McChesney, F. 1997. *Money For Nothing: Politicians, Rent Extraction, and Political Extortion*. Cambridge, MA: Harvard University Press.
McCormick, R. and Tollison, R. 1981. *Politicians, Legislation and the Economy*. Boston: M. Nijhoff.
Olson, M. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven: Yale University Press.
Peltzman, S. 1976. Toward a more general theory of regulation. *Journal of Law and Economics* 19: 211–40.
Posner, R. 1974. Theories of economic regulation. *Bell Journal of Economics and Management* 5: 335–58.
Posner, R. 1982. Economics, politics, and the reading of statutes and the constitution. *University of Chicago Law Review* 49: 263–65.
Riker, W. and Weingast, B. 1988. Constitutional regulation of legislative choice: the political consequences of judicial deference to legislatures. *Virginia Law Review* 74: 373–401.
Robinson, G. 1988. Public choice speculations on the line-item veto. *Virginia Law Review* 74: 403–17.
Rossiter, C. 1960. *The American Presidency*. New York: Harcourt, Brace & World.
Tollison, R. 1988. Public choice and legislation. *Virginia Law Review* 74: 339–43.
Weingast, B. and Marshall, W. 1988. The industrial organization of Congress; or, why legislatures, like firms, are not organized as markets. *Journal of Political Economy* 96: 132–63.

public enforcement of law. In this essay we consider the theory of the public enforcement of law – the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. Of course, private parties also play an important role in law enforcement, by providing information to public authorities and also by initiating their own legal actions (notably, tort suits), but to maintain focus we restrict attention here to public enforcement activity. (On private enforcement, see generally Landes and Posner 1987 and Shavell 1987c; and on private versus public enforcement, see Becker and Stigler 1974, Landes and Posner 1975, and Polinsky 1980a.)

1. OUTLINE OF ESSAY. In sections 2 through 4, we present the basic elements of the theory of public enforcement of law. Our concern is with the probability of imposition of sanctions, the magnitude and form of sanctions, and the rule of liability. In sections 5 through 16 we then examine a variety of extensions of the central theory, including accidental harms, costs of imposing fines, mistake, marginal deterrence, settlement, self-reporting, repeat offences, and incapacitation.

Before proceeding, we note that economically-oriented analysis of public law enforcement dates from the eighteenth century contributions of Montesquieu (1748), Beccaria (1764), and, especially, Bentham (1789), whose analysis of deterrence was sophisticated and expansive. But, curiously, after Bentham (1789), the subject of enforcement lay essentially dormant in economic scholarship until the late 1960s, when Gary Becker (1968) published a highly influential article, which has led to a voluminous literature.

2. THE BASIC FRAMEWORK. An individual (or a firm) chooses whether to commit an act that for simplicity is assumed to cause harm with certainty (see section 5 on uncertain harm). If he commits the act, he obtains some gain, and also faces the risk of being caught, found liable, and sanctioned. The rule of liability could be either strict – under which the individual is definitely sanctioned – or fault-based – under which he is sanctioned only if his behaviour fell below a fault standard. The sanction that he suffers could be a monetary fine, a prison term, or a combination of the two.

Whether an individual commits a harmful act is determined by an expected utility calculation. He will commit the act if that would raise his expected utility, taking into account the gain he would derive and the probability, form, and level of sanction that he would then face. We will usually first examine the assumption that individuals are risk neutral with respect to sanctions, that is, that they treat an uncertain sanction as equivalent to its expected value; but we will also consider alternative assumptions.

Social welfare is generally presumed to equal the sum of individuals' expected utilities. An individual's expected utility depends on whether he commits a harmful act, on whether he is a victim of someone else's harmful act, and on his tax payment, which will reflect the costs of law enforcement, less any fine revenue collected. If individuals are risk neutral, social welfare can be expressed simply as the gains which individuals obtain from committing their acts, less the harms caused, and less the costs of law enforcement. (The assumption that individuals' gains are always credited in social welfare could be relaxed without affecting most of our conclusions. The principal difference that altering the assumption would make is that more acts would be treated as socially undesirable and that optimal sanctions and enforcement effort would increase.)

We assume, as is conventional, that fines are socially costless to employ because they are mere transfers of money (but in section 6 we consider fines that are costly to impose), whereas imprisonment involves positive social costs because of the expense associated with the operation of prisons and the disutility due to imprisonment (which is not naturally balanced by gains to others).

The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures, or, equivalently, a probability of detection, the level of sanctions and their form (a fine, prison term, or combination), and the rule of liability (strict or fault-based).

3. OPTIMAL ENFORCEMENT GIVEN THE PROBABILITY OF DETECTION. We consider here optimal enforcement given the assumption that the probability of detection is fixed (the probability will be treated as a policy instrument in the next section). Thus, we ask about the optimal form and level of sanctions under strict and fault-based liability, and about how the two liability rules compare.

Strict liability. Assume initially that fines are the form of sanction and that individuals are risk neutral. Then the optimal fine is the harm h divided by the probability of detection p, that is, h/p; for then the expected fine equals the harm (observe that $p(h/p) = h$). If, for example, the

harm is \$1,000 and the probability of detection is 25%, the optimal fine is \$4,000, and the expected fine is \$1,000. This fine is optimal because, when the expected fine equals the harm, an individual will commit a harmful act if, and only if, the gain he would derive from it exceeds the harm he would cause. Essentially this basic and fundamental formula was noted by Bentham (1789: 173) and it has been observed by many others since.

If individuals are risk averse with regard to fines, one would expect the optimal fine to be lower than in the risk-neutral case for two reasons. First, this reduces the bearing of risk by individuals who commit the harmful act. Second, because risk-averse individuals are more easily deterred than risk-neutral individuals, the fine does not need to be as high as before to achieve any desired degree of deterrence. (However, for subtle reasons that we will not address here, the optimal fine could, in principle, be higher when individuals are risk averse.)

Next assume that imprisonment is the form of sanction, with social costs incurred in imposing sanctions. In this case, there is not a simple formula for the optimal imprisonment term; see Polinsky and Shavell (1984). The optimal term could be such that there is either underdeterrence or overdeterrence, compared to socially ideal behaviour. On one hand, a relatively low imprisonment term, implying underdeterrence, might be socially desirable because it means that imprisonment costs are reduced for those individuals who commit harmful acts. On the other hand, a relatively high term, implying overdeterrence, might be socially desirable because it means that imprisonment costs are reduced due to fewer individuals committing harmful acts, even if some of these deterred individuals would have obtained gains exceeding the harm. (The possible optimality of overdeterrence, however, strikes us as more theoretical than real.)

Now consider the combined use of fines and imprisonment. Here, the main point is that fines should be employed to the maximum extent feasible before resort is made to imprisonment. In other words, it is not optimal to impose a positive imprisonment term unless the fine is maximal. (The maximal fine might be interpreted as the wealth of an individual.) The rationale for this conclusion is that fines are socially costless to impose, whereas imprisonment is socially costly, so deterrence should be achieved through the cheaper form of sanction first. This point is noted by Bentham (1789: 183) and Becker (1968: 193); see also Polinsky and Shavell (1984). To amplify, suppose that the fine f is less than the maximal fine f_m and that a positive prison term t is employed. Raise f toward f_m and lower t so as to keep the disutility of the combined sanctions constant. Then deterrence and the amount of harm will not change, but the cost of imposing the imprisonment sanction will fall, raising social welfare. Hence, it must be optimal for the fine to be maximal before imprisonment is used. (Observe that this argument holds regardless of individuals' attitudes toward risk of either fines or imprisonment.)

