On the Psychology of Punishment

*Cass R. Sunstein**

Are juries rational or irrational? In the context of punitive damage awards, jury decisions suffer from serious problems. Jurors are intuitive retributivists, in a way that produces departures from economic theories of punishment. Their decisions are rooted in outrage, which they cannot easily translate into dollar terms. The result is a degree of unpredictability and incoherence. An understanding of this point casts light on several problems with existing institutions and offers some clues about how those problems might be solved.

I. INTRODUCTION

Judgments about punishment are typically a product of outrage about the underlying acts.[1] Inside as well as outside of law, punitive acts are undertaken when people have been wronged. When legislators penalize misconduct, they are typically responsive to the outrage of their constituents. And when juries punish unlawful acts, either through sentencing or through punitive damage awards, they are often motivated by outrage.[2]

In this essay I defend these points, and use them to make two ma-

* Karl N. Llewellyn Distinguished Service Professor, University of Chicago, Law School and Department of Political Science. This essay is a revised version of a paper presented at a conference on The Law and Economics of Irrational Behavior at George Mason University Law School on Nov. 1, 2002. I am most grateful to participants in the conference for their helpful reactions, and to Daniel Kahneman and David Schkade, my coauthors on the work discussed here.

[1] Jonathan Baron and Ilana Ritov, Intuitions *About Penalties and Compensation in the Context of Tort Law*, 7 J Risk & Uncertainty 17 (1993).

[2] See Cass R. Sunstein et al, *Assessing Punitive Damages*, 107 Yale L J 2071 (1998).

 0-226-64593-2/2004/0011-0005$10.00

jor claims about the relationship between outrage and legal punishments. The first is that it is extremely difficult to translate outrage into the terms that the legal system makes relevant. Because of the difficulty of this task, the legal system risks incoherence in the sense of erratic and unpredictable patterns. Precisely because of the unpredictability of particular awards, these patterns show a kind of irrationality. The second claim is that when people make one-shot judgments, as juries typically do, they are likely to produce patterns that they themselves would repudiate. The result is another kind of incoherence—incoherence not in the sense of unpredictability, but in the sense of patterns that are extremely hard to justify. To that extent the patterns are irrational.

Taken together, these points help to support the Supreme Court's extraordinary decision in *State Farm Mutual Insurance Co. v. Campbell*,[3] in which the Court attempted to discipline punitive awards by juries, in part by suggesting that the punitive award should ordinarily not be more than nine times higher than the compensatory award.[4] The same concerns about unpredictability and incoherence clarify the argument for a number of legal institutions, including the United States Sentencing Commission and various institutions entrusted with producing workers' compensation awards.[5] The same points also raise questions about existing practice in many domains, including the awarding of punitive damages by juries, the system of civil penalties by administrative agencies, and compensatory awards in several areas of law, involving, for example, libel, intentional infliction of emotional distress, pain and suffering, and sexual harassment. I will urge that an understanding of the dynamics of outrage casts light on problems and potential reforms in a variety of areas of law.

In making these claims, I will draw on, and attempt to generalize, a series of experimental studies of punitive damage awards.[6] The resulting work, much of it highly technical, seems to me to have broader implications for a range of issues in both law and politics. If punishment judgments are typically a function of outrage, the problems found in jury behavior might well have analogies in criminal sentencing and administrative fines. If outrage is difficult to translate

[3] 123 S Ct 1513 (2002).

[4] Id.

[5] A helpful overview is Price Fishback and Shawn Kantor, *A Prelude to the Welfare State: The Origins of Workers Compensation* (Chicago 1999).

[6] See Daniel Kahneman et al, *Shared Outrage and Unpredictable Awards*, 16 J Risk & Uncertainty 47 (1998); Sunstein, 107 Yale L J at 2071 (cited in note 2); David Schkade et al., *Deliberating About Dollars: The Severity Shift*, 100 Colum L Rev 1139 (2000); Cass R. Sunstein et al, *Do People Want Optimal Deterrence*, 29 J Legal Stud 237 (2000); Cass R. Sunstein et al, *Predictably Incoherent Judgments*, 54 Stan L Rev 1153 (2002). Many of these papers are collected, in revised and abbreviated versions, in Cass R. Sunstein et al, *Punitive Damages: How Juries Decide* (Chicago 2002).

into dollars or years, we might be able to understand seemingly unjustified disparities from both agencies and courts, and also to see what might be done about those disparities. If people's outrage, when faced with an individual case, produces patterns that people cannot accept, we might be able to identify serious problems with both civil and criminal punishment. In investigating some of these issues, I draw throughout on our empirical findings, but in a way that involves my own extrapolations, some of them admittedly speculative. One of my hopes is that the whole is larger than the sum of the parts. For purposes of the present discussion, I will speak broadly and in qualitative terms. Readers interested in numbers and statistical analysis might consult the papers from which I draw.

