ISSN 1045-6333

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

ECONOMIC ANALYSIS OF PUBLIC LAW ENFORCEMENT AND CRIMINAL LAW

Steven Shavell

Discussion Paper No. 405

02/2003

Harvard Law School
Cambridge, MA 02138

JEL Classifications: D00, D6, D8, K00, K14,K32, K42,L5, H00

ECONOMIC ANALYSIS OF PUBLIC LAW ENFORCEMENT AND CRIMINAL LAW

Steven Shavell*

ABSTRACT

This paper contains the chapters on public enforcement of law and on criminal law from a general, forthcoming book, *Foundations of Economic Analysis of Law* (Harvard University Press, 2003). By public law enforcement is meant the use of public law enforcement agents -- such as police, tax inspectors, regulatory personnel -- to enforce legal rules. A number of important dimensions of public law enforcement may be distinguished. One is the choice of the basic rule of liability: whether liability is strict or fault-based, and whether liability is imposed only if harm is done or may be imposed on the basis of acts alone (independently of the occurrence of harm). A second dimension of enforcement is the type of sanction, whether monetary or nonmonetary, notably, imprisonment. A third aspect of enforcement is the magnitude of sanctions. And a fourth dimension of enforcement is the degree of enforcement effort, which determines the probability of imposition of sanctions.

These dimensions of enforcement are discussed in the chapters that follow. In chapter 20, the basic theory of public enforcement employing monetary sanctions is discussed; in chapter 21, the basic theory of enforcement using nonmonetary sanctions is examined; and in chapter 22, extensions to the basic theory are considered.

Then, in chapter 23, functions of sanctions apart from deterrence, namely, incapacitation, rehabilitation, and retribution, are discussed. Finally, in chapter 24, the subject of criminal law is addressed against the background of the theory of public enforcement of law.

Samuel R. Rosenthal Professor of Law and Economics. Research support from the John M. Olin Center for Law, Economics, and Business is greatly appreciated.

Table of Comments

Economic Analysis of Public Law Enforcement and Criminal Law

Summary Table of Contents of

Foundations of Economic Analysis of Law

(forthcoming 2003, Harvard University Press)

Economic Analysis of
Public Law Enforcement
and Criminal Law

(part of Foundations of Economic Analysis of Law)

Steven Shavell

Chapter 20

Deterrence with Monetary Sanctions

The topic addressed here is the control of undesirable acts by the state through the use, or threatened use, of monetary sanctions. That is, the general subject is the deterrence of undesirable behavior through the use of monetary sanctions.

In the first part of the chapter, I assume for simplicity that monetary sanctions will apply with certainty -- that all parties to whom a rule should apply will be brought before social authorities and bear the intended sanctions. Then, in the second part, I assume that sanctions apply only with a probability. There I examine the use of sanctions assuming that the public must incur enforcement expense to locate and/or to convict and ultimately to penalize parties who should bear sanctions. The principal problems for society that are studied are the choice of the level of enforcement effort – which determines the probability of penalizing parties -- and the choice of the magnitude of sanctions, so as to maximize social welfare.

For convenience, I focus on the case in which parties are risk-neutral, so that parties will commit an act if the benefit to them from so doing exceeds the expected sanction. However, I also examine the case in which parties are risk-averse. In the risk-neutral case, social welfare is assumed to equal the gains parties obtain from acts, less the harm done by acts, less the costs of enforcement; in the risk-averse case the measure of social welfare also incorporates the disutility of risk-bearing. By the costs of enforcement, I mean the expenses of apprehending and convicting violators, but I assume that there is no resource cost associated with the actual imposition of monetary sanctions. This assumption is made to capture the important point that the payment of a fine is, in itself, only a transfer of purchasing power, as opposed to an expenditure of real resources.[1] (In contrast, the imposition of the nonmonetary sanction of imprisonment involves substantial direct costs. In the next chapter, the significance of this difference will be emphasized.)

[1] Of course, in fact the imposition of monetary sanctions does involve social costs, such as those involved in locating a person's assets and collecting a fine; this issue will be discussed in section 3 of chapter 22.