Fault-based liability. Assume again that fines are the form of liability. Then the same formula for the fine that we said was optimal under strict liability – namely, h/p, the harm divided by the probability of detection – will lead to compliance with the fault standard (assuming that the fault standard is optimally selected). For example, suppose that the harm is \$1,000, and that an individual would be found at fault if he failed to take a precaution costing less than \$1,000 that could have prevented the harm. Furthermore, suppose such a precaution exists that costs \$700. If the probability of detection is 25%, then a fine of \$4,000 for being at fault would induce the individual to take a precaution costing less than \$1,000; in particular, he would prefer to spend \$700 on the precaution and escape liability than not to take the precaution and bear an expected fine of \$1,000.

Observe that the optimal fine of \$4,000 in the preceding example is not unique. Any higher fine also would induce compliance with the fault standard, as would some lower fines (as long as the expected value of the fine exceeds \$700, or whatever is the expense of the precaution that costs less than \$1,000). Note that higher fines do not lead to a problem of overdeterrence because parties can escape fines by complying with the fault standard. (However, if mistakes occur in the legal process, parties might be induced to spend excessively on precautions; see section 8 below.)

If individuals are risk averse, they are more easily deterred than if they are risk neutral, so the fine does not need to be as high to induce compliance with the fault standard. Moreover, assuming that compliance occurs, no one actually is sanctioned because no one is found at fault (assuming, as we do here, that there are no mistakes). Thus, fault-based liability has the attractive feature that it can accomplish desired deterrence of harm-creating conduct without imposing risk on risk-averse individuals (Shavell 1982).

Next, consider imprisonment as the sanction; see Shavell (1987a). Here, for essentially the reasons given in the case of fines, any sanction above a threshold level will ensure compliance with the fault standard, and the minimum sanction necessary to induce compliance is higher the lower is the probability of detection. Also, fault-based liability again can accomplish deterrence without the actual imposition of sanctions, which would be socially costly (Shavell 1985).

Finally, consider the joint use of fines and imprisonment. In this case, it does not matter what the combination of sanctions is, provided that the sanctions achieve compliance with the fault standard. In particular, it is not advantageous for society to employ maximal fines before resorting to imprisonment because compliance means that sanctions are never imposed.

Comparison of liability rules. Because sanctions are not imposed under fault-based liability (in the absence of mistakes), this form of liability has an advantage over strict liability when the sanction is the fine and individuals are risk-averse, or when the sanction is imprisonment. However, fault-based liability is more difficult to administer. Namely, to apply fault-based liability, the enforcement authority must have more information than under strict liability: it must be able to calculate optimal behaviour to determine the fault standard and it must ascertain whether the fault standard was met. Under strict liability, the

authority need only ascertain harm. (Moreover, for reasons we discuss in section 7 below, strict liability encourages better decisions by injurers regarding their level of participation in harm-creating activities.)

4. OPTIMAL ENFORCEMENT INCLUDING THE PROBABILITY OF DETECTION. We now consider the optimal system of enforcement when expenditures on enforcement, and hence the probability of detection, are allowed to vary. Consideration of this issue originated with Becker (1968); the early writers on enforcement (including Bentham 1789) did not examine the issue of the choice of enforcement effort.

Strict liability. Assume first that the sanction is a fine and that individuals are risk neutral. Then the optimal level of the fine is maximal and the optimal probability is low (in a sense to be described). The basic explanation for this conclusion is that if the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. Suppose, for example, that the fine initially is \$4,000 and that the probability of detection is 25%. Now raise the fine to \$10,000, presuming that the maximal fine is at least this high, and lower the probability of detection to 10%. Then the expected fine remains equal to \$1,000, so that deterrence is maintained, but expenditures on enforcement are significantly reduced, implying that social welfare rises. This process can be continued, and social welfare augmented, as long as the fine is below the maximal level f_m. Becker (1968) suggested this result (although much of his analysis implicitly presumes that the fine is not maximal); Carr-Hill and Stern (1979: 280–309) and Polinsky and Shavell (1979) note it explicitly.

The optimal probability is low in that there is some underdeterrence; that is, the optimal p is such that the expected fine pf_m is less than the harm h (Polinsky and Shavell 1984). The reason for this result is that if pf_m equals h, behaviour will be ideal, meaning that the individuals who are just deterred obtain gains essentially equal to the harm. These are the individuals who would be led to commit the harmful act if p were lowered slightly. That in turn must be socially beneficial because these individuals cause no net social losses (their gains essentially equal the harm), but reducing p saves enforcement costs. How much pf_m should be lowered below h depends on the saving in enforcement costs from reducing p compared to the net social costs of underdeterrence that will result if p is lowered non-trivially.

If individuals are risk averse, the optimal fine is generally less than maximal, as first shown in Polinsky and Shavell (1979) (and elaborated upon in Kaplow 1992). This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit the harmful act. Another more particular explanation involves reconsidering the argument that we used in the risk-neutral case. If the fine f is less than f_m, it is still true that f can be raised and p lowered so as to maintain deterrence, but because of risk aversion, this implies that pf falls, meaning that fine revenue falls. The reduction in fine revenue reflects the disutility caused by imposing greater risk on risk-averse individuals. If individuals are sufficiently risk averse, the decline in fine revenue associated with greater risk-bearing could more than offset the savings in enforcement expenditures from reducing the probability of detection, implying that social welfare would be lower.

Next, assume that the sanction is imprisonment and that individuals are risk neutral in imprisonment, that is, the disutility of imprisonment is the same for each additional year. (We did not discuss individuals' attitudes toward the risk of imprisonment in section 3 because the points we made there did not depend on this consideration.) Then the optimal imprisonment term is maximal (Shavell 1991). The reasoning behind this result parallels that used to show that the optimal fine is maximal when individuals are risk neutral in fines. Specifically, if the imprisonment term is raised and the probability of detection lowered so as to keep the expected sanction constant, neither individual behaviour nor the costs of imposing imprisonment are affected (by construction, the expected prison term is the same), but enforcement expenditures fall.

Suppose instead that individuals are risk averse in imprisonment. In other words, the disutility of each year of imprisonment grows with the number of years in prison, perhaps because imprisonment becomes increasingly difficult to tolerate. In this case there is a stronger argument for setting the imprisonment sanction maximally than when individuals are risk neutral (Polinsky and Shavell 1997). This is because, when the imprisonment term is raised, the probability of detection can be lowered even more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but also the social costs of imposing imprisonment sanctions decline because the expected prison term falls.

