Behavioural Public Policy (2025), **9**, 293–300
doi:10.1017/bpp.2024.33

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

The outrage heuristic

Cass R. Sunstein

Harvard Law School, USA
Email: csunstei@law.harvard.edu

(Received 31 July 2024; accepted 31 July 2024)

Abstract

Many moral judgments are rooted in the outrage heuristic. In making such judgments about certain personal injury cases, people's judgments are both predictable and widely shared. With respect to outrage (on a bounded scale of one to six) and punitive intent (also on a bounded scale of one to six), the judgments of one group of six people, or 12 people, nicely predict the judgements of other groups of six people, or 12 people. Moreover, outrage judgments are highly predictive of punitive intentions. Because of their use of the outrage heuristic, people are intuitive retributivists. People care about deterrence, but they do not think in terms of optimal deterrence. Because outrage is category-specific, those who use the outrage heuristic are likely to produce patterns that they would themselves reject, if only they were to see them. Because people are intuitive retributivists, they reject some of the most common and central understandings in economic and utilitarian theory. To the extent that a system of criminal justice depends on the moral psychology of ordinary people, it is likely to operate on the basis of the outrage heuristic and will, from the utilitarian point of view, end up making serious and systematic errors.

Keywords: outrage; heuristic; moral heuristics; punitive damages; punishment

Introduction: À La Recherche Du Temps Perdu

From about 1995 to 2005, I was privileged to collaborate with Daniel Kahneman (and several others, above all David Schkade) on a series of papers on the topic of moral intuitions in the domain of punishment (Kahneman *et al.*, 1998; Sunstein *et al.*, 1998, 2000, 2002; Schkade *et al.*, 2000; Kahneman and Sunstein, 2005). Our narrow topic was punitive damages. We were concerned with sources of variability in jury judgments. But as our work continued, the viewscreen greatly broadened. We learned something about punishment judgments in general and about the cognitive psychology of moral intuitions. Among other things, we learned that people are intuitive retributivists. We learned, back in 2005, that System 1 and System 2 provide a useful way of understanding how cognitive psychology works in the moral domain, and that with respect to punishment, System 1 is in charge. Above all, we learned about the operations of what we called, offhand and very late in the game, 'the outrage heuristic'

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

(Kahneman and Sunstein, 2005). I like to think that Danny would be pleased to see an article with that name.

My purpose here is to offer an overview, an elaboration, and a generalization of our key findings, as well as to connect them to a topic on which Kahneman and I also collaborated: noise, understood as unjustified variability in judgments (Kahneman *et al.*, 2021). I hope that a brief and broad treatment, centered on the use of the outrage heuristic, will both extend and generalize our diverse arguments. I also hope that such a treatment might prove instructive for behavioral public policy in the domain of civil and criminal punishment. Behaviorally informed research on those topics remains a work in progress. As it happens, our work on noise owes a debt to a series of findings, almost two decades before, on the subject of moral intuitions, outrage, punitive intentions, and monetary measures (Kahneman *et al.*, 2021).

A personal note: Working with Danny, I was struck by a particular aspect of his method. He would consult his own intuitions, notice their limitations, wonder whether other people might share his intuitions, and if he suspected that he did, proceed to test them. He investigated his System 1 and tested it by reference to System 2. If System 1 was off the mark, Danny was off to the races.

Danny was a wise and kind person, but because he was a person, he would occasionally feel outraged. After he did, he would usually calm down – and feel both embarrassed and amused by the intensity of his feelings. Like everyone else, he could certainly feel retributive impulses (I think) and also see that they might produce foolishness, mischief, or worse. Danny had a keen intuitive sense that turning outrage into something concrete (a damage award, a monetary fine, or a jail sentence) would produce a lot of noise. He himself did not have a clear sense of how to concretize outrage. Would others? His appreciation of human frailty, evident in so much of his work, played a defining role in our project.

