnitude of the difference in the sentence that is due to receiving the high or low just deserts information can be compared to the magnitude of the difference in the sentence brought about by the high or low recidivism information. This form of the experiment can answer a somewhat more sophisticated question about the comparative strength of the relevance of the various perspectives to the respondents' sentencing decisions.

The results of these sorts of studies show that it is generally the just deserts information that drives the magnitude of the sentence assigned when the decision is being made according to what is likely to have been intuitive processes.[14] This suggests that people's intuitive responses to wrongdoing are primarily driven by retributive considerations. In one study in which respondents were allowed to acquire information relevant to the different sentencing perspectives in a serial fashion, respondents began by acquiring information relevant to the retributive perspective.[15] Some respondents then chose to acquire information relevant to the incapacitative perspective, which slightly influenced their sentencing decisions.[16]

Suppose for a moment that we accept the conclusion that people's initial responses to various transgressions are intuitively generated, and the sentence durations assigned are based on a rapid assessment of what the scenario actor justly deserves for his actions. A set of studies in which subjects' brains were imaged while they were reading scenarios adds brain-locational information about these intuitive judgments.[17] In these studies, different scenarios described certain actions, some of which were harmful in effect. The task of the subject was to allow or disapprove of the postulated action of the actor in the scenario, which sometimes involved harming another or cheating to gain advantage. For the actions that harmed others, which the investigators labeled "personal moral dilemmas,"[18] activated brain areas that previous research had demonstrated were associated with the registration of emotional and social-cognitive processes, specifically the medial prefrontal cortex, posterior cingulate/precuneus, and

---

[13] For this design, see sources cited *supra* note 12.

[14] *See* sources cited *supra* note 12.

[15] Kevin M. Carlsmith, *The Roles of Retribution and Utility in Determining Punishment*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 437, 444–45 (2006).

[16] *Id.* at 445.

[17] *See* Greene, *An fMRI Investigation*, *supra* note 6; Greene, *Neural Bases*, *supra* note 6; *see also* Greene, *Pushing Moral Buttons*, *supra* note 6.

[18] Greene, *An fMRI Investigation*, *supra* note 6, at 2107; Greene, *Neural Bases*, *supra* note 6, at 390.

superior temporal sulcus/temporoparietal junction.[19] These processes generated quick reaction-time decisions. It seems plausible that these brain areas are involved in producing the rapid, intuitively produced decisions we mentioned before.[20]

In one particular story, a different pattern of judgments and underlying brain-activity patterns appeared. The action posed in this story was to smother a crying baby, which first provoked the fast reaction-time pattern in the areas of the brain listed above. However, the story was a version of the horrible moral dilemma that is often used to illustrate the difference between a utilitarian and a deontological moral code: The baby's cries would alert enemy soldiers, who would find the hidden group and then kill them all, including the baby.[21] For this story, a second set of temporally slower brain processes occurred, taking place in areas associated with higher-order reasoning and decision conflict-management processes.[22]

Returning to our previous comments, for this case it is a *dual process theory* that is being proposed. The essence of the dual process claim, which is being made in many areas in which psychologists have investigated decision-making processes, is that, depending on the decision-making process activated or relied on, the decision can not only be a different one but an opposing one. According to this account, dual processes contribute to moral judgments. One process, produced relatively rapidly, is the product of social-cognitive and emotional responses and takes place non-optionally; this is the intuitive system discussed above. The second process involves abstract-reasoning areas of the brain, ones that developed evolutionarily later than did the emotional and social-cognitive brain areas.[23]

When this account is applied to decisions based on either intuitions or reasons, it contains another clause. The intuitive processes are always triggered into action, but the reasoning processes are not always triggered. Furthermore, when this reasoning system is activated, its results are sometimes in conflict with the intuitions of the other system. The reasoning system that is giving us the utilitarian result and overriding the impulse against killing a single other individual is acting to monitor or limit the intuitive system result. It is this override response that is possibly (but that psychologists suggest is not always or even often) triggered into action for the punitive judgments we are considering here. As

---

[19] Greene, *Neural Bases*, *supra* note 6, at 390; *see alsò* Greene, *An fMRI Investigation*, *supra* note 6, at 2106.

[20] *See* Greene, *An fMRI Investigation*, *supra* note 6; *see also* Greene, *Neural Bases*, *supra* note 6; Greene, *Pushing Moral Buttons*, *supra* note 6.

