Morality in the Law: The Psychological Foundations of Citizens' Desires to Punish Transgressions

John M. Darley

Department of Psychology, Princeton University, Princeton, New Jersey 08540; email: jdarley@princeton.edu

Annu. Rev. Law Soc. Sci. 2009. 5:1–23

First published online as a Review in Advance on July 29, 2009

The *Annual Review of Law and Social Science* is online at lawsocsci.annualreviews.org

This article's doi: 10.1146/annurev.lawsocsci.4.110707.172335

1550-3585/09/1201-0001$20.00

Key Words

just deserts, deterrence, retribution, prison sentences, culpability

Abstract

Evidence from a number of research methods converges to suggest that when a person registers a transgression against self or others, the person experiences an intuitively produced, emotionally tinged reaction of moral outrage. The reaction is driven by the just deserts–based retributive reactions of the person to the transgression rather than, for instance, considerations of the deterrent force of the punishment. In experimental games arranged so that trust and fairness transgressions occur, participants punish transgressors and experience rewarding brain states while doing so, and they punish even if they were not themselves the target of the violation. What, if any, implications does this have for the punishment component of societal systems of justice? Would it be possible to construct sentencing practices that, to some extent, incorporated citizens' sense of just punishments? What would be gained by doing so? And what would be lost?

INTRODUCTION

After a static period, research in moral psychology has reached an active stage once again and draws on many sources of influence: philosophy, several branches of psychology, work with experimental games, and most recently, neural imaging of brain functioning. There now exists a field of experimental philosophy that employs empirical investigations of moral philosophical issues.

In this article, I summarize some discoveries of this research: first, because a relatively clear picture of the naive psychology of punishment emerges; second, because the evidence from different modes of experimentation converges to support this emerging consensus; and third, because these discoveries begin to have implications for legal institutions and legal practices, a topic of particular interest to those who study the intersections of the law and social sciences.

The notion that the moral norms of the community have a claim to be reflected in the legal norms governing that community is one that has frequently been taken for granted, although it sometimes has been contested by other claims for code-setting principles, such as law and economic considerations. What the current research activities in the area allow us is a better descriptive take on the content of citizens' moral rules and the workings of citizens' moral thinking. This in turn allows for clearer discussions of whether these rules and reasoning processes have any claim for shaping legal institutions and what the various aspects of this claim might be.

The topic of punishment provides a good place to begin seeing the relevance of people's moral thinking to legal institutions and practices. Large, modern societies generally have authoritatively established legal codes, a law-promulgating authority to set those codes, a policing system to detect violations of those codes, and a justice apparatus to adjudicate punishments for violations of those codes. The codes of most interest, in terms of moral psychology, are the criminal codes: that subset of legal codes that assign major punishments to convicted offenders. But although the primary focus here is on them, criminal codes are not the only set of legal codes that citizens perceive as doing important moral work. Negligence law deals with punishments for harms one citizen carelessly inflicts on another, and contract law deals with what one citizen owes another, how we make commitments to one another, and what penalties are to be extracted when one fails to keep one's commitments.

In this article, I first give a phenomenological characterization of the feelings persons have when they become aware of a transgression: a feeling of moral outrage. Next, I suggest that this feeling is produced by humans' intuitive systems and roughly computes what the transgressor justly deserves based on the moral wrongness of the transgression. I then turn to the experimental game literature to examine the actual use of punishments to alter the frequency of norm violations in small group settings. Finally, and rather tentatively, I consider the implications of people's normal punishment cognitions and practices for the institutions of criminal justice.

THE DESIRE TO PUNISH

Psychologists have observed that when individuals perceive that an injustice has been inflicted on themselves or others, they generally have a strong and immediate desire to punish the offender. The psychological processes that bring about this desire need to be specified. First, however, a more detailed description of the reaction proves useful.

Moral Outrage: The Phenomenology of Injustice

From a psychological perspective, the long and initially disparate list of actions that people regard as serious inflictions of injustice, actions that often are regarded as appropriate targets of criminal sanctions, are in fact unified by the characteristic reactions they produce in the socialized member of the culture. Miller (2001, pp. 534–35), in his *Annual Review of Psychology*

article on injustice, gives an excellent description of this feeling that is produced by experiencing or witnessing one of these unjust acts:

> Injustices have a transcendent quality, which is one reason that it is more legitimate to respond to an injustice than to that which is merely an insult. . . . To label an insult an injustice transforms it from a personal matter to an impersonal matter of principle. . . . In short, a personal insult that is labeled an injustice becomes a collective injustice, and avenging the injustice becomes a defense of the honor and integrity of the entire moral community. . . .
>
> The 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. . . . Individuals are committed to the "ought forces" of their moral community, as Heider (1958) termed them, and people believe that these forces deserve respect from all members of the community. The violation of these forces represents an insult to the integrity of the community and provokes both moralistic anger and the urge to punish the offender in its members. Viewed from this perspective, disinterested justice reactions are not disinterested at all, because everyone has a stake in seeing that the rules and values of the authority structure under which they live are respected.

What Produces Moral Outrage?

Let us provisionally accept that description of the feeling of moral outrage and examine how that feeling is produced. Several strands of cognitive, behavioral, and neural evidence have recently converged and allow us to specify further the psychological characteristics of this moral outrage reaction.

PUNITIVE REACTIONS AS INTUITIONS

Psychologists studying human judgment and decision processes now distinguish between two broadly different ways that people come to decisions and judgments: One involves heuristic, intuitive processes, and the other involves reasoning processes. I suggest that these desires to punish are often the product of intuitive rather than reasoned processes. (For an expanded discussion of the consequences of the view that punitive reactions are often intuitions, see Robinson & Darley 2007.)

What are the characteristics of intuitive processes? Strikingly similar to the processes involved in visual perception, intuitive processes are rapid, can proceed in parallel with other mental processes, and are automatic and effortless in operation. They are implicit; that is, they are not available to introspective analysis and are frequently emotionally loaded (Kahneman 2003; Sloman 1996, 2002; Stanovich & West 2002). Often described as ballistic in nature, they are habit driven and thus difficult to modify once the processes begin.

If, however, the intuitive system processes have the rapid and nonconscious properties of the perceptual system, the intuitive system products are not percepts but rather are judgments, decisions, and other kinds of conceptual representations (Kahneman 2003): Their products are like the products of the reasoning system.

As is well known, people experience their perceptions as simple, correct representations of "what is out there"; that is, people experience the perceptual world in the mode of a naive realist. In turn, intuitions, like perceptions, are often taken by the intuitor to be unproblematically correct. Thus, a set of intuition-produced decisions, choices, and problem solutions are experienced as summaries of the ways the world is, because the person having the intuitions is unaware of the complex, potentially incorrect, cognitive processes that produced them.

As mentioned above, when respondents receive a scenario in which some person commits a known morally wrong action, respondents experience a reaction of moral outrage; this becomes a substantial predictor of the relative punishments that will be assigned to the perpetrator of the immoral action. I suggest that this feeling of moral outrage is the conscious

registration of the intuitive reaction to instances of moral wrongdoing. Kahneman (2003, p. 701) comments that the assessment of the degree of badness represented in a stimulus is an intuitive system judgment:

> Some attributes, which Tversky & Kahneman (1983) called natural assessments, are routinely and automatically registered by the perceptual system or by [system 1 (this is the intuitive system)] without intention or effort. . . . The evaluation of stimuli as good or bad is a particularly natural assessment. The evidence, both behavioral and neurophysiological, is consistent with the idea that the assessment of whether objects are good (and should be approached) or bad (and should be avoided) is carried out quickly and efficiently by specialized neural circuitry.

To complete the story, the processes of the reasoning system are dissimilar to the processes of the intuitive system. Reasoned processes are relatively slow, serial, step-by-step in nature, often consciously monitored and controlled, effortful, and rule-governed in nature. Many, perhaps most, of our injustice-produced moral reactions are intuitively produced. Circumstances, however, may provoke the individual into reasoning about the case, and the reasoning system conclusion can override the response dictated by the intuitive system. This, then, is a dual process account, in which two different ways of thinking can, depending on circumstances, be brought to bear on one judgment or decision.