1. Certain Enforcement: Basic Theory of Liability

1.1 Introduction. Here, as just stated, I examine the theory of enforcement assuming that it occurs with certainty. I consider first the two basic forms of harm-based liability: strict liability; and fault-based liability, that is, liability for a harmful act that is judged to be an undesirable act. Then I consider analogous act-based rules. (This discussion will be in substantial respects a restatement of the discussion of strict liability and of negligence rules in chapter 8.)

1.2 Strict liability for harm. Under this rule, because a party always pays for the harm an act causes, the party's expected sanction equals the expected harm. Hence, he will commit an act if and only if his expected benefit exceeds the expected harm. That is, he will commit an act if and only if the act is socially desirable; the optimal outcome will result.[2] Note that, in general, if the sanction is less than harm, parties will sometimes act in ways that create greater harm than benefits. And if the sanction is greater than harm, there will be a chilling effect on desirable acts; parties will be discouraged from acts that create greater benefits than harm.

*Comments***.** (a) The only information required by the social authority in order to apply the strict liability rule is the level of harm.

(b) The assets of a party must be sufficient to pay for the harm; otherwise, the party will not generally be induced to act optimally and may engage excessively in harmful acts.

Risk-averse case. If parties are risk averse, they will tend to bear risk because they may find themselves in circumstances where the benefits from a harmful act are high enough to make committing it desirable, meaning that they will bear sanctions. In order to reduce the magnitude of this risk, it may be socially beneficial for the sanction to be less than harm. Moreover, if the sanction is less than harm, overdeterrence, that is, the discouragement of desirable acts, tends to be avoided.[3] (These statements presume that parties are not insured against sanctions; on such

[2] Let g be the gain, h the harm, and q the probability of harm (this notation will also be used in subsequent notes). There are two natural cases to examine: where g is enjoyed only when harm comes about (suppose a person throws a rock at a window and is trying to break it), and where g is enjoyed when the act is committed, regardless of whether harm comes about (suppose that a firm discharges a potentially harmful pollutant into a river in order to save the costs of hauling its waste to a dump site -- here it saves the costs for sure, regardless of whether the pollutant causes harm). In either case, liability equal to harm will lead to optimal behavior. In the first case, it is optimal for the act to be committed if and only if $qg > qh$, and because the sanction equals harm, the person will commit the act if and only if that is true. In the second case, it is optimal for the act to be committed if and only if $g > qh$, and again, if the sanction equals harm, the person will commit the act if and only if $g > qh$. In the text, I will not usually distinguish these cases for expositional convenience, and it will be clear that the conclusions to be noted hold for both cases, as I will sometimes explicitly note below.

[3] To be more precise, let me specify the state's problem of maximizing social welfare in a simple model where parties are risk averse. Suppose that $U(y)$ is the utility function from income y of members of a population of risk averse individuals with identical initial incomes, and among whom the gain g from committing the act (for concreteness, consider here and in many later notes the case where g is enjoyed with certainty; see note 2) is distributed according to the density $f(g)$. Then, if s is the sanction for harm, an individual will commit the act if and only if $(1 ! q)U(y + g) + qU(y + g ! s) \$ U(y)$. Thus individuals will commit the act when $g \$ g^*(y, s)$, where the critical g^* can be shown to be decreasing in y and increasing in s. It is presumed that the income y of each individual is income net of taxes, where taxes are set in order to cover the state's expenses. The state collects fine revenue and, for simplicity, is assumed to suffer harms done. Therefore, $y = z ! (1 ! F(g^*))(qh ! qs)$, where z is the initial income of each person and F is the cumulative distribution function of f; for $(1 ! F(g^*))$ is the fraction of individuals who commit the act, qh is the expected harm caused by a person who commits the act, and qs is the expected revenue collected from such a person. The social problem is then to choose s to maximize social welfare W, the sum of expected utilities, that is

$$W = F(g^*(y, s))U(y) + \int^{4} [(1 ! q)U(y + g) + qU(y + g ! s)]f(g)dg.$$

insurance, see section 6 of chapter 22.)