More particularly, our principal findings were as follows:

1. In making moral judgments about personal injury cases, people's judgments are both predictable and widely shared. With respect to outrage (on a bounded scale of one to six) and punitive intent (also on a bounded scale of one to six), the judgments of one group of six people, or 12 people, nicely predict the judgments of other groups of six people, or 12 people. These shared judgments cut across demographic differences, so that there is no difference, in the relevant cases, between rich and poor, old and young, white and African-American, and poorly educated and well-educated. People are intuitive retributivists, and their intuitions are widely shared.
2. In making punitive damage awards about personal injury cases, people's judgments, on a monetary scale, are highly unpredictable and far from shared. In this domain, there is a great deal of noise. People do not have a clear sense of the *meaning* of different points along the scale of dollars. Hence, the dollar judgments of one group of six people, or 12 people, do not well predict the dollar judgments of other groups of six people, or 12 people. The reason is simple: people are engaged in the process of 'scaling without a modulus', which is a predictable source of unjustified variability or 'noise'.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

3. As compared with the median of predeliberation judgments, the effect of deliberation is to increase dollar awards, often quite substantially. Group discussions have the remarkable effect of raising group members' judgments about appropriate punishment. And when people are outraged, group discussion makes them more outraged still.
4. People care about deterrence, but they do not think in terms of optimal deterrence. Because people are intuitive retributivists, they reject some of the most common and central understandings in economic and utilitarian theory.

Now for some details. For the 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 shall draw.

Outrageousness and noisy awards

Suppose that people are asked to assess a set of personal injury cases, libel cases, or cases involving sexual harassment or damage to the environment. Suppose, more specifically, that people are asked to rate their level of outrage on a bounded numerical scale – say, zero to six, where zero means 'completely acceptable' and six means 'absolutely outrageous'. Suppose too that people are asked to rate those cases, in terms of appropriate punishment, on a bounded numerical scale – say, zero to six, where zero means 'no punishment', and six means 'punished extremely severely'. Will people agree? Will the decision of one group of six or 12 provide good predictions about what other groups of six or 12 will do? The answer will depend on whether the social norms that govern moral outrage and intended punishment are widely shared. If they are shared, we should not expect sharp divergences in terms of both ranking and rating.

Undertaking a series of studies of citizen judgments, we found that in the tested domains, the relevant norms are indeed widely shared. In personal injury cases, at least, the judgment of any particular group of six, on a bounded scale involving both outrage and punishment, is highly 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. We also found that people's outrage judgments and punishment judgments were highly correlated. Outrage, measured on a bounded scale, predicted punitive intent, also measured on a bounded scale.

Indeed, we can go further. On a bounded scale, members of different demographic groups show considerable agreement about how to rank and rate personal injury cases in terms of both outrage and punishment. 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, we found no substantial disagreement, in terms of rating or ranking, within any group. In personal injury cases, people simply agree, again on a bounded scale measuring both outrage and punishment.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

What about dollars? Do the broadly shared norms also produce regularity in jury verdicts? One of our central findings is that they do not. With respect to dollars, both individuals and jury-size groups are all over the map. Even when moral rankings are shared – as they generally are – dollar awards are extremely variable. A group that awards a 'four' for a defendant's misconduct (with respect to outrage and punishment on a bounded scale) might give a dollar award of $500,000, or $2 million, or $10 million. A group that awards a 'six' might award $1 million, or $10 million, or $100 million. In fact, there is so much noise in the dollar awards that differences cannot be connected to demographic characteristics. It is not as if one group – whites, for example – gives predictably different awards from another – say, African–Americans or Hispanics. We cannot show systematic differences between young and old, men and women, and 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.

What accounts for this? Why do people share moral judgments but diverge on monetary awards? The best answer (developed by Kahneman, of course, though his coauthors were grateful to be along for the ride) is that the effort to 'map' moral judgments onto dollars is an exercise in 'scaling without a modulus'. 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, the agreement is substantially improved. And 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 inpunitive damage awards – and in all likelihood with other damage awards too – does not come from differences in social norms or in any relevant norms that govern punishment judgments. It comes from the 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.