[21] Greene, *Neural Bases*, *supra* note 6, at 390–91.

[22] *Id.* at 395–96. The broader implications of this story suggest that utilitarian and deontological decisions are often made via different brain processes. *Id.* at 398.

[23] The neural machinery producing these two systems is discussed in Analysis 1 in *id.* at 393–95.

Kahneman commented, the reasoning system is a rather lax monitor of the outputs of the intuitive system.[24]

III. CONTRIBUTIONS OF EXPERIMENTAL GAMING RESEARCH

The scenario sentencing studies discussed above suggest that punishment decisions are made on retributive grounds and are often made intuitively. The experimental games studies enable us to see if these decisions are carried out when the respondents are able to inflict actual punishments on actual transgressors. These experimental games that examine the respondents' punishment patterns are elaborately staged interactive occasions in which players make moves, generally via computer, that register their decisions and actions in one trial of the game. It is often the case that respondents play on repeated trials but generally they understand that these new trials are played with a rotating set of new players. Generally, either after a trial or as part of their actions on a successive trial, respondents have the opportunity to inflict punishments on other players who have committed actions that morally offend them. By and large, the actions that are experienced as violations are those that take advantage of another who has acted cooperatively and as such are violations of implicit social norms. Thus these situations are better called "trust-violation games." Generally, these games are "played for points" and players seek to win as many points as possible.[25]

In early studies, a great deal of punishment was actually inflicted. Generally the mechanism of punishment, imposing "fines," was done by subtracting points from an individual's hoard. But critics of the research pointed out that often the points were just that, "points" as in a board game, without actual cash value, so "punishments" inflicted on others existed only within the structure of the game and game results could not be generalized to behaviors we would expect in the real world. For this reason, later experiments have made the game points payable in real-money terms. Taking advantage of inflation and floating exchange rates, these games have been played for sums equal to or exceeding a day's wages. Further, those players who inflict punishment on others must "pay to play," meaning that now, a participant must purchase with her points the amounts of the fines she can inflict on another. The cost to the person who inflicts the fine is generally lower than the amount that the fine extracts from the punished individual.

In some of these games, social norms exist that create strong pressures against any player making certain advantage-taking moves, so it is rare that a real respondent will make one of them. The experimenter now has a dilemma because she wishes to see what the reaction of the other respondents would be to a respondent's advantage-taking action. When this is the case, the experimenter will

---

[24] *See* Kahneman, *supra* note 9, at 699.

[25] For a more elaborate description of the procedures of experimental games, and the motivations for, example, keeping the participants anonymous, see Darley, *supra* note 5, at 9–14.

arrange for a "confederate" to be present. The confederate is a person who the other respondents perceive is a normal player, like themselves, but the confederate is actually there to carry out the experimenter's directions to act in a way that violates the social norms of the situation. At some point during the exchanges, the confederate acts to "betray the trust" of one or more of the respondents. In other game situations, such as the well-known prisoner's dilemma, the forces toward trust-violation are high enough that an actual player will defect or otherwise violate the trust of others.

In any event, whether an actual player or a confederate inflicts it, a trust violation occurs. The person whose trust is violated then has at his disposal a way of inflicting a punishment on the person who violates his trust. Generally, this involves imposing a fine that removes points (worth money) from the violator's holdings. Usually the punisher must pay points to inflict the fine on the violator, but the cost of inflicting the fine is generally lower than the fine inflicted. This roughly models the real world situation in which choosing to inflict punishments comes with a cost to the punishing individual.

The researchers in these studies have sometimes been able to image the brains of the players at critical occasions, such as when they are deciding whether to punish a transgressor, and interesting information on the brain areas involved in these decisions has emerged. Other researchers have been able to arrange to carry out these studies in cultures that are quite different from our own. By doing this they have obtained results that give us early-stage information about the universality of the impulse to punish trust transgressions in human societies.

All this being said, the results of these studies can be quickly summarized.[26] The ultimatum game is a simple and ingenious way to investigate reactions to failures to share resources with another. In the ultimatum game, a sum of money is given to one individual who becomes the "decider." She can share as much or as little of that sum with another person, called the "responder." The responder then decides whether to accept the proposal. If she accepts it, she gets as much or as little as the decider offered and the decider keeps the rest. If she rejects the proposal, nobody keeps anything.