II. RETRIBUTION AND OUTRAGE: WHERE PUNISHMENT STARTS

Let us begin with the question of appropriate punishment. On the economic account, the state's goal, when imposing penalties, is to ensure optimal deterrence.[7] To increase deterrence, the law might increase the severity of punishment, or instead increase the likelihood of punishment. A government that lacks substantial enforcement resources might impose high penalties, thinking that it will produce the right deterrent signal in light of the fact that many people will escape punishment altogether. A government that has sufficient resources might impose a lower penalty, but enforce the law against all or almost all violators.

A. Probability of Detection

In the context of punitive damages, all this leads to a simple theory: the major purpose of such damages is to make up for the shortfall in enforcement.[8] If injured people are 100% likely to receive compensation, there is no need for punitive damages. If injured people are 50% likely to receive compensation, those who bring suit should receive a punitive award that is twice the amount of the compensatory award. The simple exercise in multiplication will ensure optimal deterrence. But do people actually want optimal deterrence? Do they accept or reject the economic theory of punishment?

Three simple experiments cast light on these questions.[9] In the

[7] See William Landes and Richard Posner, *The Economic Structure of Tort Law* 160-65, 184-85, 223-24 (Harvard 1993).

[8] See A. Mitchell Polinsky and Steven Shavell, *Punitive Damages: An Economic Analysis*, 111 Harv L Rev 869, 870-76 (1998).

[9] See Sunstein, 29 J Legal Stud 237 (cited in note 6); W. Kip Viscusi, *The Challenge of Punitive Damages Mathematics*, 30 J Legal Stud 313 (2001).

first, people were given cases of wrongdoing, arguably calling for punitive damages, and they were also provided with explicit information about the probability of detection. Different people saw the same case, with only one difference: varying probability of detection (by a factor of 20). People were asked about the amount of punitive damages that they would choose to award. The goal was to see if people would impose higher punishments when the probability of detection was low.

The basic finding was clear.[10] Varying the probability of detection had no effect on punitive awards. Even when people's attention was explicitly directed to the probability of detection, people were indifferent to it. In fact, there was a modest increase in awards when the probability of detection was high, though the difference was not statistically significant. Hence people's decisions about appropriate punishment were largely unaffected by seeing a high or low probability of detection. The evident reason for this result is that people focus on the outrageousness of the defendant's actions, not on the likelihood that they will be detected.[11] Now it is possible that altering the likelihood of detection could increase or decrease outrage. If a corporate defendant was certain to be caught, but nonetheless dumped pollutants into drinking water, it might seem to be brazen, incorrigible, a kind of sadist, entirely deserving of serious punishment. And if a corporate defendant engaged in some act in an especially stealthy way, showing a real skill at evading the law, people might want to punish it especially severely, on the ground that stealth and skill are grounds for heightened outrage. But in either case, a high or low likelihood of detection is operating not in the economic fashion, as part of a deterrence calculus, but instead as a part of an inquiry into the egregiousness of the acts.

There was something indirect about the first experiment: It showed people the probability of detection, but it did not ask them to evaluate judgments that took probability into account. The second experiment filled that gap.[12] It asked people to evaluate judicial and executive decisions to reduce penalties when the probability of detection was high, and to increase penalties when the probability of detection was low. People were asked to say whether they approved or disapproved of official decisions to vary the penalty with the probability of detection. Strikingly, strong majorities of respondents rejected judicial decisions to reduce penalties because of high probability of detection, and also rejected executive decisions to increase penalties be-

[10] See Sunstein, 29 J Legal Stud 237 (cited in note 6).
[11] We will see further evidence of this in Part III below.
[12] See Sunstein, 29 J Legal Stud 237 (cited in note 6).

cause of low probability of detection. In other words, people did not approve of an approach to punishment that would make the level of punishment vary with the probability of detection. What apparently concerned them was the extent of the wrongdoing, and the right degree of moral outrage rather than optimal deterrence.

A third study was the most direct of all.[13] This study asked people to undertake a deterrence calculus, based squarely on the compensatory award and the probability of detection. All the necessary information was placed before jurors. The result was that people did not successfully perform the elementary calculations. Errors were pervasive. From this experiment, it is not entirely clear whether people erred because they were unable to do what they were asked, or because they refused to do it, on the ground that it ran afoul of their moral convictions. But in either case, it is clear that people do not spontaneously think in terms of optimal deterrence, and indeed, they will fail to do so even if specifically requested to engage in that task.

The most general conclusion is that people are intuitive retributivists. Their moral intuitions are inconsistent with the economic theory of deterrence. Those intuitions are grounded in outrage.

B. Pointless Punishment?

Other studies support these findings. For example, Baron and Ritov studied people's judgments about penalties in tort cases involving harms resulting from the use of vaccines and birth control pills.[14] In one case, subjects were told that the result of a higher penalty would be to make companies try harder to make safer products. In an adjacent case, subjects were told that the consequence of a higher penalty would be to make the company more likely to stop making the product, with the result that less safe products would be on the market. Most subjects, including a group of judges, gave the same penalties in both cases.