Outrage and group deliberation

The findings thus far did not involve deliberating juries. They were based on the judgments of individuals placed, by computer, into small groups, with individual views being 'pooled' to create a verdict. The result was to create 'statistical juries' whose verdicts consisted of the view of the median juror, which seemed to provide a reasonable estimate of what the jury itself would do. But how does group deliberation affect outrage? In a subsequent study, involving about 3,000 people, we found that the median juror is not, in fact, a good predictor of the ultimate verdict of the jury. What we found does not falsify the findings just described; in a way, it reinforces them. But it also says a great deal about the effects of deliberation on moral outrage.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

The study tested the effects of deliberation on both punitive intentions and dollar judgments. The study involved about 3,000 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 zero to eight, where zero indicated that the defendant should not be punished at all and eight 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'. It would be reasonable to predict that the verdicts of juries would be the median of punishment judgments of jurors, but the prediction would be badly wrong.

Two findings are especially important. First, deliberation made the **lower** punishment ratings **decrease** when compared to the median of predeliberation judgments of individuals, while deliberation made the **higher** punishment 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. 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. When the median juror judgment was four or more on the eight-point scale, the jury's verdict was above the median predeliberation judgment of individuals.

Consider, for example, a case involving a man who nearly drowned on a yacht that was defectively constructed. People tended to be outraged by the idea of a defectively built yacht, and groups became far more outraged than their median members. But when the median juror judgment was less than four, the jury's verdict was below the median judgment of individuals. Consider a case involving a shopper who was injured in a fall when an escalator stopped suddenly. Individual jurors were not greatly bothered by the incident, seeing it as a genuine accident rather than a case of serious wrongdoing, and groups were more lenient than individuals.

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 case of high awards. For example, the median individual judgment, in the case involving the defective yacht, was $450,000, whereas the median jury judgment, in that same case, was $1,000,000. But awards shifted upwards for low awards as well. As extreme but actual illustrations of the severity shift, consider a few examples from the raw data:

– A jury whose predeliberation judgments were $200,000, $300,000, $2 million, $10 million, $10 million, and $10 million reached a verdict of $15 million.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

- A jury whose predeliberation judgments were $200,000, $500,000, $2 million, $5 million, and $10 million reached a verdict of $50 million.
- A jury whose predeliberation judgments were $2 million, $2 million, $2.5 million, $50 million, and $100 million reached a verdict of $100 million.

Notably, the degree of dispersion between individual predeliberation judgments did not contribute to greater or lesser shifts as a result of deliberation. In other words, juries whose members were in rough agreement (i.e., had a low standard deviation) about dollars or punishment did not show a different shift from groups whose members were in substantial disagreement about dollars or punishment. Whether they began in rough agreement or not, they showed the same severity shift for dollars, and the same leniency and severity shifts for punishment on a bounded scale.

Retribution, not optimal deterrence

Now let us return to the question of punishment. On the economic account, the state's goal, when imposing penalties, is to ensure optimal deterrence. 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.

In the context of punitive damages, all this leads to a simple theory: the purpose of such damages is to make up for the shortfall in enforcement. 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 suits should receive a punitive award that is twice the amount of the compensatory award. Simple exercises in multiplication will ensure optimal deterrence.

But there is a large question of whether social norms, or everyday morality, and the theory of optimal deterrence can fit together. Do people want optimal deterrence? Do they accept or reject the economic theory of punishment? If the outrage heuristic is at work, the answer to these questions will be a firm no. To find out, we conducted two experiments. In the first, we gave people cases of wrongdoing, arguably calling for punitive damages, and also provided people with explicit information about the probability of detection. Different people saw the same case, with only one difference: varying the probability of detection. People were asked about the amount of punitive damages that they would choose to award. Our goal was to see if people would impose higher punishments when the probability of detection was low.

In the second experiment, we 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. We wanted people to say whether they approved or disapproved of varying the penalty with the probability of detection.

Our findings were simple and straightforward. The first experiment found that varying the probability of detection did not affect punitive awards. Even when people's attention was explicitly directed to the probability of detection, people were indifferent to it. People's decisions about appropriate punishment were not influenced by seeing a

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

high or low probability of detection. The second experiment found that a strong majority of respondents rejected judicial decisions to reduce penalties because of the high probability of detection – and also rejected executive decisions to increase penalties because of the 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 – not optimal deterrence.

