

### 3. Spare the rod, spoil the child (Joseph Heath, Filthy Lucre MS)

“A liberal's idea of getting tough on crime,” Ronald Reagan once said, “is to give out longer suspended sentences.”<sup>1</sup> Not everyone thought this was funny, but everyone understood the difference in temperament that he was referring to. One of the most persistent divisions between the right and the left, conservatives and liberals, has to do with the use of punishment. Conservatives are eager to crack the whip. Bratty kids need to be spanked, rebellious teenagers sent to boot camp, drug addicts locked up in prison, criminals sentenced to hard time and terrorists hunted down and killed. Liberals, on the other hand, are always looking for an opportunity to coddle and appease those who break the rules. Bratty kids are just frustrated, they need someone to lend an understanding ear. Teenagers need more basketball courts and after-school programs. Drug addicts are ill, they need safe injection sites. Criminals need to be rehabilitated, not punished, and terrorists need more foreign aid.

Conservatives usually regard the liberal temperament as the result of either squeamishness or moral confusion. And truth be told, there is a fair bit of this going around. Consider, for example, the horrified reaction in the United States when it was discovered that “fun loving” American teenager, Michael Fay, had been sentenced in Singapore to be caned for his role in vandalizing cars with spray paint. The affair became an international incident when U.S. President Bill Clinton personally intervened in order to request clemency. Yet it is very far from obvious that a quick flogging is such a bad idea, compared to having offenders languish in adult prison for years, as is standard practice in the United States. It's certainly better than the death penalty – at least if you flog the wrong person, you can apologize afterwards. In practice, our squeamishness with regard to corporeal punishment often just leads us to invent more and more elaborate forms of psychological torture.<sup>2</sup>

But while many liberals take things a bit too far, there is one important problem with the conservative line. Punishment is not nearly as effective as most people are naturally disposed to believe. This is why people who are actually in the punishment business are often quite sympathetic to the liberal view – because they get to see, up close and personal, the futility of relying entirely upon punishment as a way of getting things done. Consider, for example, the so-called “faint hope” clause in Canada, which allows even mass murderers sentenced to life in prison a periodic opportunity to apply for parole. This generates paroxysms of anger every few years, when convicted child-killer Clifford Olson is granted a parole hearing. Even though he is always denied release, the mere fact that he is being considered is enough to leave the conservative wing fuming. Yet support for the faint-hope clause is strongest, not amongst bleeding-heart liberal sociologists who think that criminals are merely misunderstood, but with Corrections Canada, or more specifically, among prison wardens and guards. The problem is that without something positive to offer inmates – without a carrot to dangle in front of them – it is simply impossible to keep some of them under control. In many cases, no amount of coercion can replace what that one tiny, improbable ray of hope is able to achieve.

We tend to ignore this, however, because we are all subject to a cognitive fallacy, which creates the illusion that positive reinforcement isn't working, even when it is. The result is an extraordinarily common bias, which leads people to overestimate the effectiveness of punishment.<sup>3</sup> It is a consequence of a statistical phenomenon known as “regression to the mean,” or in less technical terms, the fact that an uncommon event is more likely to be followed by a common event than by another uncommon event. As a result, exceptionally bad behavior is likely to be followed by somewhat better behavior, just

---

1 A case of life imitating Hollywood: the line was cribbed from the satirical 1972 movie, *The Candidate*, starring Robert Redford.

2 A thesis most famously advanced by Michel Foucault, *Discipline and Punish* ().

3 Norman Miller; Donald C. Butler; James A. McMartin, “The Ineffectiveness of Punishment Power in Group Interaction,” *Sociometry*, 32 (1969): 24-42.

as exceptionally good behavior is likely to be followed by somewhat worse behavior – regardless of whether the bad behavior is punished or the good behavior rewarded. But a casual observer who ignores the underlying trend is likely to be fooled into thinking that the punishment worked, while the reward not only failed, but actually had the perverse effect of encouraging misbehavior. This is known as a “regression fallacy.” Unfortunately, it is a fallacy that can only be dispelled by looking at the long-term trend. This is why there is so much disagreement, when it comes to questions of punishment, between the views of experts, who actually study the long-term trends, and the verdicts of common sense.