1.3 Fault-based liability for harm. Under this rule a party who causes harm is liable and bears a sanction equal to harm only if his act was undesirable, that is, only if the social authority finds that the expected harm exceeded his expected benefits. If, for example, the expected harm is $100 and the gain $60, the act would be found undesirable, so there would be liability for harm. A party would, however, not engage in such an undesirable act, for his expected sanction would equal the expected harm and thus exceed his benefit (his expected liability would equal $100 and exceed $60). If an act is desirable, however, a party will clearly commit it, for then he will not bear liability for any harm that comes about as a result.[4]

Note again that if the sanction for an undesirable act is less than harm, then parties may sometimes commit such acts because their gain may exceed their expected sanction. If, however, the sanction for an undesirable act exceeds harm, there will not be an chilling effect on desirable acts, for these acts are not subject to sanctions under fault-based liability; hence, sanctions for undesirable acts that exceed harm will still lead to optimal outcomes.

*Comments***.** (a) The information needed by the social authority to apply the fault-based liability rule is not only the level of harm, but also its likelihood and the benefit from the act, for to determine whether an act is desirable or not, the authority must compare the benefit to the expected harm.

(b) Again, the level of assets must in general be sufficient to pay for the harm, in order that the party be induced not to commit undesirable acts.

Risk-averse case. Parties will bear no risk under fault-based liability if fault is found without error; this is an advantage of the fault-based form of liability over strict liability (again assuming that insurance against sanctions is not sold). Parties will, however, bear some risk of sanctions in the presence of uncertainty concerning findings of fault -- generated by errors in the determination of fault or by parties' imperfect ability to control their behavior. Thus, if parties are risk averse, the observations made in the case of strict liability carry over to the present rule, to the extent that the parties bear risk due to uncertainty in findings of fault. Notably, liability exceeding harm may well have a chilling effect on desirable acts.

$g^*(y, s)$

It can be shown under fairly general conditions that the optimal s is less than h, the intuition being as stated in text. Of note is that by lowering s from h, there is a gain in social welfare due to a reduction in risk-bearing by those who commit the act and might be sanctioned. This is so as long as the wealth of those who are sanctioned, $y + g - s$, tends to be lower than that of individuals in general (who have to pay higher taxes if s is lowered), for then the marginal utility of those who are sanctioned is higher than average marginal utility.

Finally, it should be observed that the expression for social welfare W reduces in the risk-neutral case to

$$W = z + \int_{qs}^{\infty} gf(g)dg - (1 - F(qs))qh,$$

that is, a constant plus total gains minus total harms. For in the risk-neutral case, we can take $U(y) = y$, so that $g^*(y, s)$ reduces to qs, and substitution in the previous expression for W yields this expression.

[4] To amplify, under fault-based liability, a party who commits an act and causes harm will be held liable if and only if the act was undesirable, that is, if and only if $g < qh$; otherwise he will not beheld liable. If the sanction $s = h$, then any party for whom $g < qh$ will face expected liability if he commits the act equal to qh, so will not commit the act; others will face no liability. Hence, if $s = h$, all undesirable acts will be deterred and all desirable acts will be committed.

Sanction equal to wrongdoer's gains.[5] A version of fault-based liability that is of interest is that under which a party who commits a harmful undesirable act bears a sanction equal to his gains.[6] This sanction is sometimes thought to be a natural one for purposes of deterring acts, as it removes a wrongdoer's gains. But although a sanction equal to gains will discourage undesirable behavior, it will, in principle, only barely do so, as parties lose no more than their gains. Consequently, the rule of sanctions equal to gains is peculiarly vulnerable to judicial error in assessment of gains, and for that reason tends to be inferior to fault-based liability with sanctions equal to harm. Specifically, under the rule with sanctions equal to gains, if the gain is underestimated even by a small amount, parties will have an incentive to engage in an act, no matter how much harm it causes. Suppose, for example, that an act creates a gain of $1,000 and harm of $1,000,000. If the gain is estimated to be $950, a party would have an incentive to engage in it, because he would profit by $50. In contrast, if liability is equal to harm, parties will be strongly discouraged from committing the act, even if there is substantial judicial error in estimating the harm.[7]

1.4 Act-based liability. Both strict and fault-based liability for harm have act-based counterparts. The act-based analogue to strict liability for harm is the rule under which a party is liable for the expected harm due to an act, regardless of whether harm actually occurs or not. Thus, if a party commits an act that will cause harm of $1,000 with probability 10 percent, he will be liable for $100 for having committed the act. It is apparent that under this rule, the party will behave just as under strict liability for harm; he will commit an act if and only if the benefit obtained exceeds the expected harm. Similarly, the act-based analogue to fault-based liability for harm is liability equal to the expected harm for undesirable acts, and it is clear that under this rule, parties will be induced not to commit undesirable acts.