A related study found no reduction in penalty even when subjects were told that the amount of the penalty would have no effect on future behavior—because the penalty was secret, the company had insurance, and the company was about to go out of business.[15] This study strongly suggests that punishment judgments are retributive in character, not tailored to consequentialist goals.

Another test of punishment judgments asked subjects, including both judges and legislators, to choose penalties for dumping hazardous

[13] See Viscusi, 30 J Legal Stud 313 (cited in note 9).
[14] Baron, 7 J Risk & Uncertainty 17 (cited in note 1).
[15] See id.

waste.[16] In one case, the penalty would make companies try harder to avoid waste. In another, the penalty would lead companies to cease making a beneficial product. Most people did not penalize companies differently in the two cases. Perhaps most strikingly, people preferred to require companies to clean up their own waste, even if the waste did not threaten anyone, instead of spending the same amount to clean up far more dangerous waste produced by another, now-defunct company. These studies indicate that when assessing punishment, people's judgments are rooted in outrage; they do not focus solely on social consequences, at least not in any simple way.

C. Cost-Benefit Analysis

A related test of punitive intuitions attempted to explore whether jurors would punish or reward companies that conducted a competent cost-benefit analysis before proceeding.[17] The test asked people to assess different scenarios involving safety precautions. In some of them, the company did no explicit cost-benefit analysis, but simply concluded that the company "thought that there might be some risk from the current design, but did not believe it would be significant." In other scenarios, companies engaged in cost-benefit analysis, with varying amounts used to value life (from $800,000 to $4 million). The key question is this: Will people reward or punish companies that have explicitly weighed costs against benefits?

The answer is that people do not react favorably to this kind of weighing.[18] A company that engages in cost-benefit balancing is very likely to face a punitive award, and the award that it faces is likely to be high. In fact, companies that place a high monetary value on human life are likely to face especially high awards. By contrast, people do not much punish companies that are willing to impose a risk on people.[19]

There is a real oddity here: If the costs of precautions outweigh the benefits, then companies should not, under ordinary understandings, be deemed negligent at all. If jurors are punishing companies in such circumstances, it must be because of a kind of moral outrage that has little to do with either efficiency or law.[20] And why are awards espe-

[16] Jonathan Baron, R. Gowda, and Howard Kunreuther, *Attitudes toward Managing Hazardous Waste: What Should be Cleaned Up and Who Should Pay for It*, 13 Risk Analysis 183 (1993);

[17] See W. Kip Viscusi, *Corporate Risk Analysis: A Reckless Act?*, 52 Stan L Rev 547 (2000).

[18] See id.

[19] See id; see also Philip Tetlock, *Coping With Tradeoffs, in Elements of Reason: Cognition, Choice, and the Bounds of Rationality* 239 (Cambridge, 2000).

[20] What underlies these judgments? I cannot fully answer that question here. But a careful look raises the possibility that people are outraged by any explicit decision to

cially high when companies place a high value on human life? The most likely answer is that jurors have a difficult time in coming up with dollar amounts to punish misconduct and that a high value operates as an anchor, leading to high punitive judgments.[21] This point directly bears on the problem of translating moral judgments into monetary terms, to which I will shortly turn.

Is it irrational to root punishment judgments in outrage? There is no simple answer. Many distinguished observers have argued in favor of retributive conceptions of punishment, and there is a clear connection between retribution and outrage.[22] To the extent that people are using a kind of "outrage heuristic," they cannot be shown to be making the sorts of errors to which ordinary heuristics sometimes lead them.[23] On the other hand, it is possible to worry about the potentially harmful social consequences of a system of punishment that is rooted in outrage. I share that worry; but to those who believe in the rationality of outrage, or in retribution generally, the issue cannot be resolved without a complex normative argument. If we have ir-

---

trade money for risks. When they are generalizing from a set of moral principles that are generally sound, and even useful, but that work poorly in some cases. Consider the following moral principle: Do not knowingly cause a human death. People disapprove of companies that fail to improve safety when they are fully aware that deaths will result—whereas people do not disapprove of those who fail to improve safety while appearing not to know, for certain, that deaths will ensue. When people object to risky action taken after cost-benefit analysis, it seems to be partly because that very analysis puts the number of expected deaths squarely "on screen." Companies that fail to do such analysis, but that are aware that a risk exists, do not make clear, to themselves or to jurors, that they caused deaths with full knowledge that this was what they were going to do. People disapprove, above all, of companies that cause death knowingly. I suggest, then, that a genuine heuristic is at work, one that imposes moral condemnation on those who knowingly engage in acts that will result in human deaths. The problem is that it is not always unacceptable to cause death knowingly, at least if the deaths are relatively few and an unintended byproduct of generally desirable activity. If government allows new highways to be built, it will know that people will die on those highways; if government allows new power plants to be built, it will know that some people will die from the resulting pollution; if companies produce tobacco products, and if government does not ban those products, hundreds of thousands of people will die; the same is true for alcohol. Much of what is done, by both industry and government, is likely to result in one or more deaths. Of course, it would make sense in most or all of these domains, to take extra steps to reduce risks. But that proposition does not support the implausible claim that we should disapprove, from the moral point of view, of any action taken when deaths are foreseeable.