## §

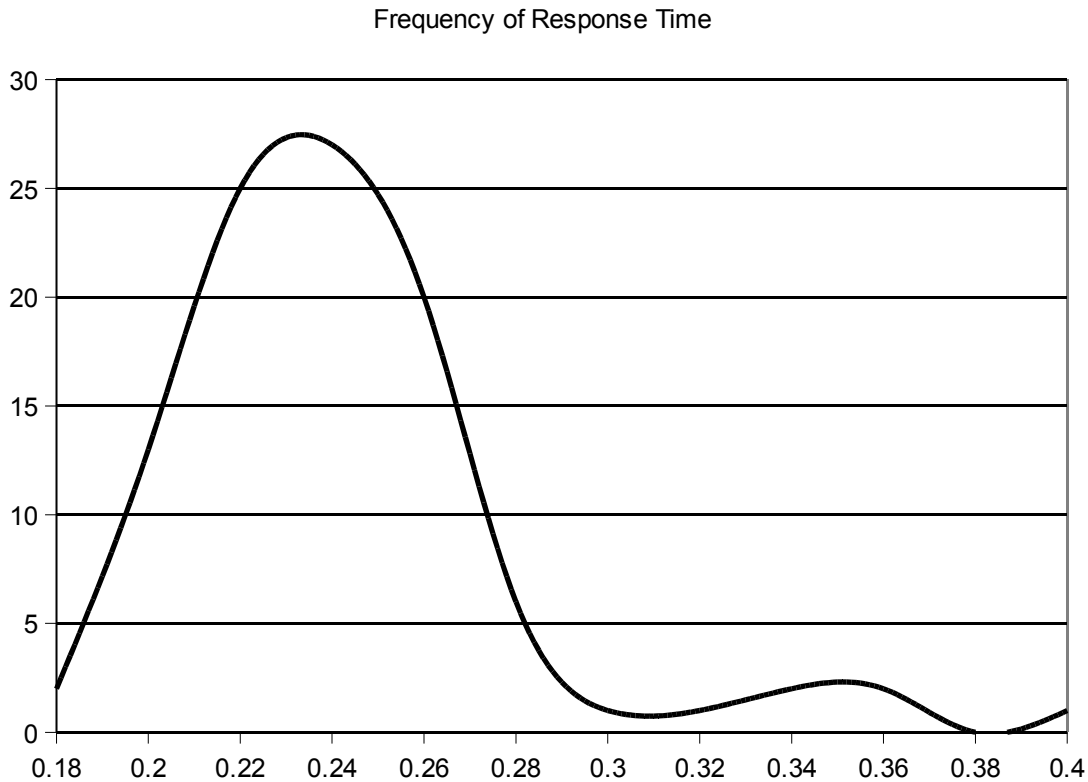
I think people who understand statistics are destined to be driven nuts by the rest of us. A colleague of mine once told me, in a state of extreme agitation, how upset she gets every morning on the subway, when forced to stare at an advertisement for a local community college boasting that 90% of their graduates find a job within a year of graduation. “It’s base rate neglect!” she cried. “Base rate neglect! 90% find *a* job. But the unemployment rate is only 5%. Does going to this college double your chances of being unemployed? Or do only losers go there?”

Regression fallacies are somewhat more subtle. Imagine a person trying to perform a moderately demanding task. For concreteness, let me use myself as an example. Like many people my age, I grew up playing “twitch” video games and never stopped. Unfortunately, now that I’m getting older (and thanks to the wonders of the internet), I find myself increasingly having to suffer the indignity of getting shot up by 14-year old hitscan freaks. So I thought I might submit myself to a reflex test, to see how degraded my skills have become with age. There are lots of these available on the internet, but I chose one that solicits demographic information from all users (and presents aggregate data on the thousands of people who have taken it, broken down by age, gender, and several other variables). I was relieved to discover that I’m still faster than most people in the 11-20 age cohort.

The test itself is rather simple. You stare at a traffic light. It starts on red, and after a variable delay, switches to green. As soon as it goes green, you click the mouse. It then reports your response time in milliseconds. To get a nice clean average, I did the test 100 times and kept track of my scores. If you plot my performance out on a graph, it takes something like the shape of the familiar “Bell curve” (Figure 3.1). The horizontal axis shows my response time, in milliseconds, while the vertical axis shows the numbers of times that a response of that speed occurred. My response was between 230 and 250 ms on 27 of the 100 tries, which is where the curve peaks. Responses between 210-230 were slightly less frequent (25), as were responses between 250-270 (20). Beyond that, I had 13 responses between 200-210, then about a dozen very fast or very slow reactions – with a few extremely slow ones, when I was asleep at the switch (this is why there is a “long tail” on the right-hand side).