Comments. (a) The social authority needs to know more in order to apply act-based rules than harm-based rules. To apply act-based strict liability, the authority needs to know not only the harm -- which it does not observe if harm does not come about -- but also the probability of harm. By contrast, to apply harm-based strict liability, it needs only to know the harm that has occurred. With regard to act-based fault liability, the social authority also faces the disadvantage that it does not observe harm (but it needs to know the probability of harm and the benefits from the act under either harm-based or act-based fault liability).

(b) The level of assets that a party needs to have in order to be motivated to act appropriately is lower under act-based liability than under harm-based liability. Under act-based liability, to be properly motivated, a party needs assets equal only to the expected harm rather than the actual harm (in the example, assets of $100 rather than $1,000).

[5] I do not consider the analogue to this rule under strict liability, namely, liability equal to gains for desirable acts as well as undesirable acts. Such a rule would obviously be perverse, as it would remove any incentive to engage in desirable acts.

[6] More precisely, under the rule in consideration, if a party obtains his gain g only if he does harm (see note 2), the sanction imposed on the party equals g. However, in the case where the party obtains g for sure when he acts, the sanction under the rule in consideration is interpreted to be g/q, so that the expected sanction equals g.

[7] On the advantage under discussion of sanctions equal to harm rather than the wrongdoer's gains, see Polinsky and Shavell 1994.

Risk-averse case. Risk-averse parties bear less risk under act-based liability because sanctions equal the expected harm rather than the realized harm.

1.5 Actual use of rules. In fact, we often observe use of harm-based sanctions, both on a strict basis and according to fault. Penalties may be imposed by the state for spills of toxic materials, for failure to pay proper taxes, and for many other harmful events. Perhaps more often, however, we see that public law enforcement involves act-based sanctions. This is typically the case with violation of safety, environmental, and many financial regulations, where sanctioned behavior is that which creates a positive expected harm but need not do actual harm.

2. Law Enforcement with a Probability: The Optimal Probability and Magnitude of Sanctions

2.1 Introduction. Here it will be assumed that it is costly to identify and penalize liable parties, so that society has to choose a level of enforcement effort, which will determine the probability of applying sanctions, as well as the magnitude of sanctions. In determining the social-welfare-maximizing choice of the probability and magnitude of sanctions, I will for simplicity assume that liability is strict and based on harm, for the major points to be made do not depend on the nature of the rule of liability (except as remarked in section 2.6 on fault-based liability).

2.2 Behavior given the probability and magnitude of sanctions. How will a person behave who will face a sanction only with a probability if he commits an act? If the person is risk neutral, he will evaluate the sanction in terms of its expected value. Hence, the person will commit an act if and only if his benefit exceeds the expected sanction.

Risk-averse case. If the individual is risk averse, he will commit an act if and only if his expected utility is raised by so doing, and in general he will not be equally deterred by different combinations of sanction and probability with the same expected value; he will be more deterred the higher the magnitude of the potential sanction in the combination, the expected sanction held constant. For example, a risk-averse person will be more deterred by a sanction of $1,000 borne with probability 20 percent than by a sanction of $500 borne with probability 40 percent even though their expected values, $200, are equal. The reason is that, for a risk-averse party, the disutility of sanctions rises more than in proportion to their size; when the sanction rises from $500 to $1,000, its disutility more than doubles.[8]

Comments. (a) *Probability versus magnitude of sanction.* It is sometimes asked whether an increase in the probability or an increase in the magnitude of sanctions would make a greater difference in deterrence. But this question is incomplete as stated, for it is not explicit about the degree of change of these two factors. Obviously, if the magnitude of the sanction rises by much more than the probability, an increase in the magnitude would exert a greater effect on

[8] More generally, if U is the utility of income function of a risk-averse person, y is income, g is the gain from the act, p is the probability of a sanction, and s is the magnitude of the sanction (this notation will be used in many later notes as well), the person's expected utility if he commits the act will be $EU = (1\ !\ p)U(y + g) + pU(y + g\ !\ s)$. If p falls to kp, where $k < 1$, and s rises to s/k (so that the expected sanction is still ps), the person's expected utility becomes $(1\ !\ kp)U(y + g) + kpU(y + g\ !\ s/k)$. Differentiating the latter expression with respect to k yields $p\{(s/k)UN(y + g\ !\ s/k)\ !\ [U(y + g)\ !\ U(y + g\ !\ s/k)]\} > 0$ because UN is decreasing. Hence, the lower is k, the lower is expected utility, and therefore the greater is deterrence.

deterrence than would an increase in the probability, and conversely.