Last, suppose that individuals are risk preferring in imprisonment, that is, the disutility of each year of imprisonment falls with the number of years in prison. This assumption seems particularly important: the first years of imprisonment may create special disutility, due to brutalization of the prisoner, or due to the stigma effect of having been imprisoned at all. Additionally, the fact that individuals discount the future disutility of imprisonment makes earlier years of imprisonment more important than later ones. If individuals are risk preferring in imprisonment, the optimal sanction may well be less than maximal (Polinsky and Shavell 1997). In particular, the type of argument used above does not necessarily apply. When the sanction is raised, the probability that maintains deterrence cannot be lowered proportionally, implying that the expected prison term rises. Because the resulting increased cost of imposing imprisonment sanctions might exceed the savings in enforcement expenditures from lowering the probability, the optimal prison term might not be maximal.

When the probability of detection is set optimally, together with the imprisonment term, underdeterrence may well result, not only to save enforcement expenditures, but also to reduce the costs of imposing imprisonment sanctions. (In theory, however, overdeterrence could be optimal for the reason mentioned in section 3 – that overdeterrence could reduce the costs of imprisonment.)

Now consider the situation when both fines and imprisonment are employed as sanctions. Recall that under the optimal enforcement policy, the fine must be maximal, for otherwise it cannot be desirable to employ imprisonment. The main point we wish to make is that, unlike when imprisonment is used alone, the optimal imprisonment term may not be maximal even if individuals are risk neutral or risk averse in imprisonment. Suppose that individuals are risk neutral in imprisonment and fines. Then if the imprisonment term is raised and the probability of detection is lowered so as to keep the expected imprisonment term constant, deterrence declines because the expected fine falls (due to the reduction in the probability). Hence, to maintain deterrence, the probability cannot fall proportionally. But this implies that the expected prison term, and the costs of imposing imprisonment, are higher than previously. Only if the savings in enforcement costs are sufficiently large, therefore, is it socially desirable to raise the imprisonment sanction.

Fault-based liability. The least expensive way to accomplish compliance with the fault standard is to use the highest possible sanction and, given this sanction, the lowest probability of detection that deters individuals who would be at fault. The reason is that, if all individuals who would be at fault are deterred, the only cost incurred is associated with the setting of the probability; this cost is minimized by using the maximal sanction and a correspondingly low probability. Note that this is true regardless of whether the sanction is a fine or imprisonment and regardless of individuals' attitudes toward the risk of fines or of imprisonment.

Comparison of rules. As we emphasized earlier, under fault-based liability sanctions are not actually imposed (in the absence of mistakes), which is an advantage over strict liability when the sanction is a fine and individuals are risk averse, and is always an advantage when the sanction is imprisonment. Moreover, this advantage of fault-based liability implies a second advantage: it may allow a further savings in enforcement expenditures over that under strict liability. For example, suppose that the sanction is a fine and that injurers are risk averse. Then, as we have emphasized, the optimal fine under strict liability is generally not maximal, due to risk bearing by injurers. This implies that the probability of detection needed to achieve any given level of deterrence is higher than if the fine were maximal. Under fault-based liability, however, the optimal fine is maximal despite the risk aversion of injurers (because the fine is not actually imposed), meaning that a lower probability can be employed. However, to decide which liability rule is preferable, these advantages of fault-based liability would have to be weighed against the disadvantages of this rule relative to strict liability that we mentioned at the end of section 3 (one of which, as noted, will be discussed in section 7).

This concludes the presentation of the basic theory of public enforcement of law. We now turn to various extensions and refinements of the analysis.

5. ACCIDENTAL HARMS. As we noted at the outset, we assumed that individuals decide whether or not to commit acts that cause harm with certainty, that is, they decide whether or not to cause intentional harms. In many circumstances, however, harms are accidental – they occur only with a probability. For instance, if a driver speeds, he only creates a likelihood of a collision; or if a firm stores toxic chemicals in a substandard tank, the firm only creates the probability of a harmful spill.

Essentially all that we have said above applies in a straightforward way when harms are accidental. If individuals are risk neutral, sanctions are monetary, and the expected sanction equals harm, then induced behaviour will be socially optimal; further, the optimal magnitude of sanctions is maximal if individuals are risk neutral because this allows enforcement costs to be saved; and so forth. Our general conclusions from above can thus be interpreted to apply both when harms are intentional and cause harm for sure, as well as when actions alter the risk of harm.

There is, however, an additional issue that arises when harm is uncertain: a sanction can be imposed either on the basis of the commission of a dangerous *act* that increases the chance of harm – storing chemicals in a substandard tank – or on the basis of the *actual occurrence of harm* – only if the tank ruptures and results in a spill. In principle, either approach can achieve optimal deterrence. To illustrate, suppose that the substandard tank has a 10% chance of rupturing, in which case the harm would be $10 million; the expected harm from using the tank therefore is $1 million. If injurers are risk neutral and sanctions are imposed only when harm occurs, deterrence will be optimal if, as usual, the expected sanction equals the harm of $10 million. Alternatively, if sanctions are imposed on the basis of the dangerous act of using the substandard tank, deterrence will be optimal if the owner of the tank faces an expected sanction equal to the expected harm due to his use of the substandard tank, $1 million.

Several factors are relevant to the choice between act-based and harm-based sanctions (Shavell 1993). First, act-based sanctions need not be as high to accomplish a given level of deterrence, and thus offer an underlying advantage over harm-based sanctions because of limitations in parties' assets. In the example in the preceding paragraph, the owner of the storage tank might be able to pay the $1 million required if sanctions are act-based but not the $10 million required if sanctions are harm-based. Second, and closely related, because act-based sanctions need not be as high to accomplish deterrence, they offer an advantage over harm-based sanctions when parties are risk averse. Third, act-based sanctions and harm-based sanctions may differ in the ease with which they can be applied. In some circumstances, act-based sanctions may be simpler to impose (it might be less difficult to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil into the ocean); in other circumstances, harm-based sanctions may be easier (a driver who causes harm might be caught without difficulty, but not one who speeds). Fourth, it may be hard to calculate the expected harm due to an act, but relatively easy to ascertain the actual harm if it eventuates; if so, this constitutes an advantage of harm-based liability.

6. COSTS OF IMPOSING FINES. We inquire in this section about the implications of costs borne by enforcement authorities in imposing fines. Our principal observation is that such costs should raise the level of the fine.

We assume for simplicity that the probability of detection is fixed, that liability is strict, and that individuals are risk neutral. In this setting, recall from section 3 that the optimal fine is h/p, the harm divided by the probability of detection.

Now let there be a public cost k of imposing a fine. The optimal fine then becomes $h/p + k$; the cost k should be added to the fine that would otherwise be desirable (Becker 1968:192 and Polinsky and Shavell 1992). The intuition behind this result is that, if an individual commits a harmful act, he causes society to bear not only the immediate harm h, but also, with probability p, the cost k of imposing the fine – that is, his act results in an expected total social cost of $h + pk$. If the fine is $h/p + k$, the individual's expected fine is $p[(h/p) + k] = h + pk$, leading him to commit the harmful act if and only if his gain exceeds the expected total social cost of his act.

For example, suppose the harm is \$1,000, the probability of detection is 25%, and the cost of imposing the fine is \$500. An individual's harmful act causes society to bear expected total social costs of \$1,125 – the harm of \$1,000 plus a 25% chance of the \$500 cost of imposing a fine. If the fine is set equal to \$4,500, the individual's expected fine also will equal \$1,125.