The Inability to Recover Grounds for Intuitions

In the case of the moral reactions discussed here, evidence suggests that the rapid, intuitive processes produce an initial reaction, essentially the feelings of moral outrage mentioned above. Several researchers have demonstrated results that support various elements of this claim. In Haidt's (2001) well-known work on moral dumbfounding, he constructed a number of scenarios involving actions that people would consistently identify as morally wrong (for instance, cooking and eating the family pet that just died, or brother-sister intercourse). However, he also constructed the stories to make clear that no harm, physical or psychic, was inflicted on any of the story characters. Respondents did immediately identify those actions as wrong. But when challenged by a questioning experimenter to produce reasons why these acts were wrong, the respondents would generally cite a harm that the action produced. But, as the researcher then pointed out, the story ruled out the possibility of the harms that the respondents cited. Eventually the respondents realized that they could not produce the harm-based reasons about why they judged the act wrong. Characteristically, they still claimed the acts morally wrong, but they just could not produce reasons for their wrongness. This is characteristic of intuitively produced responses; they are strongly held, but the sources of information on which the judgments are made are sometimes not retrospectively accessible to the decision maker.

Further evidence that many moral responses are intuitively reached comes from work on "trolley car" scenarios. The core of the scenario told to respondents is that a trolley car has broken loose, is running downhill on its tracks, and will kill five workers who are further down the tracks. Their deaths, however, can be averted if a person in the story diverts the trolley car to a sidetrack. If this is done, one worker on the sidetrack will die. Sometimes, respondents are asked whether or not the person should divert the trolley, and other times they are told that the person did divert the trolley and are then asked whether that action is acceptable or not.

Multiple versions exist of these problems, which vary in exactly how the trolley can be diverted. Researchers have given selected versions of the problem to respondents and find that respondents vary in their responses. For instance, if the diversion involves throwing a switch, a very high percentage of subjects approve of throwing it, even though it kills one worker. If, however, the respondent, now imagining the person standing on a footbridge over the tracks, is told that the person

accomplishes the diversion by pushing onto the tracks a large man, also standing on the footbridge (his body, crushed by the trolley, will halt the trolley's forward progress, saving the five workers), far fewer of the respondents report that they approve of the action of pushing the man in front of the trolley. This and similar patterns of agreement and disagreement seem odd. The appropriate moral question should be: Does someone sacrifice one to save five? The two problems should produce the same answer, regardless of the morally irrelevant detail of accomplishing this by throwing a switch or pushing a bystander.

Hauser's research group (Hauser et al. 2007) has confronted subjects who have responded to both stories, with the usual contradiction between their decisions. Both involve sacrificing one to save five. Again, the subjects are unable to produce reasons that led them to the different conclusions but continue to insist that their two decisions are correct. It is this inability to report reasons for decisions coupled with the perseverance in the decisions that lead these researchers to characterize these decisions as intuitively made.

Neural Imaging Evidence

Greene and colleagues (2001, 2004) have done brain imaging work that supports the idea of a rapid intuitive system that generates initial reactions to scenarios of transgressions. The researchers brain imaged respondents who were presented moral problems and whose task was to approve or disapprove of an actor's postulated response to the problem. These problems included trolley car problems and simple moral judgments, such as whether it is appropriate to kill another or cheat to gain advantage. In this first study, they found that a set of problems titled "personal moral dilemmas" activated brain regions that previous research had associated with both emotion and social cognition activities. Other problems, labeled "impersonal moral dilemmas," caused increased activity in areas associated with abstract reasoning and problem solving. The trolley car problem that posed the choice of letting the trolley continue on its tracks and kill five persons or throwing a switch to shunt the trolley to a track on which it would only kill one was an impersonal moral dilemma. It engaged the more cognitive areas involved in moral reasoning, but not the emotional/social areas that form the other half of the complete moral processing system. The footbridge example, however, in which the possible trolley-stopping action involved pushing a specific other individual in the path of the oncoming trolley car, is a personal moral dilemma case: The action required is directly personal, in which "I act directly to harm a specific other person." That dilemma activated both the cognitive and emotional/social brain areas.

These personal violation cases generally drew quick-reaction-time decisions: The action in question was judged wrong. These cases induced heightened brain activity in the emotion and social cognition areas (specifically, the medial prefrontal cortex, posterior cingulate/precuneus, and superior temporal sulcus/temporoparietal junction). Greene et al. (2009) suggest that these rapid intuitions are produced in response to actions that involve direct infliction of harm on others, actions that fit the template of "I directly harm you." But because throwing a switch does not directly harm the person who is eventually killed by the act of throwing that switch, it therefore does not trigger the negative signal associated with direct violations.

A famous case illustrates how reasoned decision-making processes sometimes are triggered into action and can override the more automatic intuitive responses. This story consists of a set of villagers, hiding from soldiers who will certainly kill all the villagers if they are discovered. A baby in the group starts to cry. Stopping the baby from crying will surely kill it, but if that is not done, the whole group will be discovered and killed. This horrific, complex case provoked the usual, rapid emotional responses, and the suggestion is that they were produced by one's repugnance at inflicting direct and lethal harm to a helpless baby. Probably

triggered by the inevitable calculation, however, that all would be killed, including the baby, if action were not taken, a second set of temporally slower brain processes occurred, taking place in areas associated with higher-order reasoning and decision conflict management processes.

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 nonoptionally. 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 social-cognitive and emotional brain areas and are not always triggered into action. This, I suggest, is the complex reasoning system. Furthermore, when this reasoning system is activated, its results are sometimes in conflict with the intuitions of the other system. The conflict is perhaps resolved by some assessment of the competing strengths of the two sets of signals. So 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 possible but that, I suggest, is not always, or even often, produced for the punitive judgments described in this review.

PUNISHMENT JUDGMENTS ARE DRIVEN BY JUST DESERTS CONSIDERATIONS

Just Deserts versus Utilitarian Considerations

Psychologists do not claim that these intuitive judgments are made at random. Thus, experiments that vary the input information are necessary to determine what sorts of information are used in the punishment decisions. Asserting that these judgments are intuitions does not mean they are not based on some properties of the scenarios; it simply generates certain constraints on how the research is to be done to determine which properties are used in the decisions. The considerations that lead to intuitions—the information consulted by the mind—are not recoverable by the decision maker. Nisbett & Wilson (1977) first demonstrated this in a series of famous experiments showing that people had only limited insight into the considerations governing their choice behavior. When asked why they made a particular choice, participants tended to formulate an answer spontaneously and after the fact according to what they thought the stereotypical, sensible considerations would have been, rather than engaging actual memories that accessed and recalled the actual considerations. Nisbett & Wilson concluded that although people often do have judgment processes that follow clear rules and do look to specified kinds of input information, they only have limited insight into those processes. Intuitive decisions are often of this inaccessible decision-making sort, and so we cannot rely on what people report about the information they process to make these kinds of decisions.

Psychological researchers' standard practice here is to take what is somewhat grandly called a policy-capturing approach (Cooksey 1996). Essentially, the researcher presents a scenario describing a crime and elicits a severity of punishment judgment from the respondent. A simple version of a study would have two groups of respondents assign punishment to a story varying in information relevant to just deserts. If those who received information that from a just deserts perspective would lead to a more severe sentence actually assigned a more severe sentence, then we infer that just deserts information is a driver of sentence severity judgments. A more complex design would orthogonally vary just deserts and, say, incapacitation information and test whether either or both sources of information made a difference in the assigned severity of a sentence.

To make this more concrete, consider an example. The just deserts stance assigns sentences according to the moral culpability of the perpetrator. Thus, if the vignette is adjusted such that the moral culpability is either high or low,

the retributivist will adjust his or her punishment accordingly. By contrast, a person largely concerned with removing dangerous criminals from circulation would be relatively unmoved by this variable and far more sensitive to the frequency with which this criminal had committed crimes in the past or was characterologically likely to do so in the future. Across samples of respondents, variations in information are included in vignettes; patterns of response emerge, and, thus, we can infer which sentencing perspective drives people's actual sentencing decisions.

Several studies have been conducted using this paradigm, and the results indicate that respondents generally sentence based on a just deserts perspective. In one study (Darley et al. 2000), the just deserts motive was compared to the incapacitation motive, and the just deserts manipulation was the major determinant of the sentence assigned. Manipulations in magnitude of the just deserts punishment both directly increased the final sentence assigned and indirectly increased the sentence assigned by increasing the moral outrage that the subject felt at the criminal's act. If the actor had committed a similar crime before, then the subject perceived that the actor was more likely to commit similar offenses in the future. Because of this, the subject perceived the committed offense as more serious and was more morally outraged by the offense committed, but there were no unmediated effects of the likelihood of recidivism on the severity of the prison sentence assigned.