A natural and well-posed question, however, is how a given percentage increase in the probability of sanctions compares in importance to the same percentage increase in the magnitude of sanctions. If parties are risk neutral, any named percentage increase in the probability of sanctions has an identical effect to an equal percentage increase in the magnitude of sanctions -- for a given percentage increase in either the probability or the magnitude of sanction will raise the expected sanction by exactly that percentage. If there is a 20 percent probability of imposition of a sanction of $500 and the probability doubles to 40 percent, the expected sanction will double, from $100 to $200; and likewise if the sanction doubles to $1,000 (and the probability remains at 20 percent), the expected sanction will double to $200. Thus, a risk-neutral party will be affected in the same way by either type of change.

If parties are risk averse, however, they will be more affected by a percentage increase in the magnitude of sanctions than by an equal increase in the probability of sanctions. A risk-averse party will be more deterred by the sanction of $1,000 with probability 20 percent than by the sanction of $500 with probability 40 percent. The reason is, as was just noted, that risk-averse parties suffer disutility more than in proportion to increases in the magnitude of sanctions.[9]

Still, one often encounters the notion that the probability of sanctions (or, as it is frequently expressed, the certainty of sanctions) matters more than their magnitude. Although this disagrees with our conclusions for both risk-neutral and risk-averse individuals, it could be the case that probability matters more due to the ineffectiveness of large sanctions, notably, the fact that people may be unable to pay large amounts.

(b) *Perception of the probability of sanctions.* Information that individuals have about the probability of sanctions will often be imperfect. Enforcement authorities generally do not publish data on the likelihood of punishment. Moreover, the probability of sanctions is frequently variable, depending on the circumstances of the violation, so that even if enforcement authorities were forthcoming, there would inevitably be substantial imperfection of knowledge about the probability. In addition, individuals often experience difficulty in assessing and interpreting probabilities, especially small ones, sometimes failing to discriminate among them, sometimes inflating their importance, and sometimes essentially ignoring them. These observations suggest the need for caution in applying what would appear to be the effect of the probability of sanctions on behavior.[10]

(c) *Perception of the magnitude of sanctions.* Information about the magnitude of sanctions may also be imperfect. This is most likely to be true when the sanction is decided upon by a court or other tribunal that enjoys discretion over sanctions, so that there is no set

[9] Specifically, let us assume as in the previous note that expected utility $EU = (1 \,!\, p)U(y + g) + pU(y + g \,!\, s)$. We want to show that the (negative of) the elasticity of EU with respect to p is less than that with respect to s. The elasticity of EU with respect to p is $[p/EU][dEU/dp] = p[U(y + g \,!\, s) \,!\, U(y + g)]/EU$, and the elasticity of EU with respect to s is $[s/EU][dEU/ds] = \,!psU'(y + g \,!\, s)/EU$. We therefore need to show that $sU'(y + g \,!\, s) > [U(y + g) \,!\, U(y + g \,!\, s)]$, but this holds because U' is decreasing.

[10] See Bebchuk and Kaplow 1992, Garoupa 1999, and Sah 1991 on perceptions of the likelihood of sanctions and learning about them. For empirical evidence on knowledge of expected sanctions, see, for example, Wilson and Herrnstein 1985.

magnitude of sanctions, but only a distribution of them. In many contexts, however, sanctions are stipulated and well known in advance.

(d) *Level of wealth of a party.* The level of wealth of a party imposes a ceiling on the maximum sanction. Thus, the lower is the probability of sanctions, the lower is the maximum expected sanction, so that it might be impossible to deter a person from committing an act even if his benefit from it is quite modest if the probability of sanctions is small. For example, consider a risk-neutral individual with wealth of $5,000 who would obtain a benefit of $100 from an act. It would be impossible to deter this person from committing the act if the likelihood of sanctions is 1 percent, for then the highest expected sanction that he could face is 1% H $5,000 = $50.