[21] See Cass R. Sunstein, *Hazardous Heuristics*, U Chi L Rev (forthcoming 2002).

[22] See David Owen, *The Moral Foundations of Punitive Damages*, 40 Ala L Rev 705 (1989); Marc Galanter and David Luban, *Poetic Justice*, 42 Am U L Rev 1393 (1993); Jean Hampton, *The Retributive Idea*, in Jean Hampton and Jeffrie Murphy, *Forgiveness and Mercy* 111 (Cambridge 1988).

[23] The key papers can be found in Daniel Kahneman, Paul Slovic, and Amos Tversky, eds, *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge 1982).

rationality here, it is not irrationality in any simple sense. But for the legal system, the use of outrage does lead to serious problems, as we shall now see.

III. THE TRANSLATION PROBLEM

Punitive judgments are rooted in outrage; but do people agree about the appropriate level of outrage? Imagine that diverse Americans are confronted with a case of clear wrongdoing. Should we expect a sharp divergence if they are asked to answer, on a bounded numerical scale, "how bad was the underlying conduct?" I now offer evidence suggesting that people's outrage is widely shared, but that the consensus breaks down when people are asked to translate their outrage into dollars. The Supreme Court's evident concern about arbitrary punitive awards[24] has a sound basis in the psychology of punishment, as we shall now see.

A. Shared Outrage

A series of studies of citizen judgments demonstrates that at least in some domains, people agree about the degree of outrage that appropriately fits social misconduct.[25] At least if people use a bounded scale (of, say, 0 to 6 or 0 to 8) with accompanying verbal descriptions ("not at all outrageous" for 0 and "extremely outrageous" for 6 or 8), a high degree of social agreement is likely. In personal injury cases, the judgment of any particular group of six is likely to provide a good prediction of the judgment of any other group of six. In this sense, a "moral judgment" jury is indeed able to serve as the conscience of the community.

In one study, people were asked to assess the outrageousness of the defendant's conduct on a bounded scale, and separately to say how much the defendant should be punished on that scale. Two striking facts emerged. The first was an extraordinary degree of correlation between judgments of outrageousness and judgments about appropriate punishment—a finding that confirms the suggestion in Part I that punishment judgments are rooted in outrage.[26] The second was a high degree of regularity in both sets of judgments, so that people tend to rank and to rate diverse cases in essentially the same way. At least in a set of highly varying personal injury cases, people's punishment judgments do not significantly diverge, and the assessment of one jury is a good predictor of the assessment of another.

[24] *State Farm Mut Ins Co v. Campbell*, 123 S. Ct. 1513 (2002).

[25] See Sunstein, 107 Yale L J at 2071 (cited in note 2); Schkade, 100 Colum L Rev at 1139 (cited in note 6).

[26] See Sunstein, 107 Yale L J at 2071 (cited in note 2).

Indeed we can go further. Members of different demographic groups show considerable agreement about how to rank and rate personal injuries cases.[27] Thousands of people were asked to rank and rate cases. Information was elicited about the demographic characteristics of all of those people. As a result, it is possible, with the help of the computer, to put individuals together, so as to assemble all-male juries, all-female juries, all-white juries, all-African-American juries, all-poor juries, all-rich juries, all-educated juries, all less-educated juries, and so forth. Creating "statistical juries" in this way, there were no substantial disagreements, in terms of rating or ranking, within any group. In personal injury cases, people largely agree with one another.

Subsequent work has broadened this finding, showing that people agree on how to rank tax violations, environmental violations, and occupational safety and health violations.[28] From this evidence, it seems reasonable to hypothesize that in a wide range of domains, people will agree how to rank and rate cases. The moral norms within a heterogeneous culture are, to that extent, widely shared, and strikingly so. Now this does not mean that people will agree on how to rank cases from different categories (a point to which I will return). Nor does it mean that small groups will always agree on how to do the ranking. Nor does it mean that demographically diverse groups will agree about how to rate cases in contentious areas of the law—consider sexual harassment or racial discrimination. But the findings do suggest that within category, disagreement about both outrage and punishment is the exception, not the rule.

B. Erratic Dollar Awards

There is a consensus about the appropriate level of outrage. But even when that consensus exists, there is no consensus about appropriate punishment in terms of dollars. As we shall see, the reason for the lack of consensus lays in particular properties of the dollar scale. The scale of years in jail, used for criminal punishment, suffers from similar problems.