What does this graph tell us? The first thing we learn is that, for me, a response between 200 and 270 is a very high-probability outcome on this test, since 85 of my responses occurred within this band. So if I were to take the test again, and you wanted to guess what my score would be, picking a number somewhere in this range would be the safest bet. My average (or mean) response time over the 100 tries was 241 ms, slightly higher than the most frequent (or modal) response time of 234, due to the fact that the distribution is slightly skewed to the right. That because there’s only so fast you can get, on this test, whereas there’s no limit to how slow you can be. Thus deviations in the “very slow” direction were much larger than deviations in the “very fast” direction.

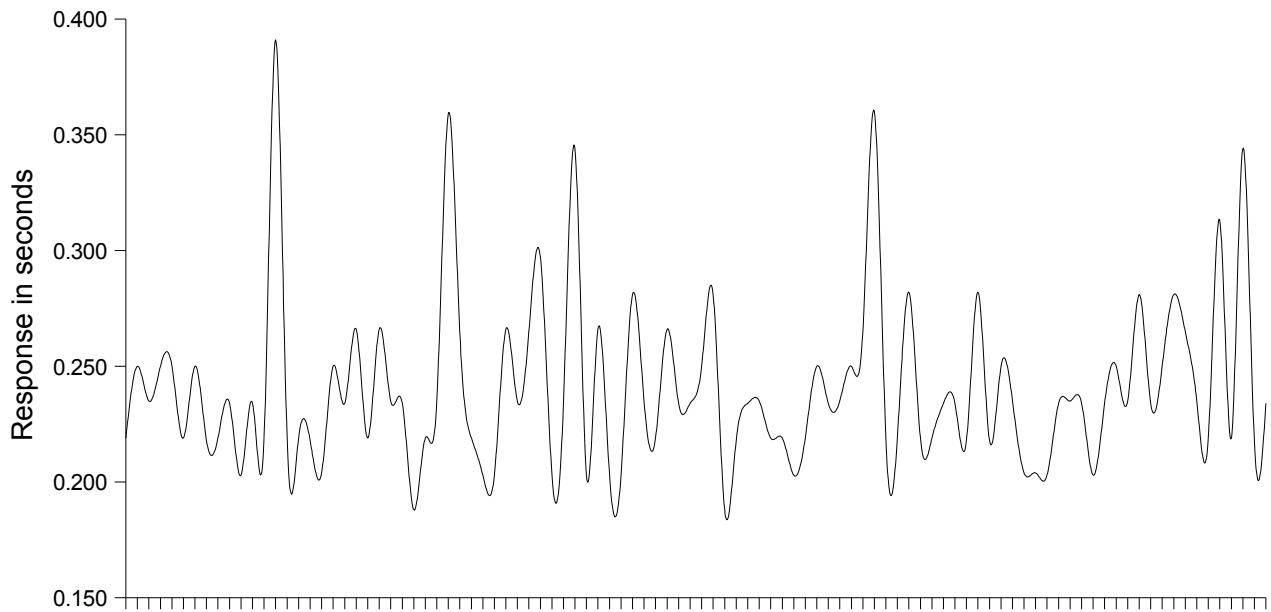
There’s one thing, however, that you can’t see by looking at this graph. This has to do with the pattern in my responses. Consider Figure 3.2, which shows my response time on the vertical axis, for each of my 100 tries (tracked on the horizontal axis). This graph shows what my scores were *in the sequence in which they were achieved*. Again, it looks exactly as one might expect it to look – a lot of responses clustering around the modal outcome of 234, with the occasional spike or dip, representing either a very fast or very slow response. These “outliers,” however, tend to be isolated occurrences.