The level of wealth of a party not only determines the maximum sanction that can be imposed on a party, it also may influence how he reacts to the risk of sanctions generally, for the degree of risk aversion is usually thought to depend on wealth. The more wealthy a party is, the less averse to risk, and thus the less he tends to be deterred by a given probability and magnitude of sanction.[11]

2.3 Optimal sanctions when the probability of their imposition is a given. Let me now address the question about the socially best magnitude of sanction, taking the probability of imposition of sanctions as a given. The assumption that the probability of sanctions is taken as given is relevant in many contexts, as those who decide on the magnitude of sanctions may not have control over enforcement effort. For example, a judge or administrative officer who sets the fine for a regulatory infraction may take the enforcement budget and its allocation as a given. Further, in many areas of enforcement, the probability of sanctions for a particular type of infraction is set by overall policy and is not independently variable (see section 5 of chapter 22). In any case, the problem of determining the optimal sanction given the probability of sanctions is a subpart in a theoretical sense of the problem of finding the optimal probability and sanction, for to find the optimal probability, one must in general find the optimal sanction for each probability.

If parties are risk neutral, optimal behavior will be induced if the expected sanction equals the expected harm, for then a party will compare his benefit to the expected harm. Consequently, the sanction, when imposed, must exceed harm; in particular, the sanction is governed by a fundamental probability-related *multiplier* -- the sanction must equal the harm multiplied by the inverse of the probability of its imposition.[12] Thus, if the harm is 100 and the probability of sanctions is 50 percent, the sanction should be multiplied by 1/.5 = 2, so the sanction should equal 200 (and thus the expected sanction equals 100); if the probability of sanctions is 25 percent, the sanction should be multiplied by 1/.25 = 4, so the sanction should equal 400 (again the expected sanction equals 100); and so forth. In this way, parties will behave optimally; the situation will be as if they faced liability equal to the harm.

[11] I will comment generally on the actual effect of sanctions (both monetary and nonmonetary) on deterrence in section 2.3 of chapter 21.

[12] If harm is h and the probability of proper imposition of the sanction is p, the magnitude of the sanction should be h multiplied by $1/p$, so that the expected sanction is $p(h/p) = h$, resulting in optimal behavior under strict liability (and fault liability).

Risk-averse case. If parties are risk averse, the optimal sanction tends to be lower than it is when parties are risk neutral. The reasons are essentially as indicated above in section 1.2. First, because parties for whom the act is socially desirable will often commit it, they will bear risk, which is socially undesirable in itself. Second, if the sanction equals its optimal level in the risk-neutral case, risk-averse individuals will tend to be overdeterred. Lowering the sanction ameliorates both of these problems.[13]

Comments. (a) *Practical ability to impose high sanctions reflecting the probability of their imposition.* The theme of this section is that sanctions should be scaled upward to reflect the likelihood of escaping liability. There are several problems, however, that may be faced in actually imposing such sanctions. First, there may be resistance to inflating sanctions on grounds of fairness; the notion that the magnitude of sanctions should be proportional to the gravity and moral quality of an act is a widely held notion of fairness, and this notion does not accord weight to the likelihood of escape from sanctions. For example, the fair punishment for an act such as littering might be thought quite modest (perhaps no more than $10 or $20) because an act of littering is not considered to be seriously harmful, even though the sanction called for by the principles discussed here would be substantial (such as $200) if the probability of catching a litterer is small.[14]

A second problem is that there may be significant difficulty in determining the probability of sanctions. For example, if a restaurant violated an ordinance about safety in its kitchen, the sanctioning authority would have to take into account such factors as the probability of inspection of the restaurant, the probability that employees would make reports to authorities themselves, the probability that customers would notice something wrong, and the like. These determinations are often difficult and lend themselves to dispute, although, as with any type of determination, they can be performed more cheaply if demands for accuracy are reduced.

(b) *Effect of wealth.* It should be borne in mind that the wealth of the party may be too low (consider especially individuals with essentially no savings, or thinly capitalized firms) for the party to be induced to act optimally. If the likelihood of being caught is low and the magnitude of the harm high, it may be impossible to induce the party to act optimally, leading to a significant problem of underdeterrence.