With respect to dollars, both individuals and jury-size groups are all over the map.[29] Even when moral rankings are shared—as they generally are—dollar awards are extremely variable. A group that awards a "5," for defendant's misconduct, might give a dollar award of $500,000, or $2 million, or $10 million. A group that awards a "7" might award $1 million, or $10 million, or $100 million. In fact, there

[27] Id.

[28] See Cass R. Sunstein et al., *Legal Coherence and Incoherence* (unpublished manuscript, 2001).

[29] See Sunstein, 107 Yale L J at 2071 (cited in note 2); Schkade, 100 Colum L Rev at 1139 (cited in note 6).

is so much noise, in the dollar awards, that differences cannot be connected with demographic characteristics. It is not as if one group—whites for example—give predictably different awards from another—say African-Americans or Hispanics. We cannot show systematic differences between young and old, men and women, well-educated and less well-educated. The real problem is that dollar awards are quite unruly, from one individual to another and from one small group to another.

C. Why Are Awards Erratic?

These findings raise an obvious question: why are erratic dollar awards found amidst shared moral judgments? The best answer involves the problem of translating outrage into dollars. More particularly, the answer is that the effort to "map" moral judgments onto dollars is an exercise in "scaling without a modulus."[30] In psychology, it is well known that serious problems will emerge when people are asked to engage in a rating exercise on a scale that is bounded at the bottom but not at the top, and when they are not given a "modulus" by which to make sense of various points along the scale. For example, when people are asked to rate the brightness of lights, or the loudness of noises, they will not be able to agree if no modulus is supplied and if the scale lacks an upper bound. But once a modulus is supplied, agreement is substantially improved. Or if the scale is given an upper bound, and if verbal descriptions accompany some of the relevant points, people will come into accord with one another.

The upshot is that much of the observed variability with punitive damage awards—and in all likelihood with other damage awards too—does not come from differences in levels of outrage. It comes from variable, and inevitably somewhat arbitrary, "moduli" selected by individual jurors and judges. If the legal system wants to reduce the problem of different treatment of the similarly situated, it would do well to begin by appreciating this aspect of the problem. The point applies to many legal problems, including criminal sentences, pain and suffering awards, administrative penalties, and damages for libel, sexual harassment, and intentional infliction of emotional distress. In these areas as well, those entrusted with the task of "mapping" lack a modulus with which to discipline their decisions. An empirical study of pain and suffering awards finds that no less than forty percent of the variance cannot be explained by differences in case characteristics.[31] A legal system that does not give guidance for "map-

[30] Schkade, 100 Colum L Rev at 1139 (cited in note 6).

[31] David Leebron, *Final Moments: Damages for Pain and Suffering Prior to Death*, 64 NYU L Rev 256 (1989).

ping" is bound to create similar problems in other areas. Indeed, the rise of guidelines for criminal sentencing can be understood as responsive, at least in part, to exactly this problem.

D. The Effects of Deliberation

The study just described involved individual judgments, aggregated, with the aid of the computer, so as to produce statistical jurors. A subsequent study tested the effects of deliberation on both punitive intentions and dollar judgments.[32] The study involved about 3000 jury-eligible citizens; its major purpose was to determine how individuals would be influenced by seeing and discussing the punitive intentions of others. To test the effects of deliberation on punitive intentions, people were asked to record their individual judgments privately, on a bounded scale, and then to join six-member groups to generate unanimous "punishment verdicts." Hence, subjects were asked to record, in advance of deliberation, a "punishment judgment" on a scale of 0 to 8, where 0 indicated that the defendant should not be punished at all, and 8 indicated that the defendant should be punished extremely severely. After the individual judgments were recorded, jurors were asked to deliberate to a unanimous "punishment verdict."

Two findings are especially important. First, deliberation made the lower punishment ratings decrease, when compared to the median of pre-deliberation judgments of individuals—while deliberation made the higher punishments ratings increase, when compared to that same median. When the individual jurors favored little punishment, the group showed a "leniency shift," meaning a rating that was systematically lower than the median predeliberation rating of individual members.[33] But when individual jurors favored strong punishment, the group as a whole produced a "severity shift," meaning a rating that was systematically higher than the median predeliberation rating of individual members.[34] When the median juror judgment was less than four, the jury's verdict was below the median judgment of individuals.[35]

The second important finding is that dollar awards of groups were systematically higher than the median of individual group members—so much so that in 27% of the cases, the dollar verdict was as high as, or higher than, that of the highest individual judgment, predeliberation. The basic result is that deliberation causes awards to increase, and it causes high awards to increase a great deal. The effect of deliberation, in increasing dollar awards, was most pronounced in the

[32] Schkade, 100 Colum L Rev 1139 (cited in note 6).
[33] Id. at 1152, 1154-55.
[34] Id.
[35] Id.

case of high awards. For example, the median individual judgment, in a case involving a defective yacht, was $450,000, whereas the median jury judgment, in that same case, was $1,000,000.[36] But awards shifted upwards for low awards as well.[37]