A second line of evidence from this study also suggests that respondents' sentences are derived from a just deserts stance. Subjects originally assigned sentences to a series of crimes that they thought were the appropriate sentences. They then were given descriptions of how a person "who was sentencing from an incapacitive stance, or from a just deserts stance, would sentence." When subjects assumed the just deserts stance, their pattern of sentencing closely fit the sentences they gave before being instructed. This suggests that their ordinary sentence-generating system was a just deserts one. The pattern generated when they were working from an incapacitative perspective was quite different from their unprompted sentence pattern.

In a second study (Carlsmith et al. 2002), the just deserts motive was tested against the motive to achieve general deterrence. Again, the sentences assigned were driven by the just deserts stance rather than the deterrence stance. In fact, the deterrence-relevant information had no effect, direct or mediated, on sentence durations. One possible reason is that it may strike the respondents as unfair to increase a sentence on a single offender who has already offended to achieve deterrence at a societal level.

Harm or Wrong: What Is Computed in Just Deserts Reactions?

What aspects of the actor's transgressive intentions, actions, and action outcomes are the observers reacting to in determining their just deserts responses? Often, those writing about people's perceptions of crime seriousness do not specify what they take to be the psychological determinants of seriousness. The two standard considerations are the harm done by the crime and the moral culpability of the criminal. For instance, Wasik & von Hirsch (1990) suggest that both contenders are involved, without precisely specifying the relationship: "The sentence shall be in just proportion to the seriousness of the criminal conduct: that is, to the conduct's harmfulness or potential harmfulness and to the offender's degree of culpability in committing the conduct" [p. 510, Section 1(1)].

In reality, however, the conduct's harmfulness can have little to do with its moral culpability. But the standard way of creating cases for respondents to rate causes these two variables to be linked together, which causes real difficulties in determining which is driving the respondents' ratings of crime seriousness. The problem is that if one samples actually occurring crimes, then the sample will mainly consist of crimes in which the harm caused by the crime, the murder, the rape, or the robbery was within

the range of harm that the criminal intended to bring about. Generally, the greater the harm intended, the more consequential the norm violated by committing that harm. When a person hits another, a norm is violated, but when one kills another, the norm violation is much more severe. In the standard study, successful crimes are studied (Rosenmerkel 2001, Warr 1989); harm and culpability almost perfectly covary, and disentangling which is driving respondents' ratings of crime seriousness is hard to do. In one study (Alter et al. 2007), however, "completed attempt" cases were used to disentangle harm from culpable conduct and thus demonstrated that it was the culpable moral wrongness of the conduct that drove the respondents' punishment severity assignments. The logic can be illustrated by a pair of the scenario cases. In one, a man deliberately shoots and kills another man, committing an intentional homicide. Now consider a second case in which the would-be murderer shoots, with identical intent to kill and good aim, but by an improbable coincidence—much loved by moral philosophers—a passing bird is hit by the bullet instead. This is a completed attempt case, but the intended target escapes unharmed. The shooter in both instances acts with the same degree of moral wrongfulness, but he inflicted harm in only the first case.

As the reader has probably intuited, the respondent inflicts very high punishment on both actors because the moral culpability is about equally high in both cases, although there is no harm in the completed attempt case. Consequently, it is moral culpability rather than harm that is mainly driving the punishment judgments. However, the occurrence or nonoccurrence of the harm matters. In the completed attempt case, the actor generally gets a punishment that is slightly reduced from that assigned to the successfully completed crime.

If we add another scenario to this pair, an interesting contrast is revealed. Some person, having taken all the proper precautions, shoots a gun at an artificial target, and via another improbable coincidence, some fool gets between the shooter and the target. Even though the fool dies, no culpability accrues to the shooter. This is roughly what Hart (1948–1949) meant when he commented that murder was a defeasible concept. The accusation of murder can be rendered void by the demonstration that the intent to murder was not present.

A recent neural imaging study (Young et al. 2007) adds to this claim. Respondents were brain imaged while reading that an actor either (*a*) intentionally puts a toxin in her friend's coffee, (*b*) intentionally puts what she thinks is a toxic substance (but is actually sugar) in her friend's coffee, or (*c*) unintentionally puts what she thinks is sugar but is actually a toxin in her friend's coffee. If the above argument is correct, people will wish to punish both the person who intentionally puts the toxin in the coffee (case a) and the person who intentionally tries to put the toxin in the coffee but, unbeknownst to her, uses sugar instead (case b). People will not punish the actor who draws from the container marked sugar, accidentally accesses the toxin, and tragically kills her friend (case c). These are the results that the researchers obtained: The first two actions were judged punishable, while the last, although it had bad consequences, was judged to be not punishable. The imaging evidence demonstrates activation of brain areas that process what cognitive developmental researchers call theory of mind information. Sorting out blame in this situation requires information not only as to what the actor is doing, but also as to what she believes she is doing. This enables the judgment that the actor deserved blame for administering what she thought was toxin to her friend and deserved no blame for administering what she thought was sugar.

To summarize, people react to transgressions with feelings of moral outrage—based on their intuitive assessments of the degree of the transgressor's moral wrongness—that motivate their assignments of punishment. Is it then natural to wonder if there are ways that we can see what actions follow from these cognitions about punishment? Will people actually administer punishments that mirror their desires to punish? It turns out that there is an experimental arena that begins to answer these questions.

PUNISHMENT IN THE WORLD: EXPERIMENTAL GAMES

Recently, experimental psychologists and behavioral economists have been running what are called experimental games. For our purposes, the key aspects of these games are that the participants can win—or sometimes lose—significant amounts of money, and the structure of the games can create chances for participants to take actions that the other participants will experience as transgressions. The games are arranged so that, at least sometimes, other participants are able to inflict punishments on those who transgress. Sometimes the focus of the game lies in seeing the patterns of transgressions and punishments that naturally arise. Generally, in those games, all the participants are actual subjects. In other games, the interest lies in seeing the frequency and magnitude of the punishments inflicted on a transgressor. Occasionally, then, the situation is rigged by having one of the apparent participants be a confederate of the researcher, who transgresses at some strategic moment.

The Anatomy of Punishment Games

Transgressions are generally self-serving exploitations of game structure in ways that benefit self and unfairly disadvantage others. The familiar example is a defecting choice in a prisoner's dilemma game. Other transgressions include profiting from, rather than reciprocating, generous gestures, free riding on others, or taking more than one's share of a common resource. The latter is Hardin's (1968) paradigm of the tragedy of the commons. Ingenious researchers have arranged to brain image players in the game, creating the possibility of linking patterns of brain activation to events occurring in the game. For instance, they have imaged people in the process of deciding whether to punish another's transgression.

Generally, the players' identities are kept secret from one another, and often they play only one trial against a specific other player. This removal of the "shadow of the future" by arranging that there will be no future interaction with one's current game trial partner is important because it defeats a theory of rational choice explanation for a player retaliating if transgressed against to earn the potentially valuable reputation of being a person "whom one better not shortchange in the future."

Originally, these games were played for low stakes, points with no value, points that with a low probability might turn into money, and so on. This left their results open to the criticism that game researchers were studying inconsequential decisions in which actions like retaliation were almost costless. More recently, these games have been played for major monetary stakes, and the willingness of game players to expend significant sums of money to punish if they feel unfairly treated tends to persist.

Transgressions and Punishments in Small Groups

In a number of transgression games, experimenters have demonstrated that if a transgressor, apparently deliberately, violates social norms, the victim of the transgression will punish the transgressor and will do so even if accomplishing this requires the victim to expend resources to do so. As an example, Sanfey et al. (2003) imaged subjects who were the responder in an ultimatum game. In that two-person game, the experimenter gives a sum of money to a proposer, who proposes a split of the money between himself and the responder. The respondent accepts or rejects the proposal. If it is accepted, the money is split, the responder gets what the proposer offered him, and the proposer keeps the rest. If rejected, both get nothing.

Past studies (Sanfey et al. 2003) have demonstrated several interesting findings. The modal offer from the proposer is often 50%. The responder is likely to accept an offer above about 25% to 30% of the total and is likely to reject an offer below that. Unfair offers differentially activated the bilateral anterior insula, dorsolateral prefrontal cortex, and anterior cingulated cortex. Interestingly, these brain areas are less

activated, and offers are less likely to be rejected if the responders perceived the unfair offer to come not from a person but from a computer program that randomly produces various offers.