2.4 Optimal sanctions when the probability of their imposition is also optimally determined. One of the basic insights that apply to optimal law enforcement when the state chooses both the probability of imposing sanctions and their magnitude is that a low probability-high magnitude sanction policy is socially advantageous. The reasons are two-fold: A social savings in enforcement effort can be achieved by allowing sanctions to be imposed only with a low probability; and sanctions can be raised to avoid dilution of deterrence from the low probability of sanctions.[15] This strategy for conserving enforcement resources without sacrificing

[13] This can be shown along the lines sketched in note 3.

[14] Issues of fairness in sanctions are discussed in chapter 27. On fairness and the economic theory of law enforcement, see Polinsky and Shavell 2000a and Kaplow and Shavell 2002b, chapter 6.

[15] Note that that the rise in the sanction does not increase enforcement expenditures; this is an aspect of the maintained assumption of this chapter that the imposition of monetary sanctions does not involve resources costs.

deterrence has the apparent implication that enforcement effort and probabilities of sanctions should be very low, but be accompanied by very high sanctions. Such a draconian conclusion will shortly be seen to hold if parties are risk neutral. But this strong conclusion does not hold if parties are risk averse (or if any of a variety of other factors are relevant, as will be noted later), even though the conclusion contains an important element of the truth about optimal policy under all circumstances.

Suppose that parties are risk neutral. In this case, it is optimal for the fundamental strategy for saving enforcement resources just mentioned to be employed to the fullest extent, meaning that the sanction should be as high as possible, equal to the entire wealth of an individual. To understand why, suppose that the sanction is less than maximal. Then the sanction can always be raised and the probability lowered proportionally, so that deterrence is not altered; but as the lower probability will mean a savings in enforcement costs, the change must raise social welfare. For example, suppose that the wealth of individuals is $10,000, the likelihood of sanctions is 10 percent, and the sanction is $1,000. Thus, in particular, the expected sanction is $100. Now if the sanction is raised to $2,000 and the probability of sanctions is lowered to 5 percent, the expected sanction and deterrence will be unchanged, and equal $100, but enforcement expenses will be lowered. Indeed, if the sanction is raised to the maximum, $10,000 and the probability of sanctions is reduced to 1 percent, deterrence will be unchanged and more enforcement expenses will be saved. The conclusion, therefore, is that sanctions should be raised until they are maximal.[16]

What is the optimal probability of imposing the sanction? It might at first seem that the best probability is such that the expected sanction equals the harm. In the example under discussion, this would mean that if the harm from the act is $100, the expected sanction should be the same, so that the probability *p* should satisfy $p \times \$10{,}000 = \$100$, implying that the best *p* is 1 percent. But in fact the optimal probability should be lower than 1 percent. In general, the optimal expected sanction is less than the harm. The reason for this conclusion, another basic insight about optimal enforcement when the probability of sanctions is chosen along with the magnitude of sanctions, is essentially that, because of the cost of enforcement, it is better to compromise and not achieve perfect behavior, but rather to permit a degree of underdeterrence in order to save enforcement resources.[17] If the cost of enforcement is significant, it may be best to

[16] To establish this conclusion formally, observe that social welfare in the risk-neutral case, the benefits obtained from committing acts less harm and enforcement costs, is given by

$$W = \int_{ps}^{\infty}(g - h)f(g)dg - c(p),$$

where $c(p)$ is the enforcement cost of setting the probability equal to p. (It is assumed here for simplicity that an act causes harm with certainty, rather than only with a probability, and this will also be assumed in subsequent notes.) Clearly, if s is not maximal, s can be raised to the income y of individuals, and p can be lowered to $p(s/y)$, so that the expected sanction $[p(s/y)]y$ remains ps. Hence, the integral in W does not change but $c(p)$ falls, so that W rises, meaning that raising s to y and lowering p increases welfare; thus the optimal sanction must be maximal. Note that this conclusion that the optimal sanction is maximal does not depend on the magnitude of the harm. Becker 1968 first suggested the conclusion (although much of his analysis presumes the sanction is not maximal) and it is noted explicitly in Carr-Hill and Stern 1979 and Polinsky and Shavell 1979.

[17] To amplify the point that some degree of underenforcement is desirable, suppose in the example that the expected sanction is $99 instead of $100 -- which would be the case if the probability is .99 percent instead of 1 percent.

allow substantial underdeterrence to reduce costs of enforcement.