These findings create many puzzles. For present purposes, the key point is that the translation problem is not cured by deliberating bodies. On the contrary, the problem of unpredictability is increased, not decreased, by the existence of deliberation. If we are seeking an explanation for the movements that are observed, the best answer lies in the phenomenon of group polarization.[38] This is the pervasive process by which group members end up in a more extreme position in line with the predeliberation tendencies of group members. It is predicted, according to group polarization, that high levels of outrage will be increased by deliberation, and that low levels of outrage will be decreased by deliberation. Nor do such movements present any real puzzles for rationality. A central reason for group polarization involves the exchange of information within the group. In a group that favors a high punishment rating, group members will make many arguments in that direction, and relatively few the other way. Speaking purely descriptively, the group's "argument pool" will be skewed in the direction of severity. Group members, listening to the various arguments, will naturally move in that direction.

In the context of actual dollar awards by juries, a particular finding deserves emphasis.[39] As I have noted, the highest awards increased by the largest amount, but all awards increased. This might appear to be a surprise. An understanding of group polarization might suggest that low awards would drop and high awards would be raised, with the difference pivoting around some neutral point, say, $60,000. But this is not what was observed. Why did dollar awards systematically increase? A possible explanation, consistent with group polarization, is that any positive median award suggests a predeliberation tendency to punish, and deliberation aggravates that tendency by increasing awards. But even if correct, this explanation seems insufficiently specific. The striking fact is that those arguing for higher awards seem to have an automatic "rhetorical advantage" over those arguing for lower awards. A subsequent study of supported this finding, suggesting that given prevailing social norms, people find it much easier to defend higher awards against corporate defendants than the opposite.[40] If this is so, then processes of deliberation will naturally lead

[36] Id. at 1152.
[37] Id.
[38] See Roger Brown, *Social Psychology: The Second Edition* (Simon & Schuster 1990).
[39] See Schkade, 100 Colum L Rev 1139 at 1149-1151 (cited in note 6).
[40] Id at 1161-62.

to jury awards that are systematically higher than the award of the median individual member in advance of deliberation.

My major goal here, however, is not to investigate the sources of movements in awards, but to suggest that group deliberation does not solve the translation problem. Deliberating juries, no less than statistical juries, show a high degree of consensus about appropriate punishment (and hence outrage). Deliberating juries, even more than statistical juries, show a high degree of variability in terms of appropriate dollar awards.

IV. OUTRAGE, PUNISHMENT, AND CONTEXT

I now turn to the largest puzzles of all. Thus far, it has seemed as if people's moral judgments are quite stable. But that proposition was established only by looking at a set of personal injury cases. Here the relevant level of outrage is both predictable and coherent, in the sense that people's judgments are not much affected by whether they are seeing cases in isolation or simultaneously and in the context of other cases from the same category. But is this coherence maintained when people look at cases from different categories? Suppose, for example, that people are evaluating personal injury cases and cases involving commercial fraud, and that similar people are evaluating personal injury cases and cases involving rape and murder. Would the judgments about personal injury cases remain stable across these various contexts? Are judgments about cases different, depending on whether those cases are seen in isolation or in the context of cases from other categories?

We do not have full answers to these questions; but suggestive evidence has started to emerge.[41] It appears that people agree on how to rank cases within categories and that their judgments about particular cases are not affected by seeing them in isolation or alongside other cases within the same categories.[42] It also appears that people have a kind of implicit ranking of categories themselves; they think that murder is worse than rape, that rape is worse than assault, and that assault is worse than libel. But when people are trying to rank cases from different categories, they have far more difficulty, in the sense that they are unsure exactly what to do. They are not certain, for example, whether a relatively bad income tax violation is worse than a relatively not-so-bad occupational safety and health violation.[43] This lack of certainty translates into a lack of consensus. People agree

[41] See Sunstein, 54 Stan L Rev 1153 (cited in note 6).
[42] See Sunstein, 107 Yale L J at 2071 (cited in note 2).
[43] See Sunstein, *Legal Coherence and Incoherence* (cited in note 25).

much more on how to rank cases within a category than how to rank cases across categories. Note that I am putting aside the evident difficulties in deciding what counts as a "category." It is easy to design experiments in which people will simply disagree about whether, for example, a comparatively serious tax violation is worse, or less bad, than a lawless act that harms the environment. Hence, the social norms that govern cross-category comparisons are not as widely shared as the social norms that govern within-category comparisons. It follows that judgments about outrage, and about appropriate punishment, are more variable across categories than within categories.

Perhaps this is not big news. A more striking finding is that people's judgments about cases, taken one at a time, are very different from their judgments about the same cases, taken in the context of a problem from another category.[44] For example, people were asked to assess a case involving a personal injury, on a bounded punishment scale and also on a dollar scale. People were also asked to assess a case involving financial injury, on a bounded punishment scale and also on a dollar scale. The financial injury involved relatively egregious misconduct, such as a violation of trust by a trustee, for the benefit of a favored client; the personal injuries were relatively less egregious, such as an injury caused to a driver when a steering system failed. The basic goal was to ask people to assess, in isolation but then in comparison, a financial injury case that would seem outrageous for its type, and a personal injury case that would seem less outrageous for its type.