Additionally, observe that, not only does the state bear costs when fines are imposed, so do individuals who pay the fines (such as legal defence expenses). The costs borne by individuals, however, do not affect the formula for the optimal fine. Individuals properly take these costs into account, because they bear them.

7. LEVEL OF ACTIVITY. We have been assuming that the sole decision that an individual makes is whether to act in a way that causes harm when engaging in some activity. In many contexts, however, an individual also makes a choice about his *activity level* – that is, not only does he choose whether to commit a harmful act when engaging in an activity, he also chooses whether to engage in that activity, or, more generally, at what level to do so. For example, besides deciding how to behave when driving (whether to exercise care in changing lanes), an individual also chooses how many miles to drive; the number of miles driven is the individual's level of activity. Similarly, not only does a firm decide how to conduct its operations during production (whether to pollute), it also chooses its level of production; the output of the firm is its level of activity.

The socially optimal activity level is such that the individual's marginal utility from the activity just equals the marginal expected harm caused by the activity. Thus, the optimal number of miles driven is the level at which the marginal utility of driving an extra mile just equals the marginal expected harm per mile driven. The determination of the optimal level of activity presumes that individuals act optimally when engaging in the activity – for example, that they drive with appropriate care. Thus, in determining the optimal level of activity, an individual's marginal utility from the activity is compared to the expected harm that results when he acts appropriately when engaging in the activity. Often this expected harm is positive even when a party acts optimally, because the cost of precautions (or, equivalently, the gain or savings from not taking precautions) exceeds the expected reduction in harm that taking the precautions would bring about.

Will parties' choices about their activity levels be socially correct under the two major forms of liability? The answer is that under strict liability, their choices about activity levels will be correct, but under fault-based liability, generally they will participate in activities to a socially excessive extent. Under strict liability, parties will choose the optimal level of activity because they will pay for all harm done. They will choose the optimal number of miles to drive because they will pay for all harm per mile driven. Under fault-based liability, however, parties generally do not pay for the harm they cause because, as we have discussed, they will be induced to behave so as not to be found at fault. Consequently, when deciding on their level of activity, they will choose an excessive level. They will not take into account the harm that each additional mile of driving causes, and therefore they will drive too much.

The interpretation of the preceding points in relation to firms is that under strict liability, the product price will reflect the expected harm caused by production, so that the price will reflect the full social cost of production. Hence, the amount purchased, and thus the level of production, will tend to be socially optimal. However, under fault-based liability, the product price will not reflect harm, but only the cost of precautions; thus, the amount sold, and the level of production, will be excessive.

A related comment is that safety regulations and other regulatory requirements are often framed as standards of care that have to be met, but which, if met, free the regulated party from liability. Hence, regulations of this character are subject to the criticism that they lead to excessive levels of the regulated activity. Making parties strictly liable for harm would be superior to safety regulation with respect to inducing socially correct activity levels.

The tendency of parties to choose an excessive level of activity under fault-based liability, but not under strict liability, constitutes a fundamental advantage of strict liability; it was first emphasized in Shavell (1980) and Polinsky (1980b). This advantage, note, is stronger the greater is the harm engendered by engaging in the activity (given that behaviour is optimal when engaging in the activity). Thus, for activities for which expected harm is likely to be substantial, the disadvantage of fault-based liability will be significant.

8. MISTAKES. Errors of the two classic types can occur in public enforcement of law. First, an individual who should be found liable might mistakenly not be found liable – a Type I error. Second, an individual who should not be found liable might mistakenly be found liable – a Type II error. For an individual who has been detected, let the probabilities of these errors be ϵ_1 and ϵ_2, respectively.

Given the probability of detection p and the chances of Type I and Type II errors, an individual will commit the wrongful act if and only if his gain g net of his expected

fine if he does commit it exceeds his expected fine if he does not commit it, namely, when $g - p(1 - \epsilon_1)f > -p\epsilon_2 f$, or, equivalently, when $g > (1 - \epsilon_1 - \epsilon_2)pf$.

The first point to note is that, as emphasized in P'ng (1986), *both* types of error reduce deterrence: the term $(1 - \epsilon_1 - \epsilon_2)pf$ is declining in both ϵ_1 and ϵ_2. The first type of error diminishes deterrence because it lowers the expected fine if an individual violates the law. The second type of error, when an individual is mistakenly found liable, also lowers deterrence because it reduces the difference between the expected fine from violating the law and not violating it. In other words, the greater is ϵ_2, the smaller the increase in the expected fine if one violates the law, making a violation less costly to the individual.

Because mistakes dilute deterrence, they may reduce social welfare (see generally Kaplow and Shavell 1994a). Specifically, to achieve any level of deterrence, the probability p may have to be higher to offset the effect of errors. It should also be noted that, were one to take into account an individual's decision whether to engage in an activity (like driving), Type II errors have the additional effect of discouraging socially desirable participation in the activity.

Now consider the optimal choice of the fine. Given any probability of detection, the dilution in deterrence caused by errors requires a higher fine to restore deterrence. If the probability and the fine are variable, then, as explained in section 4, the optimal fine is maximal. Somewhat surprisingly, the optimal fine remains maximal despite mistakes, by essentially the argument used previously: If the fine f were less than maximal, then f could be raised and the probability p lowered so as to keep deterrence constant, but saving enforcement costs.

Now consider the possible risk aversion of individuals. As we emphasized in section 4, the optimal fine under strict liability is generally less than maximal when individuals are risk averse, because lowering the fine from the maximum level reduces the bearing of risk. Introducing the possibility of mistakes may increase the desirability of lowering the fine because, due to Type II errors, individuals who do not violate the law are subject to the risk of having to pay a fine (Block and Sidak 1980). Indeed, because the number of persons who do not violate the law often would far exceed the number who do, the desire to avoid imposing risk on the former group can lead to a substantial reduction in the optimal fine.

Next, consider imprisonment and mistake. In this regard, our conclusions parallel those with respect to fines in several respects. Notably, mistakes of both types dilute the deterrent effects of imprisonment; and the optimal imprisonment term is maximal if individuals are risk neutral or risk averse in imprisonment but is generally not maximal if they are risk preferring in imprisonment.

We have not yet commented on fault-based liability. Here, an important implication of mistake is that some individuals will bear sanctions even if they comply with the fault standard. Hence, the results regarding the optimal probability of detection and the sanction under strict liability apply to some extent under fault-based liability as well. Moreover, as stressed by Craswell and Calfee (1986), individuals will often have a motive to take excessive precautions in order to reduce the chance of being found erroneously at fault.

Finally, observe that, although we have treated the probabilities of error as fixed, they can be influenced by procedural choices. For example, prosecutorial resources can be increased in order to reduce the probability of a Type I error, or the standard of proof can be raised to reduce the chance of a Type II error (although this presumably increases Type I errors). Because the reduction of both types of error increases deterrence, expenditures made to reduce errors may be socially beneficial (Kaplow and Shavell 1994a).