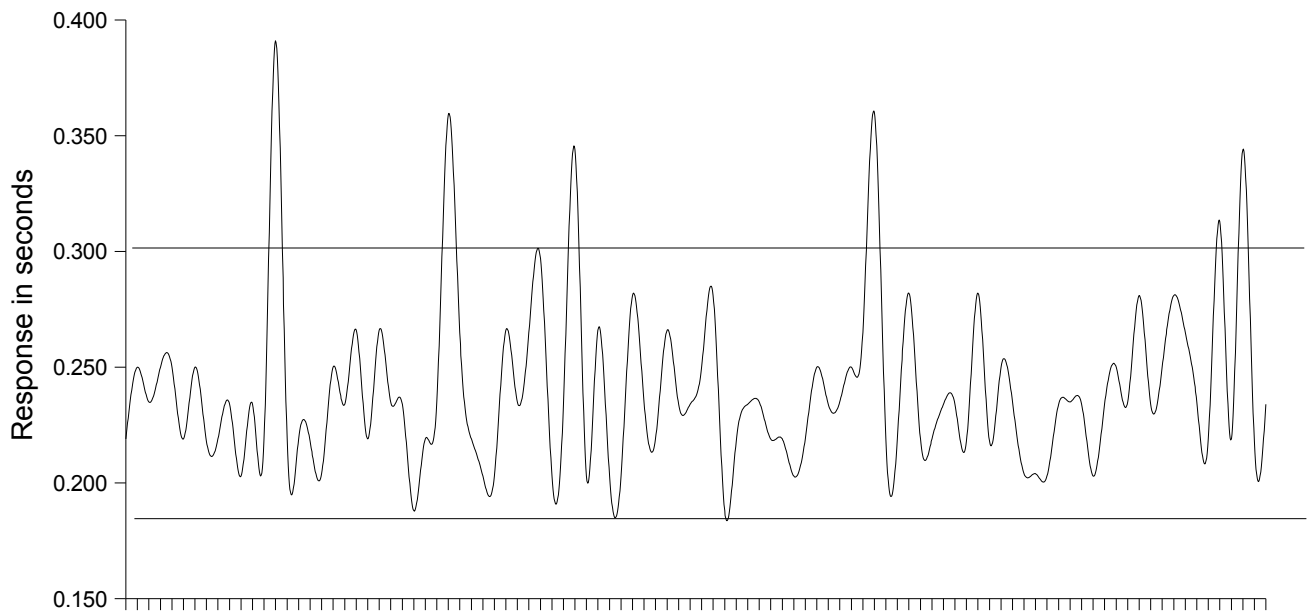
**Figure 3.1 Reaction time**

Not one of the very slow responses (above 270) is followed by another very slow response, and neither are any of the very fast responses (below 200). There is a very simple reason for this. As Figure 3.1 shows, the probability of my response being between 200 and 270 is about 85%. The probability of having a single response outside that range is 15%, but the probability of having two responses in a row outside that range is much lower. In fact, it's sufficiently improbable (2.25%), that it didn't occur even once in this particular sequence of 100 scores. This is why the very good and very bad scores take the form of isolated spikes.

The term "regression to the mean" is used to describe what happens immediately after the very bad or very good score is achieved. The follow-up tends to be closer to the average. This is not due to the "laws of psychology," but rather the "laws of probability." Low probability events are most likely to be followed by high probability events, rather than by more low-probability events, simply because the former are in general more likely to occur than the latter. "Common things are common," as the saying goes. In this example, bad performances were always followed by better ones – not because I felt chastened and resolved to do better, or because the terrible score helped to focus my mind, but simply because the bad performance was a significant deviation from my average performance, and was therefore unlikely to begin with. As a result, it was followed by an improvement, simply because the improvement was an independently more probable event (regardless of when it occurred). For the same reason, my occasional lightning-fast response was in each case followed by a slower one. Common things are common.



**Figure 3.2 Variation in response time**



**Figure 3.3 Introducing Punishment and Reward**

Now suppose that one were to introduce a little “punishment and reward” scheme into this test, in order to try to improve my performance. Suppose that every time it took me more than 300 ms to respond, I was subjected to a mild electric shock. And suppose that every time I responded in less than 200 ms, I was rewarded with some sort of snack. Suppose further that this punishment and reward system was completely useless (as it probably would be, since 20 years of video game playing is about

the longest course of operant conditioning for fast mouse clicking that can be devised). Or if you like, suppose that an observer is tricked (Milgram-style) into thinking that I am being punished and rewarded, when in fact I am not. The question is, what would it *look like* the effects of the punishments and rewards were?