Here is what emerged.[45] When each of the two cases is judged in isolation, the financial injury case receives a more severe punishment rating and a higher dollar award. But when the two cases are seen together, there is a significant judgment shift, in which people try to ensure that the financial award is not much higher, and for many respondents is lower, than the personal injury award. The upshot is that people's decisions about the two cases are very different, depending on whether they see the case alone or in the context of a case from another category.[46]

Exactly the same kind of shift was observed for judgments about two problems calling for government regulation and expenditures: research on bone marrow cancer among the elderly and protection of coral reefs by banning of cyanide fishing.[47] Looking at the two cases in isolation, people are willing to pay about the same to protect coral reefs, and register more satisfaction, on a bounded scale, from doing

[44] See Sunstein, 54 Stan L Rev at 1173-1178 (cited in note 6).
[45] Id at 1176.
[46] Id.
[47] Id at 1176-1177.

that. But looking at the two cases together, people will be quite disturbed at this pattern, and will want to pay significantly more to protect elderly people from cancer and will also register more satisfaction from doing that. Here too there is a significant shift in judgment.

In these findings, the translation problem is not the source of the difficulty. People's evaluations shift depending on whether they see cases in isolation or in the context of cases from other categories. What accounts for these shifts? Let me offer a preliminary account. When people see a case in isolation, they naturally normalize it by comparing it to a set of comparison cases that it readily calls up. If people are asked whether a German Shepherd is big or small, they are likely to respond that it is big; if they are asked whether a Volkswagen Bug is big or small, they are likely to respond that it is small. But people are well aware that a German Shepherd is smaller that a Volkswagen bug. People answer as they do because a German Shepherd is compared with dogs, whereas a Volkswagen Bug is compared with cars. So far, so good; in these cases, everyone knows what everyone else means. We easily normalize judgments about size, and the normalization is mutually understood. Another example is John Stockton, who is about six feet tall, and hence a small basketball player. What happens, in ordinary communication, is innocuous. It does not breed error or confusion.

In the context of legally relevant moral judgments, something similar happens, but it is far from innocuous. When evaluating a case involving financial injury, people apparently normalize the defendant's conduct by comparing it with conduct in other cases from the same category. They do not easily or naturally compare that defendant's conduct with conduct from other categories. Because of the natural comparison set, people are likely to be quite outraged by the misconduct, if it is far worse than what springs naturally to mind. The same kind of thing happens with the problem of bone marrow cancer among the elderly. People compare that problem with other similar problems, and conclude that it is not so serious, within the category of health-related or cancer-related problems. The same is true with personal injury cases (normalized against other personal injury cases) and problems involving damage to coral reefs (normalized against other cases of ecological harm).

When a case from another category is introduced, this natural process of comparison is disrupted. Rather than comparing a cancer case involving the elderly with other cancers, or other human health risks, people see that it must be compared with ecological problems, which (in most people's view) have a lesser claim to public resources. Rather than comparing a financial injury case to other cases of business misconduct, people now compare it to a personal injury case, which (in most people's view) involves more serious wrongdoing. As a result of

the wider viewscreen, judgments shift, often dramatically. It follows that if people's informed judgments are taken to be the criterion, punitive damage awards are likely to be too high in financial injury cases and too low in personal injury cases. Some data supports this suggestion.[48] Similar shifts could be produced with many other pairs of categories. For example, punitive damages awards involving libel might well be higher, in isolation, than punitive damage awards involving racial discrimination; but there is likely to be a reversal if the two cases are put together.

It is reasonable to hypothesize that the comparative situation may alter judgments in another way, by reducing the anchoring effects of compensatory damages on punitive awards. In the real world of punitive damages, unlike our experiment, compensatory awards are generally much larger in financial cases than in cases of physical injury.[49] As a consequence, a case of financial damage with a large compensatory anchor (say $10,000,000) is expected to receive a higher punitive damage award than a case of physical injury with a smaller anchor (say $500,000), when the two are judged in isolation. When cases of the two kinds are directly compared, many people will be more strongly influenced by the relative prominence of the harms than by the relative size of the anchors. Preliminary evidence[50] supports this hypothesis, which suggests that two distinct mechanisms may cause punitive awards for financial cases to be higher in the current system than they would be if jurors were given a richer context: anchoring on high dollar numbers, and masking of the low prominence of the category through the effect of normalization.

I believe that this uncovers a serious problem with current practice in many domains of law. The problem is that when people assess cases in isolation, their viewscreen is usually narrow, indeed often limited to the category to which the case belongs, and that as a result, people produce a pattern of outcomes that makes no sense by their own lights. In other words, the overall set of outcomes is one that people would not endorse, if they were only to see it as a whole. Their considered judgments reject the very pattern that they have produced, because of a predictable feature of human cognition. The result is a form of incoherence.