From a rational choice perspective, the responder should accept any amount offered, since she will end with more than she would get otherwise. However, the studies show that this is not what happens. One team of researchers administered the ultimatum game in over a dozen different cultures.[27] At a broad level of description, responders in all cultures showed an increasing rate of offer rejection as the offer diminished from a 50–50 split of the resources. In some cultures, the rejections began quickly when the split offered was only slightly below equality,

---

[26] *See id.*

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

while in other cultures rejections become frequent only when the offers became more disadvantageous to the responder, such as a 70–30 or 75–25 split.[28]

In another study of the ultimatum game done in the United States, responders' brains were imaged.[29] 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.[30]

In a trust-game study in which one trial was conducted against multiple other players, a "trusting" player would transfer money to another player with the implicit but strong expectation that the profits of this transfer would be shared by the "trustee."[31] Several times the trustee shared none of the profits. The betrayed trusting player then could "fine" the trustee in two unit increments, a large share of his unshared profits, but had to pay one unit for every two units he fined the trustee.[32] Consistently, betrayed trusting players chose to expend their resources to inflict punishments on an untrustworthy other player. Doing this, as the authors commented, "activated the dorsal striatum, which has been implicated in the processing of rewards that accrue as a result of goal-directed actions."[33] This supported their hypothesis that "people derive satisfaction from punishing norm violations" and anticipate satisfaction from doing so.[34]

In yet another scanning study, a participant observed another participant behave in an untrustworthy way toward a third participant who could not inflict any punishment on the double-crossing actor.[35] 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.[36] This latest result, demonstrating that a third party is sometimes willing to expend resources to punish a person who has transgressed against another, is an interesting one. It is referred to in the literature as either "third party punishment" or "altruistic punishment." The latter name suggests that there are no rewards achieved by the punisher and even some costs incurred. It is true that there are no financial rewards for the third party, but since the third party does have the satisfaction of inflicting a punishment on the offender, the term "third party punishment" is preferable.

---

[28] *Id.* at 1769–70.

[29] *See* Sanfey, *supra* note 6.

[30] *Id.* at 1756–57.

[31] *See* de Quervain, *supra* note 6, at 1254.

[32] *Id.* at 1255.

[33] *Id.* at 1254.

[34] *Id.*

[35] *See* Ernst Fehr & Urs Fischbacher, *Third-Party Punishment and Social Norms*, 25 EVOLUTION & HUM. BEHAV. 63 (2004).

[36] *Id.* at 66.

The research using the variation of the ultimatum game called the "dictator"[37] game demonstrates the generality of this effect. In this version of the game, the dictator has decision-making power such that whatever share of the total he proposes is what the recipient gets, and the recipient has no option of refusing and thus destroying the share the dictator keeps for himself. As the researchers reported, "[a]t each transfer level below 50, roughly 60% (n = 22) of players C [third parties] chose to punish the dictator A."[38] As the shares proposed by A grew more selfish, the amount of punishment imposed on A by C became larger. Past a 50–50 split, every 10 units awarded by A to himself caused C player to assign a fine of 8.4 units on A.[39]

The researchers also tested the frequency and magnitude of third party punishments inflicted on non-cooperators playing a variation of the prisoner's dilemma game. Here too, third party punishments were frequent. Forty-six percent of third parties punished a player who defected when the other player cooperated.[40] In general, the punishments inflicted by the third party punishers reduced the profitability of the transgressions but were not high enough in frequency and magnitude to make the transgressions unprofitable. As the researchers noted, in the real world, where more than one third party is likely to witness a person's transgressions, third party punishments from multiple sources "are likely to be powerful enforcers of social norms."[41]

For our purposes, the experimental games literature is generally consistent with and usefully extends the results of the scenario studies. In the scenario studies, it was possible to ascertain the goals that the respondents sought to reach when they assigned punishment durations to crimes described in the scenarios, but it was perfectly possible to be skeptical about whether the respondents actually would inflict those hypothetical punishments when faced with real offenders. The respondents in the experimental games were quite willing to inflict real punishments on other players. And in the third party punishment games, they would punish players who had transgressed against others rather than themselves. Many see these experimental games as demonstrating the natural utility of punishment practices for controlling moral deviance in small, face-to-face groups and, by extrapolation, see the "naturalness" of punishment mechanisms and institutions in larger scale societies.