Figure 3.3 shows my sequence of scores with an overlay representing the punishment and reward system. The upper line shows the threshold beyond which my behavior is punished, the lower line shows when my behavior is rewarded. In this series of scores, two of my responses would have been rewarded, and six punished. When one looks at the “incentive effects” of this system, it *looks* as though the punishments are working. After all, each time I am punished, my performance improves. With the rewards, on the other hand, not only does it look like they are not working, it looks as though they are having the opposite effect. Each time I am rewarded, my performance deteriorates, since the fast click is followed by a slower click. Of course, the idea that this has anything to do with the punishment and reward system is entirely an illusion. It’s regression to the mean that is doing all the work. Yet every intervention *seems* to be having an effect, since the subsequent result is quite different from the one that occasioned the reward or punishment.

The only way to tell whether the punishments and rewards are actually working is to look at the long-term trend, to see if my average performance improves. Again, casual observation is likely to be misleading. In Figure 3.3, since there were so many episodes of punishment followed by improvement, one could easily be misled into thinking that my average performance improved. The frequency of very slow response also appeared to decline. Yet in fact, my overall performance got slightly worse over the course of the 100 tries – I got slower on average, rather than faster. One can only discern this by subjecting the results to rather careful quantitative scrutiny (or in less fancy terms, “asking Excel to plot the trend line”).

Most of the time, however, we don’t collect long-term data on individual performance. This means that all we have to go on is our individual impressions. Yet because of regression to the mean, these individual impressions are almost always wrong.

## §

As a university teacher, I am intimately familiar with the business of punishment and reward. After all, I spend a fair number of working hours each year giving out grades, which are basically just sublimated rewards and punishments. In a university humanities course, the schema is pretty simple. Anyone who is at all serious about doing the work will pass, the question is simply what the grade will be. For the average student, a grade of “C” represents a punishment, “B” is neutral, and an “A” is a reward. It takes only a few encounters with crying students, devastated by the C that will prevent them from getting into medical school, to realize that what you’re really doing is punishing and rewarding, not all that different in kind from administering electrical shocks and handing out treats.

After a few years of giving out A’s, B’s and C’s, I started to notice a trend. Many of the students who got a C on their first assignment would show significant improvement over the course of the semester. Clearly they were chastened by the experience, and decided to “pull up their socks,” get more serious about their studies. On the other hand, many of the students who received an A on their first assignment would show marked deterioration. Obviously the A had encouraged them to slack off, or start coasting. Perhaps they were misled into thinking that the course was going to be easy, and that serious effort would not be required? Either way, it seemed clear that I wasn’t doing students any favors by giving them high grades on their first assignment.

As the years went by, I actually developed a fairly well-elaborated theory about the incentive effects of grades. The salutary effects of giving out low grades figured rather prominently in this theory. Unfortunately the theory was completely, 100% based upon a regression fallacy. I had simply ignored the fact that most students are B students, and that most of the work they turn out is B work. In fact, that’s just what B work is – work of the quality done by the majority of students. Of course, every so

often a typical B student will write an exceptionally bad or an exceptionally good piece of work. They will also tend to “return to form” shortly thereafter, and produce a piece that is merely average. So it is natural that one would see improvement among a certain percentage of C students, and deterioration among a certain percentage of A students, even if none of them had even looked at their grades.

It wasn't until I read about “regression to the mean” that I realized how thoroughly fooled I had been by the phenomenon. Unfortunately, I didn't learn about it from any teacher's manual, or even a treatise on social psychology. I learned about it from reading a dog training manual.<sup>4</sup> The authors were trying to explain why so many owners wind up beating their dogs, even though positive reinforcement is usually a more effective training strategy. The problem is that the owner's perception of what is working and what isn't working is usually false. The tendency to overestimate the power of punishment becomes especially tragic when it comes to dealing with submissive behavior. It's actually impossible to eliminate such behavior through negative reinforcement, since the punishment simply elicits more of the behavior that the trainer is seeking to extinguish.

What makes the fallacy especially insidious is the fact that, not only is the world organized in such a way as to generate the impression that punishment is more effective, it actually rewards us for thinking that it is. Again, this is something that I know well from my line of work. Like many teachers, I want students to do well (at least in part because, when they do well, it suggests that I am doing my job well). Thus I am disappointed when performance declines, and I am happy when students improve. As a result, giving out low grades tends to be followed by personal satisfaction, as students improve, while giving out high grades generates disappointment, as students regress to the mean. Daniel Kahneman summed up the situation perfectly, when he wrote that, “because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them.”<sup>5</sup> In other words, the world *conditions* us to prefer punishing others to rewarding them.