We can find that incoherence not only in jury verdicts, but also in administrative fines and in criminal sentencing, where no serious

[48] See Jonathan Karpoff and John Lott, *On the Determinants and Importance of Punitive Damage Awards*, 42 J L & Econ 527, 539 (1999).

[49] Id at 538-39, showing mean awards of $14.8 million in fraud cases, and $20.6 million in business negligence cases, but $6.2 million in product liability cases, $1.6 million in malpractice cases, and $991,000 in motor vehicle accident cases.

[50] See Sunstein, *Legal Coherence and Incoherence* (cited in note 25).

effort has been made to ensure that the overall pattern of outcomes makes the slightest sense.[51] Indeed there is reason to believe that the pattern, in many domains, is quite senseless. And it may not be too much of a stretch to suggest that the same is true of reactions, some of the time, by both individuals and institutions—that people are quite outraged about behavior that, in a broader or different comparison set, would outrage them little or not at all.

What should be done by way of legal reform? I cannot answer that question here. If outrage is the appropriate basis for punishment, then steps should be taken to ensure that outrage reflects a wide viewscreen rather than a narrow one, so as to reduce the risk that the legal system will produce punishment patterns that people reject. Perhaps the United States Sentencing Commission can be understood partly in this light; and perhaps an emphasis on incoherence suggests directions in which the Commission might go in the future. To date, the Commission has made little effort to ensure that penalties cohere across categories.

In the context of punitive damages, the claims I have made suggest that judges might take a stronger role in overseeing jury awards, in part to ensure that those awards cohere with what has been done in other areas of the law. The natural implication is that judges should decrease unjustifiably high awards and increase unjustifiably low ones; and indeed the Supreme Court's unexpectedly ambitious ruling in the *State Farm* case is a clear step in the direction of imposing discipline on jury awards.[52] It might be tempting to reject this suggestion by emphasizing the populist credentials of the jury and by fearing judicial usurpation of the jury's functions. But the translation problem, and the risk of incoherence from one-shot judgments, demonstrates that this concern is misplaced, because the decisions of any particular jury do not produce community sentiment about what patterns of punishment make sense.

Outside of the domain of punitive awards, it would make sense to try to systematize civil fines in general, so as to ensure that the penalties imposed by, for example, the Environmental Protection Agency fit well with the penalties imposed by say the Internal Revenue Service, the Fish and Wildlife Service, and the Occupational Safety and Health Administration. The discussion thus far suggests that any effort at systematization would present a daunting task. But incoherent judgments are extremely likely in the administrative arena as well, and at least it would be worthwhile to attempt to correct the most egregious anomalies.

[51] See Sunstein, 54 Stan L Rev at 1189-1196 (cited in note 6).

[52] *State Farm Mut Ins Co v. Campbell*, 123 S. Ct. 1513 (2002).

V. CONCLUSION

In this paper I have urged that punishment judgments are rooted in outrage and that people do not naturally think in terms of optimal deterrence. With respect to punishment, people are intuitive retributivists. I have also suggested that punishment judgments are rooted in outrage, which is, in an important respect, stable across individuals or at least small groups. But for purposes of operating a legal system, punishments that are based on outrage present two key problems. The first is that the legal system does not attempt to measure outrage directly, but instead requires people to translate their moral opprobrium into the unbounded scale of dollars. This act of translation produces unpredictable and arbitrary awards. The second problem is that outrage is category-specific. People's level of outrage is a function of comparison cases. When they confront a case in isolation, they evaluate it by comparing it not to the full universe of cases, but to a natural set of similar cases. When cases from other categories are introduced, their outrage is shifted. The result is that when making decisions in isolation, people produce patterns of outcomes that they themselves repudiate once those decisions are seen together.

These findings raise a number of problems. Economically oriented observers reject the idea that punishment should be rooted in outrage, which could easily result in too much and too little deterrence. For those who believe that punishments should not be an outgrowth of outrage, it is wrong to base civil and criminal punishments on ordinary intuitions. In addition, the translation problem ensures a high degree of unexplained noise in punishments, resulting in unclear signals to possible defendants and also ensuring that similarly situated plaintiffs and defendants will not be treated similarly. And for those who would like to take outrage seriously, and who believe in retributive goals, the existence of incoherence raises serious problems of its own, above all because it suggests that one-shot judgments by juries will not reflect the levels of outrage that would come from a wider viewscreen.

For many purposes, outrage is highly desirable from the social point of view. But when operationalized into legal terms, it tends to produce punishments that are both unpredictable and incoherent, and to result in systems that fall far short of rationality. I do not suggest that an understanding of the psychology of punishment clearly supports any particular set of legal reforms. But such an understanding helps to explain many problems with existing institutions, and offers a number of clues about how those problems might be solved.