---

[37] *Id.* at 65.

[38] *Id.* at 68.

[39] *Id.*

[40] *Id.* at 73.

[41] *Id.* at 85.

IV. IMPLICATIONS FOR PENAL PRACTICES IN OUR SOCIETY

A central question for those who are involved in criminal justice policy making is the degree to which citizens' perceptions are relatively fixed on one hand, or to some extent malleable on the other hand, on such issues as: what actions are sufficiently grave to count as crimes, what duration of punishment these crimes deserve, and what the correct purposes are for those punishments. The findings sketched in this article, and the interpretations made of them, may have implications for these questions.

Our suggestion is that people's thinking about these issues generally proceeds at the intuitive level, and by now the reader will understand that this claim is a far-reaching one. Citizens presented with a description of a crime generally have a rapid and automatic perception of what punishment that crime deserves and can associate that with a desired level of our culture's usual metric for expressing punishment severity: prison term duration. Coupled with this judgment, people generally have strong convictions that this punishment is one that society is required to carry out; we find it hard to imagine a society that does not punish actors who commit the prototypical serious criminal acts such as rape, murder, theft, and fraud. Elsewhere we have argued that a society that has in place a criminal justice code that is persistently deviant from the shared moral intuitions of the community risks losing the unforced obedience of its members to its laws, and we think that a society that consistently does not punish violations of core prohibitions, such as violence toward others, theft of property, fraud, and failure to keep contracts, would quickly become predatory.[42]

Many reformers consider the current punishment practices of the United States to be absurdly punitive.[43] Our long duration sentences, generally primitive prison conditions, absence of any rehabilitative training, and a long list of other practices, such as a parole system oriented toward detecting trivial violations of parole conditions and returning the inevitable offenders back to prison, seem to these critics to be self-defeating. But it may be that this is exactly what common citizens wish it to be, given that this account of the "psychology of punishment," sketched above, finds a retributive strain in people's thinking.[44] It is imperative to discuss whether pessimism about change is warranted. We think that the evidence suggests that change toward more socially functional practices will be possible, although achieving them will require energetic advocacy and frequent demonstrations of the practicality of the changed practices. The present view

---

[42] *See* Paul H. Robinson & John M. Darley, *The Utility of Desert*, 91 NW. U. L. REV. 453 (1997).

[43] *See, e.g.*, MICHAEL TONRY, THINKING ABOUT CRIME: SENSE AND SENSIBILITY IN AMERICAN PENAL CULTURE, at vii (2004) ("From capital punishment to three-strikes-and-you're-out to the highest imprisonment rates in the Western world by a factor of five, the United States stands alone in what it does to its citizens to prevent crimes and punish criminals.").

[44] *See* sources cited *supra* note 12.

allows for and can even explain a good deal of disagreement on aspects of the punitive subsystem of the criminal justice system, which is a good thing since we see that a good deal of actual disagreement exists within the society of the United States as well as in other societies. Several elements in the psychological account given here create the spaces in which disagreement can exist.

Recall that we said that there are two general kinds of thought processes that are mobilized by people making decisions on criminal punishments. The intuitive processes, it was suggested, deliver punishment-duration judgments normally based on just deserts intuitions and deliver these judgments automatically in a way that creates strong perceptions of "rightness" on the part of those who hold these perceptions.[45] Further, the studies cited demonstrate a great deal of consensus on these judgments among citizens.[46]

But this is a dual system situation. It is the existence of the reasoning system that creates the first possibility by which individuals, or more likely opinion-sharing groups, can come to conflicting judgments on what constitutes a crime and what are the appropriate purposes of criminal "punishments." If a person has previously reasoned to a general conclusion about some criminal issue, and if the new case is one that should be decided by that general conclusion, then it may be so decided. Whether it actually is decided in this fashion depends on whether the person recognizes that the general conclusion previously reached is applicable to the current case. It is this recognition that causes the reasoning process to be triggered into action, which then overrides the assessment produced by the person's intuitions.[47]

Judges are one group of people who we would expect to have developed a carefully thought-through and consciously-reasoned approach to assigning punishments to offenders. John Hogarth's monumental study, begun in 1965 and published in 1971, examined different influences on the sentencing practices of seventy-one full time magistrates in the province of Ontario.[48] The study is complex, detailed, and not easily summarized, but it is clear that there are some quite different philosophies of punishment held by different judges and that these different philosophies, to some extent, drive their sentencing assignments. Interestingly, in this particular study, not many judges regarded the retributive goal that we have argued is the primary driver of the sentencing judgments made by people reacting in the intuitive mode as an important purpose of sentencing.