One of the regions activated, the bilateral anterior insula, is an area implicated in studies of emotion, particularly involved in the evaluation and representation of specific negative emotional states (Sanfey et al. 2003). To sum up, offers perceived as unfair in the ultimatum game are reacted to with negative emotions and with frequent rejections of the unfair offers.

Strikingly similar results emerge from research on what are called trustee games (de Quervain et al. 2004). Trustee games have the following dynamic: The experimenter gives both A and B, for instance, ten units. A decides whether to send B no units or 10 units. If A sends 10 units, the experimenter quadruples the amount sent. B now has the original 10 units and 40 from the quadrupling, resulting in a total of 50 units. In the other scenario, A sends no units, in which case there is nothing for the experimenter to quadruple. Note that the sending of 10 units by A signals to B that A is willing to completely trust B in order to create the possibility of a maximal joint gain for the two players. The sending of zero units by A signals to B that A is completely unwilling to trust B in order to increase the sum they might share.

Assume A sends 10 units. B now has a two options structured by the experimenter: send back 25 units, in which case A's initial trust in B is reciprocated and B has proved to be trustworthy, or send back nothing, committing what A will experience as a serious violation of the social norms prevailing in the situation.

At the end of the trial, A has the opportunity to punish B by fining him a variable number of units: In one condition, A must pay a price in units to administer the fine; in the second condition, he can fine without paying a price; in a third condition, a symbolic punishment can be delivered that extracts no fine from B; and in a final condition, B is understood to be a computer that randomly returns 25 or zero units to A.

When the experiment begins, each of 15 A actors will play seven trials with what they perceive to be different B actors. The results indicate that all but one of the 15 A subjects trusted B and transferred 10 units to B; it is the behavior of these subjects that we examine (the experimenters actually had to rig the play of B—actual players would generally act in a trustworthy way, and the interest here is in norm-violating behavior). In three of the seven trials, the B player proved trustworthy (A thus earned 75 units during the game; this stash provided the funds that the subject could keep after the experiment, but A could also use them to pay to administer punishment to B). In four trials, the four B players were untrustworthy. In one of these, B was identified as a computer playing randomly. A did not punish the computer. In the remaining three trials, A thought B was a person who therefore had seriously violated the social norms of the situation. In these three trials, it was the cost and kind of punishment that A could administer that varied. When punishment could be administered without cost, the average punishment was more than 35 units. When it cost A one unit to administer two units of punishment, the average punishment was about 23 units—notice this is rather close to the 25 units that B, if trustworthy, owed to A. A also punished B even when it was possible to do so only by assigning symbolic points that cost B nothing.

Brain activation patterns in the caudate nucleus, an area associated with registration of reward information, revealed heightened activity while A was in the process of actually punishing B, but not during the symbolic punishment: "Taken together, our findings suggest a prominent role of [the] caudate nucleus, with possible contributions of the thalamus, in processing rewards associated with the satisfaction of the desire to punish the intentional abuse of trust" (de Quervain et al. 2004, p. 1256). This association of the caudate nucleus with satisfaction in administering actual punishment is supported by the finding that the stronger the caudate activity of the brain, the more units the subject expended on buying punishment to

inflict on the trust abuser. 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.

The Cross-Cultural Universality of the Punishment Response

Most of these trust games have been staged in Western, capitalist cultures, and it is plausible that costly punishment patterns found in such cultures are not necessarily characteristic of all cultures. In one remarkable study, however, experimenters administered the ultimatum game in 15 different cultures, chosen to represent a wide variety of human production systems (Henrich et al. 2006). The researchers report that all cultures showed the same general pattern of increased likelihood of responder rejection of offers as proposer offers grew more one-sided. Against this general picture, however, a quite marked cultural variation occurred: Some cultures showed a high rate of punishment that appeared rapidly when the proposer crossed the 50-50 split line, whereas other cultures showed much lower rates of punishment and the responder rejected the offer only after a split line was reached that was considerably more disadvantageous to the responder, such as 70-30 or 80-20 division.

By taking another behavioral measure—a measure of the level of altruistic sharing that respondents in each culture displayed—the researchers were able to predict another measure of the strength of the social norm against trust violations in cultures. The behavioral measure was taken in a dictator game, which is a variant of the ultimatum game, and was the percentage of the money the decider gave to the respondent. Here, both understood that the respondent simply got what was offered and could not punish the decider by blowing up the game. Thus, deciding to share higher fractions of the joint sum were altruistic actions. The researchers found that the cultures in which a higher degree of altruism was displayed, leading to the inference that sharing was normative in those cultures, were also the cultures in which not reciprocating trust was punished most heavily by an observer in a trust game.

Third-Party Punishments

"Punishment by an observer" needs explaining. Trust games can be expanded to include more than two individuals, and this is often done to study what are called third-party punishment propensities. In these studies, the third party, C, is given a role in which C observes A and B playing some kind of trust game and then has the power to administer a punishment to any of the first two players who violate social norms.

Substantial proportions of these third parties who witness the transgression but are not the victim of it actually administer punishment. Furthermore, third parties will punish even when the game is arranged so that they will never play against the transgressor again and they are assured that all parties in the experiment will be kept in a state of anonymity from each other (Fehr & Fischbacher 2004). The last two conditions are important in that they remove any possible rational reason to administer the punishment (from the self-interested perspective of rational choice theory), such as paying a small cost to demonstrate to others that you are not to be transgressed against in the future. In addition, third parties are willing to pay some costs to buy punishment fines to impose on those who violate norms mandating cooperation and trustworthy behaviors. For instance, in one study (Kahneman et al. 1986), the researchers ran a dictator game, which resembles the ultimatum game, except the responder (now the receiver) had no choice but to take the split offered by the proposer (now the decider). 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 dictator who inflicted that unfair outcome a fine of $5.00 by paying a cost of $1.00. In such cases, 74% of third-party

witnesses chose to do so. Some researchers have therefore labeled this "altruistic" punishment, a correct label from a purely economic perspective but perhaps not from a psychological perspective. These third parties are willing to pay to experience the reward of punishing the transgressor.

Think about the effect of third-party punishments of transgressions in the context of a small group situation, in which several observers of transgressions would be present—the punishment that a single third-party observer will inflict may be less than the punishment that the injured party will inflict. The combinatorial logic of this is complicated because, if the multiple observers were jointly aware of the presence of others who were available to punish, each observer is likely to reduce the amount of punishment they inflict, particularly if the punishment is costly to the observer. Still, if many punish, its magnitude can be high.

Punishment Can Control Normative Deviance in Actual Groups

Here, one may speculate on the human propensity to punish as an effective mechanism for the control of transgressions in at least small, face-to-face groups. But it turns out that we need not entirely rely on speculation. Remarkably, one imaginative trust and sanctioning game experiment (Gürerk et al. 2006) has actually created small groups in which punishment of transgressions is sometimes allowed and differences in outcomes observed. Twelve anonymous participants were run at once. To begin, each had a choice of joining one of two groups that were to play a 30-round public good game. In this public good game, each round consists of each subject being endowed with 20 units (units are cashed in for money at the end of the game) and all simultaneously contributing as much or as little of their endowment to the pubic good pool as they wish. The total contributed to the pool (multiplied by a factor of 1.6) is then shared out equally among all, regardless of the size of their contribution to the common pool. Given this, those who contribute above average amounts to the pool are taken advantage of by those who free ride by contributing little or nothing.

One difference between the two groups exists: In one, it is possible to punish free riders, and in the other, it is not. All subjects learn about this difference at the outset and, based on this information, initially choose a group. In the group with punishment, the punishment mechanism is the now familiar fine of three units, which costs the punisher one unit to administer. Therefore, the punishment is costly. In the group with punishment, after the contributions to the common pool are made, all subjects see a display of the total contributed and are given the amount that is the common distribution that each individual will receive. At this point, each subject is given 20 more units and can choose to keep the units or buy tokens with which to inflict a punishment on any other members of the group. Each punishment token costs one unit and punishes the target three units. So punishments of up to 60 units can be inflicted by one person onto others. In this group, subjects now inflict punishments on others.

In the no-punishment group, all see the display of what each did. To match the treatment given to the punishment group, all subjects are given 20 tokens, but these are simply kept by the subjects.

Another complexity is added and appears central to the eventual outcomes of the study. After a trial, subjects from each group see not only their results but also the results of the other group being run concurrently, and each person can migrate from one group to the other.