## §

It is important to distinguish the regression fallacy from the so-called gambler's fallacy. Regression fallacies occur when we ignore the fact that some outcomes are more likely than others. Gambler's fallacies occur when we imagine that some outcomes are more likely than others, when in fact they are not. If you want to see someone commit a gambler's fallacy, go to the corner store and buy one of those lottery tickets that allow you to pick your own numbers. Select an extremely “patterned” sequence of numbers, such as 1, 2, 3, 4, 5, 6. “Come on,” the clerk will say, “what are the chances of the numbers being drawn in a row like that?” To which you can then reply, “Exactly the same as the chances of any other sequence of numbers I pick being drawn in a row.”

When it comes to a lottery of this sort, each outcome is equally probable, and so all bets are equally good. For example, picking last week's winning number is no better and no worse than anything else you could pick. Similarly with a slot machine – a machine that hasn't paid out in a while is no more likely to pay out in the near future than any other. With craps, on the other hand, certain outcomes are more likely than others, simply because there are two dice being thrown, and so various ways in which certain sums can be produced. There are six different ways of getting a seven, but only two ways of getting an 11, making a seven three times more probable than an 11.<sup>6</sup> Thus an 11 is more likely to be followed by a seven than by another 11, simply because a seven is more probable than an 11 in the first place. As a result, a sequence of dice throws in a craps game would exhibit regression to the mean. It is important to recognize, however, that throwing an 11 does not make a seven any more probable, or an 11 any less probable. The regression to the more common outcome is not a causal relation, it is merely a description of a pattern that often occurs, in cases where the background

4 The Monks of New Skete, *The Art of Raising a Puppy* (New York: Little, Brown & Co., 1991).

5 Autobiographical sketch for Nobel Prize.

6 Ref. Huygens, *On Ratiocination in Dice Games* .

probabilities are not the same for all outcomes.

Thus it is important to be able to make predictions based upon regression to the mean without committing the gambler's fallacy. If someone flips a coin 10 times and gets 10 heads in a row, this does not make her any more or less likely to get a head on the next coin toss. Although the probability of flipping 11 heads in a row is very small, once you've flipped 10 in a row, the probability of flipping an 11<sup>th</sup> is still what it always was, 50%. Thus the fact that someone is on a streak does not increase the probability that the streak will end (or that it will continue). The same is true with regression to the mean. A craps player who has rolled ten 11s in a row is less likely to roll an 11 than a seven, but this has nothing to do with the fact that he is on a streak, it's just that the seven is and always has been more probable than the 11.

So we shouldn't get carried away with predicting regression. For instance, most actively managed mutual funds don't beat the market – they perform worse than index or other passively managed funds. And to the extent that fund performance is random (i.e. not the result of human intelligence), which is typically the case, a fund that has performed exceedingly well one year is likely to perform worse the following year. Thus it is not a particularly good investment strategy to buy funds on the basis of past performance. On the other hand, it is not a particularly bad strategy either. The mere fact that a fund manager got lucky one year does not make him any less likely to get lucky the following year. Although he is unlikely to repeat his own past year's performance, he is still just as likely to beat the market as he ever was. To the extent that individual consumers have no basis for choosing one fund over another, they might as well choose based on past performance, since this amounts to a randomizing strategy. But in order to avoid disappointment, they should form their *expectations* by looking at the average performance of similar funds, not the past performance of the fund that they purchase.

## §

We've seen that people in general, and economists in particular, have a general tendency to overestimate the importance of "extrinsic" motives, or external incentives, in guiding individual conduct. This cognitive bias is often compounded by a regression fallacy, which leads them to overestimate the effectiveness of punishment as a means of providing such incentives. The result can be an extremely caustic moral temperament, one that often shades over into cruelty.