Hogarth had the magistrates rate the relative importance of five of the classical purposes in sentencing: reformation, general deterrence, individual

---

[45] *See* sources cited *supra* note 12.

[46] *See* Robinson & Kurzban, *supra* note 2; *see also* Paul H. Robinson, Robert Kurzban & Owen D. Jones, *The Origins of Shared Institutions of Justice*, 60 VAND. L. REV. 1633 (2007).

[47] *See* Kahneman, *supra* note 9, at 697–701.

[48] JOHN HOGARTH, SENTENCING AS A HUMAN PROCESS 15–16 (1971).

deterrence, incapacitation, and punishment.[49] The definition of "punishment" he gave to the magistrates was: "The attempt to impose a just punishment on the offender, in the sense of being in proportion to the severity of the crime and his culpability."[50] This surely is the retributive impulse that we previously argued drives the intuitive response to crimes made by the subjects in the psychological studies that we reviewed.

However, only about 10% of the magistrates rated the punishment purpose as very important. In comparison 57% rated reformation as the most important purpose.[51] General deterrence was rated as very important by 38% of the magistrates, probably reflecting the general belief held by most criminal justice scholars that having a known sanction in place for crimes serves as a general deterrent against crime commission.

Hogarth asked the magistrates about the sentencing presumptions that they held, and the answers seemed to reveal how their different sentencing strategies were turned into concrete sentencing decisions.[52] For instance, when asked about what situations would lead to a presumption that prison was appropriate, 64% reported that crimes of violence would create this presumption while 22% presupposed that prison was appropriate for all crimes that involved violations of trust.[53] Overall, the magistrates rated prisons as highly effective in inflicting punishment.

An interesting argument is made about magistrates who are oriented toward reformation goals in sentencing. They regard probation as a highly effective sentence to give, which makes sense since it would allow for some reformative activities. But, Hogarth points out, realities are such that the magistrate must send "at least some offenders to prison."[54] This causes him some dissonance: he does not want to regard himself as a punitive judge. "The easiest way out of the dilemma is to see prisons as therapeutic institutions. If prisons could be seen as hospitals, then the magistrate has resolved all discrepant beliefs."[55] And, oddly to our current views of what takes place inside prisons, the magistrates rated prisons as possible locations for rehabilitation.

Hogarth has many more things to say about the differences in the sentencing decisions of his magistrates, but the general conclusion is clear. Different subgroups of magistrates have thought carefully about the reasons that they sentence, come to different ways of reasoning, and applied that reasoning to the cases that are presented to them. One other thought is possible to extract: The

---

[49] Sixty-eight judges rated the importance of the five purposes on a scale ranging from very and quite important to some, little, and no importance. *Id.* at 70–71.

[50] *Id.* at 70.

[51] *Id.* at 71.

[52] *Id.* at 77–79.

[53] *Id.* at 78.

[54] *Id.* at 77.

[55] *Id.*

system of presumptions, in which recognition by the magistrate that the offense the defendant has been charged with is a crime of violence causes the judge to have a presumption toward assigning a prison sentence to the criminal, may show us how what is originally an elaborate chain of reasoned thinking is moving toward being more automatic, and more intuitive, in character.[56] Decision researchers comment that many decisions, repetitively made, move toward being made more intuitively.