Recall that at the beginning of the experiment, participants receive a description of the rules and procedures of both societies. One group is described as the society in which each individual has no "influence on the earnings of the other persons." The other society is described as one in which members have an influence on the earnings of the other group members "by assigning positive and negative tokens" (no more will be said about the positive tokens, since relatively little was made of the possibilities that they create). Participants signal which

group they will join and learn about the migration possibilities.

Happily, in each set of the 7 sets of 12 subjects, enough subjects opted to join each society so that the experiment could launch. About 65% chose the punishment-free society. The fact that a majority of subjects chose the nonpunishment society may not be broadly generalizable. The subjects had just been given a great deal of information and probably had not been able to intuit the full ramifications of it. A group in which "others could not influence earnings" may have sounded better to many than one in which earnings could be influenced by others.

Trials began, and rather quickly low rates of donation to the common pool began to emerge in the punishment-free group. People from that group also began to migrate to the group with punishment. To some extent, this must have been because, observing the punishment society's displayed results, the punishment-free group participants could see that the average size of the contributions to that common pool had started high and climbed higher, moving to about 90% by the tenth trial and stabilizing at that level for the rest of the game.

The engine that drives the difference between the two groups is the availability of punishment in the one society, coupled with the actual frequent use of punishment administered to low contributors by those who contribute highly to the common pool. The researchers suggest that a norm to punish developed in this society: About 63% of those in the (growing) group participated in punishing low contributors, and the high-punishing group included some who had been free riders when they had been in the society in which punishment was not possible. Of those who themselves were high contributors to the common pool, almost three-fourths administered punishment tokens to discipline low contributors. This meant that, during the first few trials, the total points extracted by the high contributors were low, given the costs of the punishments they were expending on those who were tending toward low contribution rates. But as trials progressed, the points extracted per trial by those who were active punishers of free riding approached 52, the theoretical maximum possible.

In the punishment-free society, deterioration continued. The rate of contributions to the common pool deteriorated, by the eighth trial oscillating around 10% of the maximum. In the last few trials, no participant contributed anything to the common pool. Each individual simply maximized his individual take, with nothing gained from the possibilities of growth provided by cooperating to contribute to the common pool. It is interesting to speculate why anybody stayed in this group; perhaps they stayed because they gave very few of their 20 initial units to the common pool and received 20 units later in the trial, thus accumulating around 40 units per trial without switching groups.

This study demonstrates that in certain circumstances the existence of a punishment mechanism allows group members to sanction transgressions in ways that effectively reduce the rate of transgressions so that the individuals in that group achieve high rates of mutual profit. Many researchers are amazed that a study so complex in procedure and necessity of displays, and so demanding on the subjects' mental processing, could successfully be brought off. It is not a criticism, therefore, to point out that we are not quite sure what the certain circumstances were that led to the triumph of the group with the punishment mechanism. Is the effect robust over alterations in punishment cost parameters, group sizes, and other specifics of the experimental detail? Most groups are aware of the dynamics and outcomes in the group within which they exist, but are much less aware of the dynamics and outcomes of other groups. To what extent did the results of this study depend on the constant possibility of migration to the other group and on the clear, transparent window into the inner life of the other group? All these questions and others require research attention. Still, the experiment is a remarkable achievement. It demonstrates that in a group that has the possibility to punish norm-transgressors, enough participants will expend resources to punish those transgressors so that the group members

extract a reasonable share of the resources that are available to them. Second, participants will migrate into this group, exiting a group in which the absence of the possibility of punishing transgressors is producing a low rate of resource extraction.

THE PSYCHOLOGIC OF PUNISHMENT: A SUMMARY

This review has suggested that when a person personally experiences a transgression or becomes aware of a transgression committed against another person, he or she has a reaction of moral outrage. That reaction is of course a result of cognitive processing, but it also brings with it considerable emotional energy. It is experienced as what Heider (1958, p. 219) calls an "ought force." It is not that I want or demand that the transgressor be punished, although that is true; rather, a "suprapersonal objective order requires the punishment." Generally, this reaction is produced by rapid intuitions that pop into the mind rather than resulting from complex conscious reasoning. Like most intuitions, it can be overridden by conscious reasoning, but Carlsmith (2006) has demonstrated that even when people are caused to reason about appropriate punishments, they often come to results congruent with their intuitions. Punishment decisions, therefore, are what psychologists call dual process decisions in that two or more different processes, sometimes alternately engaged, can make them.

Intuitions process information, and the information processed when punishments are being determined, I suggest, is information about what the transgressor justly deserves for the offense committed. Just deserts sentiments do fit with the characterization of the "ought force" of the impulse to punish.

Functional magnetic resonance imaging (fMRI) studies have contributed a good deal to our understanding of punishment responses. First, they support the notion of dual processing of decisions and give some information on the different brain areas involved in the alternate processes. Researchers have suggested that the initial rapid processing is put into action by a template that reacts to a direct, personal violation of another, such as is typified by a direct infliction of a physical harm. This rapid processing is the likely producer of the moral outrage reaction. Second, imaging work has demonstrated that administering punishment to the transgressor is a rewarding act that activates reward centers in the brain and that this is true whether the victim of the transgression administers the punishment or instead a third-party witness of the transgression administers it.

Trust games prove extremely useful in studying transgressions and punishments. From them one learns the shape and content of social norms involving cooperation, particularly implicit understandings of, for instance, how one person, taking a chance and acting for the common good, creates obligations for others to reciprocate. Because we have come to understand these norms of cooperation and trust, the experimental trust games create the platform for a number of the imaging studies of the brain processing of transgressions. Furthermore, as implied above, trust games with multiple players in multiple roles have demonstrated that not only victims of transgressions, but also third-party witnesses will often take the opportunity to punish transgressors.

Trust games also let us create brief worlds within which small groups briefly live and in which we can observe some of the simpler patterns of social interactions that emerge when resources are being produced and later shared among the producers of the resources, and what happens when those who have shirked the task of producing resources then claim the resources produced by group effort. As one notices, all the characteristics of punishment cognition that these empirical studies have discovered link together to produce mechanisms that are fairly efficient for the control of norm violations via the infliction of punishments in at least small groups, the members of which can reasonably monitor each other's behavior and detect at least conspicuous norm violations. Normative transgressions by individuals who ought to be

bound by those norms are punished, which can mean that the frequency of transgressions is reduced. The fact that people are moved to punish when others, rather than themselves, are the direct victims of transgressions considerably increases the efficiency of punishment as a force to reduce norm violation for several reasons. First, whether or not the direct victim of the transgression is in a position to punish, punishment is likely to occur. Second, if punishment is inflicted by several group members, the total magnitude of the punishment is likely to be high. Third, if the group members become aware of the others punishing, the punishment process is mutually reinforcing. Fourth, people are likely to join in the condemnation, which will convince others, including the transgressor, that the act in question was an actual violation of a norm actually held by the group.

Learning Punishment Norms: Evolved Predispositions or Current Learning?

There are several candidates for explaining why this rather integrated patterning of punishment cognitions and practices is characteristic of a large number of societies—as shown in particular by the experimental game research. First, it is possible, perhaps likely, that there is an evolutionary component to the emergence of these punishment cognitions and practices; specifically, there may be an evolved predisposition toward their emergence, such that they are learned rapidly from early experiences as children. Some have suggested that there is a universal moral grammar similar to the Chomsky-postulated universal language grammar (Mikhail 2007).

According to this argument, these norms and the cognitions that embed them are generally rules for the sharing of resources that are the products of individual and group effort. The metaphor: Humankind had a long existence in small, hunter-gatherer bands whose members' lives depended on reciprocal altruism and the observance of rules for sharing common goods.

Other scientists—while probably not denying the possibilities of an evolved propensity for learning these sorts of rule sets but perhaps more struck by the differences in the rule sets prevalent in different cultures—might place more emphasis on the learning mechanisms by which these norms are transmitted to present-day children in present-day norm-learning environments. These researchers are inclined to carry out careful observational studies of play groups, daycare centers, and the early grades of primary schools to see how children are taught these rules (Walton & Sedlak 1982, Much & Shweder 1978). A recent discussion of the comparative utilities of these two explanatory possibilities, advocating the evolutionary account, can be found in Robinson et al. (2007).

IMPLICATIONS: CRIMINAL JUSTICE SYSTEMS IN COMPLEX SOCIETIES

Lastly, what implications, if any, does this array of discoveries about the psychology of punishment have for legal system policies in various societies and, more specifically, for modern societies with complex capitalist economies? If one were to allow the considerations arising from these studies to affect the criminal justice system, what might we do? In discussing this, I shall occasionally present evidence that bears on the feasibility of the suggestions.