9. GENERAL ENFORCEMENT. In many settings, enforcement may be said to be *general* in the sense that several different types of violations will be detected by an enforcement agent's activity. For example, a police officer waiting at the roadside may notice a driver who litters as well as one who goes through a red light or who speeds, and a tax auditor may detect a variety of infractions when he examines a tax return. To investigate such situations, suppose that a single probability of detection applies uniformly to all harmful acts, regardless of the magnitude of the harm. (The contrasting assumption is that enforcement is *specific*, meaning that the probability is chosen independently for each type of harmful act.)

The main point that we want to make is that in contexts in which enforcement is general, the optimal sanction rises with the severity of the harm and is maximal only for relatively high harms; this point was first made in Shavell (1991) (Mookherjee and Png 1992 is closely related). To see this, assume that liability is strict, the sanction is a fine, and injurers are risk neutral. Let $f(h)$ be the fine given harm h. Then, for any given general probability of detection p, the optimal fine schedule is h/p, provided that h/p is feasible; otherwise – for high h (all h such that $h/p > f_m$) – the optimal fine is maximal. This schedule is obviously optimal given p because it implies that the expected fine equals harm, thereby inducing ideal behaviour, whenever that is possible.

The question remains whether it would be desirable to lower p and raise fines to the maximal level for the low-harm acts for which h/p is less than maximal. The answer is that if p is reduced for the relatively low-harm acts (and the fine raised for them), then p – being general – is *also* reduced for the high-harm acts for which the fine is already maximal, resulting in lower deterrence of these acts. The decline in deterrence of high-harm acts may cause a greater social loss than the savings in enforcement costs from lowering p. To express this point differently, p must be sufficiently high to avoid significant underdeterrence of high-harm acts (for which fines are maximal). But since this p also applies to less harmful acts, the fines for them do not need to be maximal in order to deter them appropriately.

The result that, when enforcement is general, sanctions should rise with the severity of harm up to a maximum also holds if the sanction is imprisonment and if liability is fault-based. The underlying reasoning is the same as that given above.

10. MARGINAL DETERRENCE. In many circumstances, a person may consider which of *several* harmful acts to commit, for example, whether to release only a small amount of a pollutant into a river or a large amount, or whether to kidnap a person or also to kill the kidnap victim. In such contexts, the threat of sanctions plays a role in addition to the usual one of attempting to deter individuals from committing harmful acts: for individuals who are not deterred, sanctions influence which harmful acts individuals choose to commit. Notably, such individuals will have a reason to commit less harmful rather than more harmful acts if expected sanctions rise with harm. Deterrence of a more harmful act because its sanction exceeds that for a less harmful act is sometimes referred to as *marginal deterrence* (it apparently was so named by Stigler 1970).

Other things being equal, as observed by Beccaria (1764: 32) and Bentham (1789: 171), it is socially desirable that enforcement policy creates marginal deterrence, so that those who are not deterred from committing harmful acts have a reason to moderate the amount of harm that they cause. This suggests that sanctions should rise with the magnitude of harm and, therefore, that sanctions should not generally be maximal. However, fostering marginal deterrence may conflict with achieving deterrence generally: in order for the schedule of sanctions to rise steeply enough to accomplish marginal deterrence, sanctions for less harmful acts may have to be so low that individuals are not deterred from committing some harmful act.

We have two additional observations to make about marginal deterrence. First, marginal deterrence can be promoted by increasing the probability of detection as well as the magnitude of sanctions. For example, kidnappers can be better deterred from killing their victims if more police resources are devoted to apprehending kidnappers who murder their victims than to those who do not. (But in circumstances in which enforcement is general, the probability of detection cannot be independently altered for acts that cause different degrees of harm.) Second, marginal deterrence is naturally accomplished if the expected sanction equals harm for all levels of harm; for if a person is paying for harm done, he will have to pay appropriately more if he does greater harm. Thus, for instance, if a polluter's expected fine would rise from $100 to $500 if he dumps five gallons instead of one gallon of waste into a lake, where each gallon causes $100 of harm, his marginal incentives to pollute will be correct. For formal analyses of marginal deterrence, see Shavell (1992), Wilde (1992), and Mookherjee and Png (1994).

11. PRINCIPAL-AGENT RELATIONSHIP. Although we have assumed that an injurer is a single actor, injurers often are more appropriately characterized as collective entities, and specifically as a principal and the principal's agent. For example, the principal could be a firm and the agent an employee; or the principal could be a contractor and the agent a subcontractor.

When harm is caused by the behaviour of principals and agents, many of the conclusions of our prior analysis are not fundamentally altered; they simply carry over to the sanctioning of principals. Notably, if a risk-neutral principal faces an expected fine equal to harm done, he will in effect be in the same position *vis-à-vis* his agent as society is *vis-à-vis* a single potential violator of law (see Newman and Wright 1990 on a closely related point). Consequently, the principal will behave socially optimally in controlling his agents, and in particular will contract with them and monitor them in ways that will give the agents socially appropriate incentives to reduce harm (but see Arlen 1994). Another result that carries over to the principal-agent context is that it will often be desirable for society to tolerate some degree of underdeterrence in order to conserve enforcement resources.

A question about enforcement that arises when there are principals and agents is the allocation of financial sanctions between the two parties. It is apparent, however, that the particular allocation of sanctions does not matter when, as would be the natural presumption, the parties can reallocate the sanctions through their own contract. For example, if the agent finds that he faces a large fine but is more risk averse than the principal, the principal can assume it; conversely, if the fine is imposed on the principal, he will retain it and not impose an internal sanction on the agent. Thus, the post-contract sanctions that the agent bears are not affected by the particular division of sanctions initially selected by the enforcement authority.

The allocation of monetary sanctions between principals and agents would matter, however, if some allocations allow the pair to reduce their total burden. An important example is when a fine is imposed only on the agent and he is unable to pay it because his assets are less than the fine; see Sykes (1981) and Kornhauser (1982). Then, he and the principal (who often would have higher assets) would jointly escape part of the fine, diluting deterrence. The fine therefore should be imposed on the principal rather than on the agent (or at least the part of the fine that the agent cannot pay).

A closely related point is that the imposition of imprisonment sanctions on agents may be desirable when their assets are less than the harm that they can cause, even if the principal's assets are sufficient to pay the optimal fine; see Polinsky and Shavell (1993). The fact that an agent's assets are limited means that the principal may be unable to control him adequately through use of contractually-determined penalties, which can only be monetary. For example, a firm may not be able, despite the threat of salary reduction or dismissal, to induce its employees never to rig bids. In such circumstances, it may be socially valuable to use the threat of personal criminal liability and a jail sentence to better control agents' misconduct.

12. SETTLEMENTS. Although we have thus far assumed that when a liable injurer is discovered he is sanctioned in some automatic fashion, in practice he must be found civilly or criminally liable in a trial, and before this occurs, he commonly settles in lieu of trial. (In the criminal context, the settlement usually takes the form of a *plea bargain*, an agreement in which the injurer pleads guilty to a reduced charge.) Given the prevalence of settlements, it is important to consider how they affect deterrence and the optimal system of public enforcement, and whether settlements are socially desirable.