Here as elsewhere, outrage rules the roost. For ordinary people, punishment judgments are rooted in the outrage heuristic. As intuitive retributivists, people make judgments about the outrageousness of behavior, and those judgments are highly predictive of their intent to punish. When monetary judgments become noisy, it is because of the difficulty of scaling without a modulus – a problem that besets judgments in many domains of law and policy.

Outrage everywhere

When confronted with serious wrongdoing, people feel outraged. Their level of outrage operates a heuristic, which predicts people's punitive intentions. Those who use it are intuitive retributivists. They care about deterrence, of course, but where ideas about optimal deterrence diverge from retributive thinking, people will be inclined to reject those ideas. Placed in a deliberating group, people who begin with high levels of outrage will end up being more outraged than they were when they started to talk. Because outrage is based on relevant comparison sets and thus category-specific, people in the legal system might well produce patterns of outcomes that they would reject if only they were to see them.

These various psychological findings should be jarring, especially to those who favor utilitarian or welfarist approaches to law. To the extent that a system of criminal justice depends on the moral psychology of ordinary people, it will operate on the basis of the outrage heuristic, and will, from the utilitarian point of view, end up making serious and systematic errors.

Let me end on a personal note. Working with Danny, on these projects and others (Kahneman *et al.*, 2021), was an honor of a lifetime. He was the most creative person I have ever met. He was the most precise with language. He had the highest standards. And as he often said, he did not have sunk costs. He had no problem deciding that a long period of work had proved utterly useless and that beloved drafts and cherished chapters had to be tossed away. Actually he enjoyed that. He cared about one thing above all others: the truth.

But he also had a ton of fun. Starting a draft, fixing a draft, discarding a draft – all that was fun. For Danny, discussing a project might have been the most fun of all. On one occasion, I got immensely frustrated by the flood of ideas coming from Danny, and by my utter inability to capture them, so that they wouldn't be lost. Noticing my despair, Danny looked at me with amusement and said, 'Cass, you think by writing. I think by talking.' It was a blessing to get to talk with Danny.

Acknowledgements. Everything said here owes a great debt to Daniel Kahneman, though he is not, of course, responsible for my errors and omissions. I am grateful as well to Adam Oliver for valuable comments on a previous draft.

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press

References

Kahneman, D. and C. R. Sunstein (2005), 'Cognitive Psychology of Moral Intuitions', in J.-P. Changeux, A. R. Damasio, W. Singer and Y. Christen (eds), *Neurobiology of Human Values*, Heidelberg: Springer-Verlag, 91–105.

Kahneman, D., D. Schkade and C. R. Sunstein (1998), 'Shared outrage and erratic awards: the psychology of punitive damages', *Journal of Risk and Uncertainty*, **16**(1): 49–86.

Kahneman, D., O. Sibony and C. R. Sunstein (2021), *Noise: A Flaw in Human Judgment*. New York: Little, Brown.

Schkade, D., C. R. Sunstein and D. Kahneman (2000), 'Deliberating about dollars: the severity shift', *Columbia Law Review*, **100**(4): 1139–1175.

Sunstein, C. R., D. Kahneman and D. Schkade (1998), 'Assessing punitive damages (with notes on cognition and valuation in law)', *Yale Law Journal*, **107**(7): 2017–2153.

Sunstein, C. R., D. Schkade and D. Kahneman (2000), 'Do people want optimal deterrence?', *Journal of Legal Studies*, **29**(1): 237–253.

Sunstein, C. R., D. Kahneman, D. Schkade and I. Ritov (2002), 'Predictably incoherent judgments', *Stanford Law Review*, **54**(6): 1153–1215.

Cite this article: Sunstein, C.R. (2025), 'The outrage heuristic', *Behavioural Public Policy* **9**, 293–300. https://doi.org/10.1017/bpp.2024.33

https://doi.org/10.1017/bpp.2024.33 Published online by Cambridge University Press