Obviously, the invited conclusion is that the criminal codes of societies should be broadly in accordance with the moral intuitions of the governed community. This is by no means a new suggestion. Indeed, in earlier centuries, this was a guiding principle that was taken for granted by many common law scholars. But the fact that it was taken for granted is no argument that it is correct.

What are the arguments for community norms being represented in criminal codes? Basically, that citizens will voluntarily obey laws that they perceive as moral ones. In *Why People Obey the Law*, Tyler (1990, p. 37) reviewed five studies that related respondents' perceptions of the morality of laws to their reported propensity to obey the laws and found an average correlation of 0.45. In Tyler's own survey, the first

order correlation between the belief that "law breaking is immoral" and reported compliance with the law was 0.42 (Tyler 1990, table 5.1, p. 59). In other words, people who think that the legal codes have moral content report that they are more likely to obey the laws.

These of course are correlational studies, but there is some observational evidence for the belief that the perceptions that laws represent correct moral behavior are what cause people to obey laws. Occasionally, legislatures have run the experiment in the negative, enacting codes that violate the moral codes of substantial segments of the community and punishing violations of those imposed codes. The imposition of prohibition laws in the United States criminalized the distribution and consumption of alcoholic products, thus criminalizing action patterns that large segments of the society thought were morally permitted ones. Alcohol consumption continued, in underground ways, and some analysts suggest that the Prohibition Era severed the connection that many citizens had formed between illegal actions and morally wrong actions, to the eventual detriment of their law abidingness in general. Similar interpretations have been made of the resistance of the colonized to the imposition of the legal codes imported from the codes of the colonizing and conquering powers. As Benton (2002, p. 27) remarks: "The British, for example, were notorious failures at making sense of the political structure of Iboland in eastern Nigeria. Consecutively disruptive policies were answered by continual revolt."

Revolt continues to be a response to perceived unjust workings of the criminal justice system. Following the acquittal of white police officers accused and tried for beating African American Rodney King, a beating that had been video recorded by a spectator and was widely shown on television, major protests erupted in Los Angeles (Cannon 1999) by rioters who thought the verdict was unjust.

Several psychologists (for instance, Nadler 2005) have attempted to demonstrate experimentally that if study participants discover that specific elements of legal codes lead to trial outcomes that violate the participants' moral codes, such a discovery causes those participants to be less willing to obey the law in the future. Usually, one group of study participants is exposed to a description of a transgression in which the punitive outcome seems unjust, and another group reads or sees the same transgression, but the punitive sentence set by the court seems just. Participants in the unjust condition report less willingness to obey laws in general and to, for instance, report criminal activities to the police.

The effects in these studies are small ones, but that is what one would expect. Given that the respondents had probably experienced many reports of crimes in which they thought court decisions had gotten the outcomes at least close to morally right, it was unlikely that their accrued confidence in court decisions was greatly lessened when the respondents learned only that there had been one occasion in which a court had brought about a violation of a moral intuition about what was a just outcome, even though the violation was a shocking one. What is possible to imagine, however, is the growing disenchantment that citizens would develop if they were continually confronted with case after case in which the justice system inflicted verdicts and sentences that they perceived as unjust.

The Empirically Informed Deserts Sentencing Proposal

The research reviewed here demonstrates that individuals have powerful intuitions about the just punishments for various offenses, and these intuitions are substantially driven by just deserts considerations. Some rather preliminary evidence suggested that when citizens perceive that the law assigns punishments that conflict with their intuitions, they lose respect for the law. Would it be possible to construct a sentencing system that to some extent incorporates these intuitions? Whether this is desirable requires consideration; however, whether it is even possible depends on the degree to which these punishment intuitions are shared among

people within jurisdictions or societies. Does this sort of societal consensus exist?

Frequently those advocating sentencing systems based on other criteria assert that no level of consistency that would allow an empirical deserts system even to be considered actually exists. For instance, van den Haag (1987, p. 1254) argues:

> von Hirsch (a proportionate deserts theorist) appears to believe that the comparative seriousness of crimes can be determined in all cases. Not so. Comparative seriousness can be determined only for some crimes, and it does not fully determine the comparative punishment deserved. If rape is a crime and murder is a crime, rape-murder must be more serious than either. Does rape-murder deserve the sum of the punishments meted out for rape and for murder. More? Less? Even when the crimes are nearly homogeneous, assigning seriousness is arbitrary. Is rape more serious than assault with a deadly weapon. Is burglary more serious than fraud, when fraud does more harm? What about mishandling toxic waste? Ordinal determinations of seriousness become altogether arbitrary when the series of heterogeneous crimes must be compared.

A number of other theorists have made similar impossibility claims about empirically informed deserts sentencing systems, but they may not be clear about what proportional deserts proponents are proposing (for an extended discussion of these claims and proposals, see Robinson & Kurzban 2007). Briefly, the suggestion is that people will see that a psychologically acceptable proportionate desert punishment is achieved if they examine the sentences assigned to various crimes by the justice system and find the rank ordering of sentences to be in accord with their own rank orderings and the magnitudes of the sentences within a latitude of acceptance from their own sentences. That does not mean that the absolute magnitude of the sentences has to match exactly. Some people assign the death penalty to violent murders, while others assign it a life sentence with no possibility of parole. Many who do either will understand that the death penalty is a contested topic and find the two alternatives to be acceptable substitutes for each other. The sentence they do not choose is still within their latitude of acceptance. But, and this certainly should be acknowledged, some who assign a death penalty will not judge that a life sentence is a morally acceptable substitute. They will see it as immorally lenient. Others find a long duration prison sentence given for possession of minor amounts of marijuana to be an immorally severe sentence. So the empirical claim here is that citizens have a shared ranking of crime severity and have some general agreement on ranges of sentences that should be assigned to different crimes, but there are some disagreements on acceptable sentence ranges.

Those who think that an empirical deserts theory built around citizen consensus is impossible point out that sometimes a case will arise that will invert the standard ranking of crimes. Standardly, murder is worse than rape, but a horribly violent, sadistic rape that leaves the victim alive is worse than a mercy killing. In fact, this is what van den Haag suggested defeated the possibility of an empirical deserts sentencing system. It is certainly true that these inversions exist, but no sentencing system would require that all crimes in a category receive the same sentence. Instead, sentencing grids allow sentencing systems the leeway to adjust the magnitude of the sentence. And the adjustments are based on factors that, on examination, seem to adjust upward or downward from the standard according to the moral gravity of the specific crime committed; adjustments depend on the presence of aggravating or mitigating elements in the crime commission. Previously, we suggested that citizens have ranges of punishment magnitudes for various crimes, which create what might be called their acceptable punishment ranges. To take this analysis one step further, it is probably the case that when a citizen is asked about an acceptable range of punishments to assign, for instance, to murder, they will report a range of punishments all of which are severe. If the case turns out to be one in

which a relative kills a terminal patient with intractable pain who is asking to be killed, then the punishment is likely to be much lower than the punishment range the citizen would assign to murder. People seem to think about crimes in terms of exemplars or prototypes, whereas legal codes tend to use necessary and sufficient condition definitions. We need to recognize that people recruit variant prototypes when they discover that a killing was a mercy killing or that a rape was a statutory rape involving two consenting partners, one of whom was slightly below the age of legal consent.

Returning to the question of whether there is a sufficient consensus among citizens about crime penalties, a good many empirical studies have examined the agreement between citizens on the rank ordering of the severity of different specific instances of crimes, and the correlations found have been high (see Robinson & Kurzban 2007, pp. 1854–61 for a review of these studies). Robinson & Kurzban's (2007) own study is methodologically sophisticated and demonstrates a really remarkable consensus on rankings of 24 short scenarios describing crimes chosen to range in seriousness. These scenarios range from rape and murder through theft and burglary, to more minor crimes.

Kendall's coefficient of concordance is a statistic that measures the concordance among ratings contributed by an entire set of subjects, and it is scaled to range between 1 and 0, with 1 indicating perfect rank order agreement from all subjects. In the Robinson & Kurzban (2007) study, in which the experimenters led the subjects through a thoughtful set of procedures to allow them carefully to consider their judgments, the coefficient of correspondence was an astounding 0.95. In a second study, the experimenters used the same stimuli but reached out to a more demographically varied, larger subject population, using a Web survey format administered by a computer program. Web surveys often find some degree of inattentive responses by respondents, and this survey found some, but still the concordance was remarkably high, dropping only to 0.88. Clearly there is a high degree of agreement on the rank ordering of the wrongness of crimes within the American culture. Therefore, a strong societal consensus on an empirically created ladder of relative sentence durations for crimes exists. So a necessary condition for creating an empirically influenced, deserts sentencing system is fulfilled.