There is, for example, a fairly well-developed conservative thesis (articulated quite forcefully by Robert Kagan just prior to the American invasion of Iraq), which claims that, when push comes to shove, it is only the threat of military force that provides stability in the global order.<sup>7</sup> Europeans, according to this view, were simply deluding themselves in thinking that diplomacy and "constructive engagement" could play any useful role in foreign policy. According to Kagan, this delusion was made possible only because European diplomacy (or "soft power") was tacitly backed by American military force (or "hard power"). Of course, the consequences of this "sticks, not carrots" approach to foreign policy – in Iraq and elsewhere – were eminently predictable, and were in fact predicted by pretty much everyone who cared to think in sober terms about international relations. Carrots, as it turns out, are not just the poor man's stick, they are in many cases superior instruments of foreign policy. That's why the European Union has been so much more successful, in the past twenty years, at promoting democracy abroad (through offers of EU membership) than the United States.

Of course, it is important not to go too far in the opposite direction, and to imagine that punishment and coercion are entirely dispensable as mechanisms of social control. The sort of squeamishness about punishment that conservatives decry is a genuine phenomenon in our culture. It is all well and good to point out that two-thirds of people voluntarily cooperate in experimental collective action games, even when they have the opportunity to free-ride with impunity. This does suggest that

---

<sup>7</sup> Robert Kagan, *Of Paradise and Power* (New York: Knopf, 2003).

threats and punishment are not nearly as important as they are often cracked up to be. Yet it is important to remember that another one-third on average don't cooperate. In repeated games, the presence of these defectors leads to a rapid erosion of cooperation, simply because those who began the game cooperating become less likely to do so, as they witness the spectacle of others free-riding with impunity. "If he's not going to play nice, why should I?" they say. The only way to sustain cooperation in these games is to establish some sort of "correlation" in the strategies played by the players, which means that the cooperators must interact more often with other cooperators *and* the defectors must interact more often with other defectors. In practice, this means that cooperators must be given the option of punishing defectors, by banishing them or otherwise depriving them of access to the benefits of cooperation.

So while access to the benefits of cooperation (the "carrot" in these games) is enough to motivate most people to cooperate, it is not sufficient. Unless players are also given a "stick" to apply against those who don't behave, the system of cooperation quickly unravels. In such cases, punishment is important, not only because it helps motivate the defectors to behave better, but because it allows the cooperators to act morally without having to fear the depredations of the defectors. There is in this respect an important asymmetry between morality and immorality. Defection is corrosive, in a way that cooperation is not (which is why civilizations sometimes collapse into barbarism, but never the reverse.) Punishment is necessary, in order to create the protected space within which morality can flourish, precisely because it prevents the exploitation of the moral by the immoral.

Thus there is nothing "illiberal" or right-wing about adopting a punitive stance, when it comes to controlling genuine social deviance. The hard line taken by the Blair government in Great Britain toward "anti-social behavior" is a case in point. Anyone who has spent some time in that country knows that petty vandalism, harassment and general yobbishness in public places are a real problem. While the direct consequences of this behavior are not sufficiently harmful to warrant criminalization, it is the indirect consequences – the general insecurity that it promotes – that serve as the proper object of public concern. Having teenagers jumping up and down on the hood of your car in the middle of the night, whenever you park it on the street, is enough to make anyone want to move. Even if they aren't doing any damage, you should be able to call the police to make them stop.

Ironically, the most important social function of punishment often has less to do with the incentives it provides than with the signal it sends. Punishment is a way of communicating the message that certain forms of behavior are not merely undesirable, but *categorically prohibited*. It does so by making clear to the perpetrator that "society" is not just going to discourage you from acting a certain way, but is going to forcibly prevent you from doing so. Hence the value of anti-social behavior orders (or "ASBOs") in Great Britain (which allow police and magistrates to order specific individuals to desist from specific nuisance activities, with criminal sanctions for subsequent violations). In order to have a society committed to the public good, it is essentially that people feel comfortable and secure in public spaces. Yet despite this basic truism, and despite the overwhelming popularity of ASBOs in Great Britain, many people on the left remain extremely uncomfortable with any sort of prohibition of the constant, low-level harassment that often goes on in urban areas.

So where does that leave us? It's easy to overestimate the effectiveness of punishment. It's also easy to shy away from punishment, based upon a hypertrophied discomfort with coercion. The correct approach presumably lies somewhere in between. Incentives are complicated, and figuring out how people will respond to them is even more so. Social control is an art, not a science. The best we can hope for is a healthy pragmatism, combined with a willingness to try, but also to discard, any item in the "toolkit" of institutional strategies available to us.