There are two general reasons why parties might prefer an out-of-court settlement to a trial (see generally the survey by Cooter and Rubinfeld 1989; and for some recent discussions of plea bargaining, see, for example, Reinganum 1988 and Miceli 1996). First, a trial is costly to the parties in terms of time and/or money. Second, settlements eliminate the risks inherent in the trial outcome, a benefit to parties who are averse to such risks. These advantages of settlement to the parties suggest that settlement is socially valuable, but the effect of settlement on deterrence is a complicating factor.

Specifically, settlements dilute deterrence: for if injurers desire to settle, it must be because the expected disutility of sanctions is lowered for them (see generally Polinsky and Rubinfeld 1988). The state may be able to offset this settlement-related reduction in deterrence by increasing the level of sanctions; if so, settlements need not compromise the overall level of deterrence.

Settlements may have other socially undesirable consequences. First, they may result in sanctions that are not as well tailored to harmful acts as would be true of court-determined sanctions. For example, if injurers have private information about the harm that they have caused, settlements will tend to reflect the average harm caused, whereas trial outcomes may better approximate the actual harm. Thus, the distribution of sanctions meted out through settlements will not be as good in terms of inducing injurers to take proper precautions (high-harm injurers will be underdeterred, and vice versa). Second, settlements hinder the amplification and development of the law through the setting of precedents, a factor of occasional relevance; and settlements also sometimes allow defendants to keep aspects of their behaviour secret, which can reduce deterrence. Third, settlements for prison terms can result in increases in public expenditures on jail if defendants are risk averse in imprisonment. On the social desirability of settlement, see, for example, Shavell (1997) and Spier (1997).

13. SELF-REPORTING. We have assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations to enforcement authorities. For example, firms often report violations of environmental and safety regulations, individuals usually notify police of their involvement in traffic accidents, and even criminals occasionally turn themselves in.

We explain here why it is generally socially desirable for the structure of enforcement to be such as to encourage self-reporting; see Kaplow and Shavell (1994b). Self-reporting can be induced by lowering the sanction for individuals who disclose their own infractions. Moreover, the reward for self-reporting can be made small enough that deterrence is only negligibly reduced.

To amplify, assume for simplicity that the sanction is a fine f, that the probability of detection is p, and that individuals are risk neutral. If an individual commits a violation and does not self-report, his expected fine is pf. Suppose the fine if an individual self-reports is set just below pf, say at $pf - \epsilon$, where $\epsilon > 0$ is arbitrarily small. Then the individual will want to self-report but the deterrent effect of the sanction will be (approximately) the same as if he did not self-report. For example, suppose that the fine if an individual does not self report is \$1,000 and that the probability of detection is 10%, so the expected fine is \$100. If the fine for self-reporting is slightly below \$100, such as \$95, or even \$99, individuals will self-report but deterrence will barely be reduced.

Given that self-reporting can be induced essentially without compromising deterrence, why is self-reporting socially advantageous? There are two basic reasons. First, self-reporting reduces enforcement costs because, when it occurs, the enforcement authority does not have to identify and prove who the violator was. If a company reports that it caused pollution, environmental enforcers do not need to spend as much effort trying to detect pollution and establishing its source. Second, self-reporting reduces risk, and thus is advantageous if injurers are risk averse. Drivers bear less risk because they know that if they cause an accident, they can (and will be led to) report this to the police and suffer a lower and certain sanction, rather than face a substantially higher sanction (for hit and run driving) imposed only with some probability.

14. REPEAT OFFENDERS. In practice, the law often sanctions repeat offenders more severely than first-time offenders. We explain here why such a policy may be socially desirable.

Note first that sanctioning repeat offenders more severely cannot be socially advantageous if deterrence always induces first-best behaviour. If the sanction for polluting and causing a \$1,000 harm is \$1,000, then any person who pollutes and pays \$1,000 is a person whose gain from polluting (say the savings from not installing pollution control equipment) must have exceeded \$1,000. Social welfare therefore is higher as a result of his polluting. If such an individual polluted and was sanctioned in the past, that only means that it was socially desirable for him to have polluted previously. Raising the current sanction because of his having a record of sanctions would overdeter him now.

Accordingly, only if deterrence is inadequate is it possibly desirable to condition sanctions on offence history to increase deterrence. But deterrence will often be inadequate because, as we emphasized in section 4, it will usually be worthwhile for the state to tolerate some underdeterrence in order to reduce enforcement expenses.

Given that there is underdeterrence, making sanctions depend on offence history may be beneficial for two reasons. First, as developed in Polinsky and Shavell (1996), the use of offence history may create an additional incentive not to violate the law: if getting caught violating the law implies not only an immediate sanction, but also a higher sanction for any future violation, an individual will be more deterred from committing a violation presently. Second, as studied, for example, in Rubinstein (1979) and Polinsky and Rubinfeld (1991), making sanctions depend on offence history allows society to take advantage of information about the dangerousness of individuals and the need to deter them: individuals with offence histories may be more likely than average to commit future violations, which might make it desirable for purposes of deterrence to impose higher sanctions on them.

There is also an incapacitation-based (see section 16) reason for making sanctions depend on offence history. Repeat offenders are more likely to have higher propensities to commit violations in the future and thus more likely to be worth incapacitating by imprisonment.

15. IMPERFECT KNOWLEDGE ABOUT THE PROBABILITY AND MAGNITUDE OF SANCTIONS. Although we have implicitly assumed that injurers know the probability and magnitude of sanctions, this is not always the case. Individuals frequently have imperfect knowledge of these variables; that is, they have estimates, or more likely, subjective probability distributions, describing the true probability of a sanction and its true magnitude. They might not know the true probability of a sanction for several reasons: because the enforcement authority refrains from publishing information about the probability (perhaps hoping that individuals will believe it to be higher than it is in fact); because the probability is variable, depending on factors that individuals do not fully understand (the probability of a tax audit, for example, is influenced by a large number of considerations); and because probabilities seem to be difficult for individuals to assess. Also, individuals may have incomplete knowledge of the true magnitude of sanctions, particularly if the magnitudes of sanctions are not fixed by law, but are to some degree discretionary.

The implications of injurers' imperfect knowledge are straightforward. First, to predict how individuals behave, what is relevant, of course, is not the actual probability and magnitude of a sanction, but the perceived levels or distributions of these variables.

Second, to determine the optimal probability and magnitude of a sanction, account must be taken of the relationship between the actual and the perceived variables (on which, see, for instance, Bebchuk and Kaplow 1992, and see Kaplow 1990 on learning about sanctions). For example, suppose that there is a delay of at least a year before individuals fully comprehend a change in the probability of enforcement. Then if enforcement resources are increased so as to make the probability, say, 15% rather than 10%, there might not be a significant increase in deterrence for some time, making such an investment less worthwhile. Or, for instance, suppose that the sanction for some act, such as robbery, is variable (say from one month of jail time to ten years), and that individuals' perceptions are quite rough, not based on true averages but more on the possible range. Then increasing the average sentence might have very little effect on deterrence, but increasing the probability substantially might still augment deterrence. The psychology and learning process (see Sah 1991) by which individuals assimilate and formulate perceived probabilities of sanctions and their magnitude are important, therefore, to determining how deterrence works and what optimal policy is.