There is, of course, no perfect agreement on the moral rightness or wrongness of all actions in our society, and if some contested actions were included in the survey above, this would be discovered. Perhaps the most extreme example is abortion, considered by some to be murder and by others to be an allowable exercise of the pregnant person's freedom to make consequential life choices. Two things must be said about this: First, it illustrates the point made above that laws perceived as unjust can generate contempt for the legal system. Here, specifically, those who have lost the criminalization battle—in this instance those who have failed in their efforts to criminalize abortion—are at risk of losing the intuition that the legal codes are a legitimate guide to moral behavior and thus should be obeyed. This actually has happened. Some prolife advocates, radicalized by engaging in a set of escalating protests, have come to regard the legal justice system as immoral and now are willing to kill doctors who perform abortions. If segments of the community are in this sort of absolute disagreement on whether specific conduct is moral or not moral, then reactions to the discovery of this can lead to increased disrespect for legal codes. Second, Tyler's (1990) groundbreaking work on procedural justice suggests another way that something like respect can be retained for justice systems with components that are not perfectly aligned with citizen sentiments. The task of the authorities is to use procedures that give an unbiased and respectful treatment to the disputants. Tyler's evidence (Tyler & Huo 2002) suggests that many people, treated in this procedurally just fashion, will retain some respect for the authority system, even when a specific decision is not one they perceive as the morally correct one. Napier & Tyler (2008, p. 510) summarize: "[R]esearch on the relational model of procedural justice has

shown that people are more likely to voluntarily accept outcomes they feel are justly arrived at and to be satisfied with the authorities and institutions using just procedures." There is a debate here. Skitka (2002) interprets data as suggesting that if people are strongly convinced of the rightness of their specific moral position, then the fairness of the procedures that decide against their position will have no influence on them. But a reanalysis of the data (Napier & Tyler 2008) finds evidence that procedural fairness still effects decision acceptance.

All this said, we could now formulate a proposal for an empirically influenced criminal justice system. Because criminal codes are implemented by the relevant legislative entities, which are usually the state legislatures, this is the group that will eventually enact them. The wise legislature might establish a commission to draft a criminal code. Our argument is that the commission should include social scientists, who will have done the empirical investigations necessary to discern the general contours of the shared community intuitions about what counts as criminal. Then, the commission should lean toward criminalizing those actions that the community thinks are major moral transgressions and lean away from criminalizing those actions that the community thinks are not morally prohibited. Further, the commission should lean toward assigning relative sentence magnitudes that are consistent with the moral intuitions of the community. (For an extended discussion of the utility of a deserts-based sentencing system, see Robinson & Darley 1995.)

Deontological deserts calls for punishment in proportion to the offender's moral culpability. Moral philosophers exploring this perspective also should have a voice in the code-shaping process, given that they are engaging in an attempt to understand the shared moral codes embedded in the traditions of the community and to examine how these codes are brought to bear on particular cases. By these processes, of course, they may discover incoherent or apparently contradictory sets of judgments in the community decision patterns and may be able to suggest resolutions of those contradictions. They also may discover that the community intuitions are in some important sense morally wrong. Both of these discoveries and, more generally, examining the contrasts between the community and the philosophical version of the code and thinking about how these differences might be resolved are a necessary exercise because, according to the arguments in this review, the community code is a deontological one as well, driven by just deserts intuitions.

There is another requirement that a social science perspective suggests should be imposed on the commission and the legislature that is enacting the criminal codes: the responsibility for working to persuade the community that the laws they enact—particularly laws that are importantly different from the community perceptions of what is and is not criminal—are the morally appropriate ones. This is a more stringent version of the rule-of-law requirement that the laws be promulgated, and it is not clear that the law enactors take on this responsibility (Darley et al. 2001). Citizens tend to generate their perceptions of what the laws are from what they think the laws should be based on their moral intuitions. If this leads them to erroneous assumptions, innocent disobedience can ensue.

IS THERE ANY VIRTUE IN VENGEANCE?

The just deserts ground for determining prison sentences is the one backward-facing justification for punishment. It is a deontological determination based on what the offender justly deserves for the moral weight of the transgression he or she inflicted. The proposition of the empirical deserts proposal is that the sentiments of the governed community are an appropriate basis on which to determine these just deserts punishments. But then, thinking back on the incredible tortures inflicted on transgressors in medieval times, on the public spectacles with which executions were brought about, and that hanging was thought the appropriate punishment for a hungry child stealing a loaf of bread and burning alive the appropriate punishment for heretical views, we are driven to

examine closely any policy based on retributive sentiments.

Here, the empirical deserts proponents have something, but not everything, to say. First, it is not clear that those punishments were imposed because the governed communities demanded them. They were often punishments imposed by authorities driven by Draconian deterrence theories using the logic of making any actions the rulers wished to discourage subject to punishments of such dreadfulness that no one would dare commit these actions. Second, for reasons most sensitively analyzed by Garland (1990), modern cultures have moved away from what we now consider the barbaric inflictions of earlier societies. Further, some societies (Doob & Webster 2003) that do use proportionate deserts punishments scale their range of punishments considerably below the magnitude of the harms they are punishing. Murderers receive sentences measured in years, but generally not life imprisonment. Many of these societies do not impose the death penalty. These lower scales of punishments seem accepted by the societies in question.

Many will be concerned by the grounding of morally appropriate sentences in what we might call "the conscience of the community." But what are the viable alternatives, given that sentencing practices are generally legislative enactments, and legislatures are likely to legislate with a keen eye to at least their interpretations of the community sentiments? And, as Simon (2007) points out, a pervasive fear of crime in our society tends to motivate legislatures to drive sentences higher and higher.

The Deterrence Sentencing Program

A deterrence theorist could generate sentence magnitudes driven by crime rates: If the rate of a particular offense is increasing in a jurisdiction, increase the sentence. However, several reviews of naturally occurring experiments that test the effectiveness of this deterrence-driven sentence-setting practice are not encouraging. Doob & Webster (2003, p. 143) comment that "a reasonable assessment of the research to date—with a particular focus on studies conducted in the last decade—is that sentence severity has no effect on the level of crime in society. It is time to accept the null hypothesis."

To clarify this, few doubt that inflicting punishments for crimes has a general deterrent effect. What is doubted is that tuning criminal sentences upward because of concerns for the prevalence of one type of crime or another will reduce the rate of the crime in question. Other researchers provide some explanation for this lack. The mechanism of this crime reduction effect is assumed to be the knowledge among crime-prone citizens of the increase in punishment magnitude. As to the plausibility of this assumption, and more generally to the notion that potential criminals make deterrence calculations, Anderson (2002, p. 295) has interviewed criminals in penitentiaries and reports the following:

> The results suggest that 76% of active criminals and 89% of the most violent criminals either perceive no risk of apprehension or have no thought about the likely punishments for their crimes. Still more criminals are undeterred by harsher punishments because drugs, psychosis, ego, revenge, or fight-or-flight impulses inhibit the desired responses to traditional prevention methods.

Rehabilitative and Restorative Concerns

Our analysis of the psychology of punishment suggests that citizens desire the criminal justice system to inflict retributive punishments on offenders, in proportion to the moral weight of the offenses committed. If this were citizens' sole desire, it would be a rather dismal outcome. This solely retributive justice system would give up any chances to achieve many of the more positive social purposes that historically it has been hoped that the criminal justice system could produce, at least with some offenders. Of particular concern here are the rehabilitative purposes such as educating prisoners,

benignly altering various attitudes, weaning prisoners off drugs and alcohol, and restoring them to some functioning mode of citizenship. Obviously, what we are striving for is a system of efficient and humane punishments in which efficient punishments not only avoid the negative consequences (i.e., prisonization) that are the normal effects of prison terms, but actually achieve some positive life consequences for those in the criminal justice system. Interestingly, there is evidence that although citizens generally do see punishment as a necessary infliction on those who intentionally violate society's moral rules, they are not only willing, but are eager that other goals be reached in the criminal justice system. A series of studies by Gromet (Gromet 2009; Gromet & Darley 2006, 2009) demonstrates that when the offense committed is relatively minor, respondents impose tasks on the offender that, for instance, have the offender exert time and effort on restoring the victim to a preharm condition, listen to the victim describe the sorrows and difficulties he or she experienced, and perform other remedies proposed by the restorative justice movement. They may feel that these processes inflict sufficient punishment to satisfy the just deserts concerns, or they may feel that for minor offenses the just deserts requirements for punishment can be set aside in the service of more important goals.