There are examples of individuals and groups who have consciously decided to base their punishment decisions on some reasoned strategy that deviates from the standard thinking in their culture. And their deviant reasoning may move toward being intuitive for them. One of the most poignant recent examples comes from Lancaster County, Pennsylvania. In 2006 a neighbor entered a small Amish schoolhouse, dismissed most of the students except ten young girls, and shot and killed five of the girls before killing himself.[57] Members of the Amish community quickly took actions to signal forgiveness to the family of the killer. One author who has studied the Amish attitudes toward forgiveness pointed out that their attitude is an ingrained one; "[t]heir religious tradition predisposes them to forgive even before an injustice occurs."[58]

This quotation contains an interesting suggestion that the community has managed to so internalize the response of forgiveness that it has been transformed from a reasoned override of a retributive intuition to the intuitive product to at least some moral wrongs. Dual process theoreticians have noted this possibility, although it is difficult to recruit many examples of it. It is possible that this kind of transformation occurs most easily within a community that holds a different account of what the intuition should be and rehearses that account in their religious practices and daily belief enactments.[59]

Finally, although speculatively, on a historical note, it is worth registering that punishment scholars who study the broad changes in punishment inflictions over historical periods are struck by the changes that have prevailed in some societies.[60] In some societies, death by prolonged torture was common for certain offenses against the state. In others, the kin of offenders were executed as well as the offenders for various crimes. In more recent eras, flogging, branding, and the

---

[56] This speculation could be tested by observing the changes in a judge's decision-making patterns over time. More experienced judges would be predicted to make prison sentencing decisions more rapidly and elaborate less on their reasons for doing so.

[57] *See* Donald B. Kraybill, *Why the Amish Forgive So Quickly*, CHRISTIAN SCI. MONITOR, Oct. 2, 2007, at 9, *available at* http://www.csmonitor.com/2007/1002/p09s02-coop.html.

[58] *Id.*

[59] The notion here is that if a group is like-minded about their opinion on a particular issue—in this case the necessity for forgiveness—then over time that belief is not challenged and reasons for it need not be produced. Over time, the production of the opinion becomes automatic and its correctness is not challenged by the opinion holder. *See* CASS R. SUNSTEIN, GOING TO EXTREMES: HOW LIKE MINDS UNITE AND DIVIDE (2009).

[60] *See* DAVID GARLAND, PUNISHMENT AND MODERN SOCIETY: A STUDY IN SOCIAL THEORY (1990).

stocks were common punishments for different, specified crimes with some consensus on which crime warranted which punishment. Our culture's range of punishments in the early years of the 21st century is considerably less than the possibilities exploited in previous times. If it is the case that the punishments assigned in prior eras were those that the citizens thought were appropriate, then there has been a considerable change in public opinion about what punishments were thought fitting by citizens over the centuries. In a certain sense, the magnitude of sentences has been reduced. Many nation states have dismissed the death penalty; even the United States has moved toward reserving the death penalty for only the most serious offenses. Also, sentences that involve the infliction of acute physical pain, and sometimes bodily mutilations, have been eliminated in some, but not all, cultures.

V. CONCLUSION

In the first sections of this article, we suggested that most people's first responses to crimes were automatic and intuitive in character, normally driven by just deserts considerations and thus retributive in character. Examining evidence from experimental games, we observed that actual actions that were clear norm violations, with the norms in question being about fair distributions and reciprocating trusting behaviors for the common good, provoked retaliatory punishments. Punishments came first from the people who were the targets of the norm violations but also came from others who witnessed the norm violations even though they were not the direct victims of them. Also, the inflictions of those punishments engaged brain centers associated with the performance of rewarding actions.

Based on this, it seems right to say that in the societies studied, a "natural response" to moral norm violation is punishment generated for retaliatory purposes. In the society of the United States, this natural tendency is amplified by certain cultural forces, such as the constant portrayal of violent criminal actions by the entertainment media and the television and newspaper news outlets. This contributes to broad-based fear of crime on the part of the populace. Further, it creates fear among politicians of appearing "soft on crime," lest some other politician gain advantage when running against them.[61]

All of this has certainly contributed to what many scholars consider the excessively severe and punitive crime control practices prevalent in our society at this time. But it does not mean that those practices are inevitable. The later stages of this article turned to the question of whether this retaliatory orientation in the world of criminal justice is unavoidable. Briefly, it is not. People are capable of reasoning about criminal justice practices and absorbing information that there are more enlightened ways, for instance, of administering parole supervision that is oriented toward sustaining parolees in the work force rather than "violating them

---

[61] *See* TONRY, *supra* note 43, at 136.

out" and back into prison. Some people have already overcome their punitive intuitions and replaced them with intuitions oriented toward forgiveness or restorative justice practices.

That a tendency toward a retaliatory perspective exists is useful to know. But it is not inevitable that it dictates and controls our criminal justice practices.