16. INCAPACITATION. Our discussion of public enforcement has presumed that the threat of sanctions reduces harm by discouraging individuals from causing harm – that is, by deterring them. However, an entirely different way for society to reduce harm is by imposing sanctions that remove parties from positions in which they are able to cause harm – that is, by *incapacitating* them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their driver's licences, preventing them from doing harm while driving; businesses can lose their right to operate in certain domains, and the like. We focus here on imprisonment, but what we say applies to incapacitative sanctions generally; on the economic theory of incapacitation, see Shavell (1987b).

To better understand the role of public enforcement when sanctions are incapacitative, suppose that the sole function of sanctions is to incapacitate; that is, sanctions do not deter. (Deterrence might not occur if, for instance, individuals' gains from harmful acts exceed the expected sanctions, given the relevant range of the probabilities and magnitudes of the sanctions.) We assume that the social goal is as before, to maximize gains from acts less harm, and less the costs of enforcement and sanctions, including the costs of keeping individuals in prison.

The optimal sanction to impose on an individual who is apprehended is determined by comparing the expected harm, net of gains, he would cause if not in prison to the private and public costs of imprisonment. If the expected net harm exceeds the costs of imprisonment, he should be put in prison and kept there as long as this condition holds. Thus, the optimal sanction as a function of expected net harm is zero up to a threshold – the point at which expected net harm equals the costs of imprisonment – and then rises discontinuously to the length of time during which the person's net expected harm exceeds imprisonment costs. Jail should only be used to incapacitate individuals whose net harm is relatively high.

Two points about the incapacitative rationale are worth making. First, there is evidence that suggests that the expected harm caused by individuals declines with their age. Thus, from the incapacitative standpoint, individuals should be released from jail when their dangerousness falls below the cost of incapacitation. Second, as a matter of logic, the incapacitative rationale might imply that a person should be put in jail even if he has not committed a crime – because his danger to society makes incapacitating him worthwhile. This would be true, for example, if there were some means to predict accurately a person's dangerousness independently of his actual behaviour. In practice, however, the fact that a person has committed a harmful act may be the best basis for predicting his future behaviour, in which case the incapacitation rationale would suggest imposing a jail term only if the individual has committed an especially harmful act.

Last, we comment on the relationship between optimal enforcement when incapacitation is the goal versus when deterrence is the goal. First, when incapacitation is the goal, the optimal magnitude of the sanction is independent of the probability of apprehension. In contrast, when deterrence is the goal, the optimal sanction depends on the probability – the sanction generally is higher the lower is the probability. Second, when incapacitation is the goal, the sanction rises discontinuously with the magnitude of harm (from zero to a positive amount), but when deterrence is the goal the sanction rises continuously with harm. Third, when deterrence is the goal, the probability and magnitude of sanctions depend on the ability to deter, and

if this ability is limited (as, for instance, with the enraged), a low expected sanction may be optimal. However, a high expected sanction still might be called for to incapacitate.

17. CONCLUSION. It may be worthwhile summarizing the main points that we have made in this essay.

(a) When the probability of detection is for some reason fixed, the benchmark level of the optimal fine is the harm divided by the probability of detection, for this results in an expected fine equal to the harm. However, costs incurred by the public in collecting the fine should be added to the benchmark level of the fine, and risk aversion of injurers should usually lower the level of the fine.

(b) When the probability of detection may be varied, high sanctions may be optimal, for this allows a relatively low probability to be employed and thereby saves enforcement costs. Indeed, the optimal fine is in principle maximal for this reason if individuals are risk neutral in wealth, and the optimal imprisonment term is maximal if individuals are risk neutral or risk averse in imprisonment. More realistically, optimal sanctions are not maximal when individuals are risk averse in wealth or risk preferring in imprisonment (as well as for the general enforcement reason discussed in section 9), although the motive to set sanctions at relatively high levels in order to reduce enforcement costs still applies.

(c) Optimal enforcement tends to be characterized by some degree of underdeterrence relative to first-best behaviour, because this conserves enforcement resources. More precisely, by lowering the probability of detection from a level that would lead to first-best behaviour, the state reduces enforcement costs, and although more individuals commit the harmful act, these individuals do not cause social welfare to decline substantially because their gains are nearly equal to the harm.

(d) The use of fines should be exhausted before resort is made to costlier sanctions, notably imprisonment.

(e) An advantage of fault-based liability over strict liability is that costly sanctions – fines when individuals are risk averse, and imprisonment – are imposed less often: Under fault-based liability, injurers generally are induced (in the absence of mistakes) to obey fault standards, and therefore ordinarily do not bear sanctions. Under strict liability, however, injurers are sanctioned whenever they are caught.

(f) An advantage of strict liability over fault-based liability is that it is easier to apply. Another advantage is that injurers' activity-level decisions will generally be better: Under strict liability, injurers' activity levels will tend to be optimal because injurers will pay for harm that they cause. But under fault-based liability, their activity levels will tend to be excessive because generally they will not pay for harm that they cause (because they will be led to behave without fault).

These and other basic conclusions about the public enforcement of law that we have described in this essay are now, for the most part, well-established in the economic literature on enforcement. However, an important aspect of this subject that has received relatively little attention is the behaviour of public enforcers themselves. Because our focus was on deriving socially optimal public enforcement policies, we assumed, as is conventional, that public enforcers act so as to maximize social welfare. A logical step to take next would be to re-examine optimal enforcement policies when public enforcers behave in a self-interested manner. Although some effort has already been devoted to this inquiry, much remains to be done.

A. MITCHELL POLINSKY AND STEVEN SHAVELL

See also ACCURACY IN ADJUDICATION; BECKER, GARY S.; BENTHAM, JEREMY; CAPITAL PUNISHMENT; CRIME AS A DISTINCT CATEGORY OF BEHAVIOUR; CRIMINAL ATTEMPTS; CRIMINAL CONVICTION AND FUTURE INCOME; CRIMINAL JUSTICE; ECONOMIC APPROACH TO CRIME AND PUNISHMENT; PLEA-BARGAINING: A COMPARATIVE APPROACH; PUNITIVE DAMAGES; SETTLEMENT OF LITIGATION; THIRD-PARTY LIABILITY; VICARIOUS LIABILITY.

Subject classification: 5e.

BIBLIOGRAPHY

Arlen, J.A. 1994. The potentially perverse effects of corporate criminal liability. *Journal of Legal Studies* 23: 833–67.

Bebchuk, L.A. and Kaplow, L. 1992. Optimal sanctions when individuals are imperfectly informed about the probability of apprehension. *Journal of Legal Studies* 21: 365–70.

Beccaria, C. 1764. *An Essay on Crimes and Punishments*. Trans. and repr. Albany: W.C. Little, 1872.

Becker, G.S. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76: 169–217.

Becker, G.S. and Stigler, G.J. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Bentham, J. 1789. *An introduction to the principles of morals and legislation*. In *The Utilitarians*, Garden City, NY: Anchor Books, 1973.

Block, M.K. and Sidak, J.G. 1980. The cost of antitrust deterrence: why not hang a price fixer now and then? *Georgetown Law Journal* 68: 1131–9.

Carr-Hill, R.A. and Stern, N. 1979. *Crime, the Police and Criminal Statistics: an Analysis of Official Statistics for England and Wales using Econometric Methods*. London: Academic Press.