As the offense increases in moral magnitude, prison terms that the respondent perceives as appropriately proportionate to the offense are tentatively set, but restorative justice conferences are also recommended. If the offender successfully completes the restorative actions that the conference suggests, the respondents revise the tentative sentence downward, often to a significant degree. More generally, respondents are often willing to trade in considerable sentence duration that they thought was appropriate in order to access rehabilitative or restorative options.

On the one hand, this complicates the picture of the citizen as solely seeking a fixed measure of retributive sanctions to be inflicted on perpetrators of transgressions. On the other hand, it brings into the picture a note that most criminal justice scholars would find socially hopeful in that it creates a space for rehabilitative purposes in the criminal justice system. Research is needed to examine what the shape of this space is. Is it the case, for instance, that if the offender achieves initial restorative actions, then the retributive inflictions could be reduced to short duration activities? Could the retributional motive be satisfied by the assignment of prison sentence duration, leaving room within that duration for some sequences of reeducation or rehabilitative activities for the offender? All this requires empirical exploration.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENT

The author acknowledges the helpful readings of Adam Alter, Adam Moore, Leigh Nystrom, and Alison Newman of this material and the timely contributions of Mary Sigler on the program of the deontological philosophers. The contributions of the 2005–2006 Fellows of the Center for Advanced Study in the Behavioral Sciences are also gratefully acknowledged.

LITERATURE CITED

Alter AL, Kernochan J, Darley JM. 2007. Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. *Law Hum. Behav.* 31:319–35

Anderson D. 2002. The deterrence hypothesis and picking pockets at the pickpocket's hanging. *Am. Law Econ. Rev.* 4(2):295–313

Benton L. 2002. *Law and Colonial Cultures: Legal Regimes in World History, 1400–1900*. Cambridge, UK: Cambridge Univ. Press

Cannon L. 1999. *Official Negligence: How Rodney King and the Riots Changed Los Angeles and the LAPD*. Boulder, CO: Westview. Paperback ed.

Carlsmith KM. 2006. The roles of retribution and utility in determining punishment. *J. Exp. Soc. Psychol.* 4(4):437–51

Carlsmith KM, Darley JM, Robinson PH. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *J. Personal. Soc. Psychol.* 2(83):284–99

Cooksey RW. 1996. *Judgment Analysis: Theory, Methods, and Applications*. San Diego, CA: Academic

Darley JM, Carlsmith KM, Robinson PH. 2000. Incapacitation and just deserts as motives for punishment. *Law Hum. Behav.* 24:659–83

Darley JM, Carlsmith KM, Robinson PH. 2001. The ex ante function of the criminal law. *Law Soc. Rev.* 35:701–26

de Quervain DJF, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, et al. 2004. The neural basis of altruistic punishment. *Science* 305(5688):1254–58

Doob AN, Webster CM. 2003. Sentence severity and crime: accepting the null hypothesis. In *Crime and Justice: A Review of Research*, ed. M Tonry, 30:143–95. Chicago: Univ. Chicago Press

Fehr E, Fischbacher U. 2004. Third-party punishment and social norms. *Evol. Hum. Behav.* 25(2):63–87

Garland D. 1990. *Punishment and Modern Society: A Study in Social Theory*. Chicago: Univ. Chicago Press

Greene JD, Cushman F, Stewart L, Lowenberg K, Nystrom LE, Cohen JD. 2009. Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* 111(3):364–71

Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44:389–400

Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293:2105–8

Gromet DM. 2009. *Restoration and retribution: people's negotiation of multiple responses to wrongdoing*. PhD diss. Princeton Univ., Princeton, NJ

Gromet DM, Darley JM. 2006. Restoration and retribution: how including retributive components affects the acceptability of restorative justice procedures. *Soc. Justice Res.* 19:395–432

Gromet DM, Darley JM. 2009. Punishment and beyond: achieving justice through the satisfaction of multiple goals. *Law Soc. Rev.* 43:1–38

Gürerk O, Irlenbusch B, Rockenbach B. 2006. The competitive advantage of sanctioning institutions. *Science* 312:108–11

Haidt J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108:814–34

Hardin G. 1968. The tragedy of the commons. *Science* 162:1243–48

Hart HLA. 1948–1949. The ascription of responsibility and rights. *Proc. Aristot. Soc.* 49:171–94

Hauser M, Cushman F, Young L, Kang-Xing J, Mikhail J. 2007. A dissociation between moral judgments and justifications. *Mind Lang.* 22:1–21

Heider F. 1958. *The Psychology of Interpersonal Relations*. Hoboken, NJ: Wiley. 326 pp.

Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, et al. 2006. Costly punishment across human societies. *Science* 312(5781):1767–70

Kahneman D. 2003. A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58:697–720

Kahneman D, Knetsch JL, Thaler RH. 1986. Fairness and the assumptions of economics. *J. Bus.* 59:S285–300

Mikhail J. 2007. Universal moral grammar: theory, evidence and the future. *Trends Cogn. Sci.* 11:143–52

Miller DT. 2001. Disrespect and the experience of injustice. *Annu. Rev. Psychol.* 52:527–53

Much NC, Shweder RA. 1978. Speaking of rules: the analysis of culture in breach. *New Dir. Child Dev. Soc. Cogn.* 2:19–39

Nadler J. 2005. Flouting the law. *Tex. Law Rev.* 83:1399–441

Napier J, Tyler T. 2008. Does moral conviction really override concerns about procedural justice? A reexamination of the value protection model. *Soc. Justice Res.* 21:509–28
Nisbett RE, Wilson TD. 1977. Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84:231–59
Robinson PH, Darley J. 1997. The utility of desert. *Northwest. Univ. Law Rev.* 91:453–99
Robinson PH, Darley J. 2007. Intuitions of justice: implications for criminal law and justice policy. *South. Calif. Law Rev.* 81:1–67
Robinson P, Kurzban R. 2007. Concordance and conflict in intuitions of justice. *Minn. Law Rev.* 91:1829–907
Robinson PH, Kurzban R, Jones OD. 2007. The origins of shared institutions of justice. *Vanderbilt Law Rev.* 60(6):1633–88
Rosenmerkel SP. 2001. Wrongfulness and harmfulness as components of seriousness of white-collar offenses. *J. Contemp. Crim. Justice* 17:308–27
Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. 2003. The neural basis of economic decision making in the ultimatum game. *Science* 300:1755–60
Simon J. 2007. *Governing Through Crime: How the War on Crime Transformed American Democracy and Created a Culture of Fear*. New York: Oxford Univ. Press
Skitka LJ. 2002. Do the means always justify the ends, or do the ends sometimes justify the means? A value protection model of justice reasoning. *Personal. Soc. Psychol. Bull.* 28(5):588–97
Sloman SA. 1996. The empirical case for two systems of reasoning. *Psychol. Bull.* 119:3–22
Sloman SA. 2002. Two systems of reasoning. In *Heuristics and Biases*, ed. T Gilovich, D Griffin, D Kahneman, pp. 379–96. New York: Cambridge Univ. Press
Stanovich KE, West RF. 2002. Individual differences in reasoning: implications for the rationality debate. In *Heuristics and Biases*, ed. T Gilovich, D Griffin, D Kahneman, pp. 421–40. New York: Cambridge Univ. Press
Tversky A, Kahneman D. 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90:293–315
Tyler TR. 1990. *Why People Obey the Law*. New Haven, CT: Yale Univ. Press
Tyler TR, Huo Y. 2002. *Trust in the Law*. New York: Russell Sage Found.
van den Haag E. 1987. Punishment: desert and crime control. *Mich. Law Rev.* 85:1250
Walton MD, Sedlak AJ. 1982. Making amends: a grammar-based analysis of children's social interaction. *Merrill-Palmer Q.* 28:389–412
Warr M. 1989. What is the perceived seriousness of crimes? *Criminology* 27:795–821
Wasik M, von Hirsch A. 1990. Statutory sentencing principles: the 1990 white paper. *Mod. Law Rev.* 53:508–17
Young L, Cushman F, Hauser M, Saxe R. 2007. The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci. USA* 104:8235–40