Cooter, R.D. and Rubinfeld, D.L. 1989. Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* 27: 1067–97.

Craswell, R. and Calfee, J.E. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, and Organization* 2: 279–303.

Kaplow, L. 1990. Optimal deterrence, uninformed individuals, and acquiring information about whether acts are subject to sanctions. *Journal of Law, Economics, and Organization* 6: 93–128.

Kaplow, L. 1992. The optimal probability and magnitude of fines for acts that definitely are undesirable. *International Review of Law and Economics* 12: 3–11.

Kaplow, L. and Shavell, S. 1994a. Accuracy in the determination of liability. *Journal of Law and Economics* 37: 1–15.

Kaplow, L. and Shavell, S. 1994b. Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy* 102: 583–606.

Kornhauser, L.A. 1982. An economic analysis of the choice between enterprise and personal liability for accidents. *California Law Review* 70: 1345–92.

Landes, W.M. and Posner, R.A. 1975. The private enforcement of law. *Journal of Legal Studies* 4: 1–46.

Landes, W.M. and Posner, R.A. 1987. *The Economic Structure of Tort Law*. Cambridge, MA: Harvard University Press.

Miceli, T. 1996. Plea bargaining and deterrence: an institutional approach. *European Journal of Law and Economics* 3: 249–64.

Montesquieu, C. 1748. *The Spirit of the Laws*. Berkeley: University of California Press, 1977.

Mookherjee, D. and P'ng, I.P.L. 1992. Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* 82: 556–65.
Mookherjee, D. and P'ng, I.P.L. 1994. Marginal deterrence in enforcement of law. *Journal of Political Economy* 102: 1039–1066.
Newman, H.A. and Wright, D.W. 1990. Strict liability in a principal-agent model. *International Review of Law and Economics* 10: 219–31.
Png, I.P.L. 1986. Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* 6: 101–105.
Polinsky, A.M. 1980a. Private versus public enforcement of fines. *Journal of Legal Studies* 9: 105–27.
Polinsky, A.M. 1980b. Strict liability vs. negligence in a market setting. *American Economic Review* 70: 363–70.
Polinsky, A.M. and Rubinfeld, D.L. 1988. The deterrent effects of settlements and trials. *International Review of Law and Economics* 8: 109–116.
Polinsky, A.M. and Rubinfeld, D.L. 1991. A model of optimal fines for repeat offenders. *Journal of Public Economics* 46: 291–306.
Polinsky, A.M. and Shavell, S. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69: 880–91.
Polinsky, A.M. and Shavell, S. 1984. The optimal use of fines and imprisonment. *Journal of Public Economics* 24: 89–99.
Polinsky, A.M. and Shavell, S. 1992. Enforcement costs and the optimal magnitude and probability of fines. *Journal of Law and Economics* 35: 133–48.
Polinsky, A.M. and Shavell, S. 1993. Should employees be subject to fines and imprisonment given the existence of corporate liability? *International Review of Law and Economics* 13: 239–57.
Polinsky, A.M. and Shavell, S. 1996. Repeat offenders and the theory of deterrence. Harvard Law School, Discussion Paper No. 188, John M. Olin Center for Law, Economics, and Business; Stanford Law School, Working Paper No. 134, John M. Olin Program in Law and Economics.
Polinsky, A.M. and Shavell, S. 1997. On the disutility and discounting of imprisonment and the theory of deterrence. Harvard Law School, Discussion Paper No. 213, John M. Olin Center for Law, Economics, and Business; Stanford Law School, Working Paper No. 146, John M. Olin Program in Law and Economics. (*Journal of Legal Studies*, forthcoming, 1999.)
Reinganum, J.F. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78: 713–28.
Rubinstein, A. 1979. An optimal conviction policy for offenses that may have been committed by accident. In *Applied Game Theory*, ed. S.J. Brams, A. Schotter and G. Schwodiauer (Wurzburg: Physica-Verlag, 1979): 406–13.
Sah, R.K. 1991. Social osmosis and patterns of crime. *Journal of Political Economy* 99: 1272–95.
Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9: 1–25.
Shavell, S. 1982. On liability and insurance. *Bell Journal of Economics* 13: 120–32.
Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85: 1232–62.
Shavell, S. 1987a. The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* 77: 584–92.
Shavell, S. 1987b. A model of optimal incapacitation. *American Economic Review, Papers and Proceedings* 77: 107–110.
Shavell, S. 1987c. *Economic Analysis of Accident Law*. Cambridge, MA: Harvard University Press.
Shavell, S. 1991. Specific versus general enforcement of law. *Journal of Political Economy* 99: 1088–1108.
Shavell, S. 1992. A note on marginal deterrence. *International Review of Law and Economics* 12: 345–55.
Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36: 255–87.
Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26: 575–612.
Spier, K. 1997. A note on the divergence between the private and the social motive to settle under a negligence rule. *Journal of Legal Studies* 26: 613–21.
Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78: 526–36.
Sykes, A.O. 1981. An efficiency analysis of vicarious liability under the law of agency. *Yale Law Journal* 91: 168–206.
Wilde, L.L. 1992. Criminal choice, nonmonetary sanctions, and marginal deterrence: a normative analysis. *International Review of Law and Economics* 12: 333–44.

public franchising. Public franchising covers all situations where government agencies make use of private-sector businesses to provide services commonly supplied in many countries directly by the public sector. These include contracting out of services within the public sector (e.g., the provision of catering services within the National Health Service and the subcontracting of local-authority refuse collection in the UK) and the transfer of more substantial parts (possibly all) of a nationalized industry to the private sector for a fixed period of time (e.g., the franchising of passenger-rail services in the UK). These contracts are normally made through the use of a bidding scheme which has many ingredients of an auction. Public-sector franchising may affect a natural monopoly but can in principle be used in any situation where the government agency does not wish to lose ultimate responsibility for an operation, perhaps imposing what amount to regulatory controls, but wishes to obtain the benefits of private-sector cost efficiency – which is typically revealed to yield a 20 per cent benefit (Galal et al. 1994). Public franchise contracts may be for an operating franchise covering publically owned assets, for the operation of assets brought into the industry by the operator, or for the building, operation and transfer (BOT) of assets back to the public sector – possibly for further franchising.

Public-sector franchising is best thought of as a regulatory alternative to traditional forms of regulation by commission. It is not an alternative to regulation as such, as the schemes are likely to need a great deal of supervision and enforcement. Neither is it a new idea: the principles go back to the nineteenth century and there was a fashion for municipal franchising in the early twentieth century. There is a sense in which some economists have been contemplating a return to what is really an older form of regulating 'problem' industries.

The principles of public-sector franchising can be traced back to the Victorian social reformer Edwin Chadwick, who argued that contract management could be used to improve social welfare by substituting 'competition for the field' for 'wasteful', 'rapacious' or otherwise dysfunctional competition 'within the field'. Chadwick (1859) assembled numerous examples of market conditions where incentive structures could apparently be improved to the benefit of consumers by periodically awarding the right to be the monopoly supplier of a service to the firm willing to guarantee the lowest price for consumers. Demsetz (1968) seized on the idea as one of his demonstrations that there