Indeed, because of the costs of enforcement, it is possible that it will be optimal for there not to be any law enforcement, for society to countenance harm in order to save the costs of law enforcement -- the game of enforcement may not be worth the candle. This can be demonstrated to be true, other things being equal, if the harm from the act is below a certain threshold.

Risk-averse case. In this case, the conclusion differs from that when parties are risk neutral. The main difference is that the optimal sanction is not maximal, in general, and may be much lower than maximal. For instance, in the example discussed above, the optimal sanction might be $300 rather than $10,000, the level of a person's wealth. The reason, roughly, is that the risk aversion of individuals means that their bearing the risk of sanctions constitutes a form of social cost.[18] The optimal level of the sanction will depend, among other things, on the degree of risk aversion of parties; the more risk averse the parties, the lower the optimal sanction will tend to be.[19]

With regard to the optimal probability, two points should be made. First, the optimal probability might be higher than in the risk-neutral case: If the sanction is, in effect, constrained

Then the individuals who would be undesirably led to commit the harmful act would be those obtaining benefits of between $99 and $100 and doing harm of $100. Thus, they would be contributing only slightly to net social harm (harm minus benefit obtained) -- for they would cause net social harm of less than $1 each. On the other hand, the social saving in enforcement expenses from reducing the enforcement probability is proportional to the probability reduction. For this reason, it is always desirable for the probability to be lowered some amount below 1 percent, so that the expected sanction is below $100. Formally, differentiate W in note 16 with respect to p and set this equal to 0, yielding $s(h - ps)f(ps) = c'(p)$. Because the right side is positive, $h > ps$ must hold (whether or not s is optimal, equal to y).

[18] Another reason that the optimal sanction may not be maximal is that higher sanctions may induce violators to spend additional resources to avoid punishment; see Malik 1990. Further reasons will be given in later chapters.

[19] Further insight into the risk-averse case can be gained by considering why, precisely, the argument applying in the risk-neutral case for optimality of maximal sanctions fails when parties are not risk-neutral. Consider any situation in which the sanction is less than maximal -- consider for instance a sanction of $1,000 and a probability of imposition of sanctions of 10 percent. Now raise the sanction to wealth, $10,000. Even though individuals are risk averse, there will be *some* reduction of the probability to a level p that will leave the risk-averse individuals indifferent between bearing the $10,000 sanction with probability p and instead bearing the $1,000 sanction with probability 10 percent. But, due to risk aversion, this p will be less than 1 percent, perhaps it will be .1 percent. At the new p and the $10,000 maximal sanction, deterrence will, by construction, be preserved: Parties who commit the harmful act will be just as well off as they were when they faced the $1,000 sanction with probability 10 percent, and enforcement resources will have been saved (indeed, even more resources will have been saved than in the risk-neutral case, when p falls only to 1 percent). So why will not social welfare necessarily have been raised? The answer is that the state's *revenue* from sanctions will have fallen, as the expected sanction will be lower (such as .1% × $10,000 = $10 for each person who commits the act, instead of $100). This decline in revenue might offset the savings in enforcement costs, and, if so, will result in higher taxes and thus tend to lead to lower welfare.

The formal problem in the risk-averse case is similar to that sketched in note 3, namely, to maximize social welfare

$$W = F(g^*(y,s))U(y) + \int_{g^*(y,s)}^{\infty} [(1 - p)U(y + g) + pU(y + g - s)]f(g)dg.$$

over s and p, where $g^*(y,s)$ is defined by $(1 - p)U(y + g) + pU(y + g - s) = U(y)$. Also, $y = z - (1 - F(g^*))(h - ps) - c(p)$, where z is the initial income of each person, so the second term is taxes. Essentially this problem is solved in Polinsky and Shavell 1979. For further analysis, see Kaplow 1992.

not to be high due to the risk aversion of individuals, say to be in the range of $300, then the only way to achieve a particular level of deterrence is through use of greater enforcement than would be needed were the sanction maximal. Second, the optimal probability could also be lower than in the risk-neutral case: If the sanction must be fairly low due to risk aversion, the effectiveness of raising the probability is reduced, leading to the possibility that the optimal probability could be lower than in the risk-neutral case, or that it might not be worth controlling the activity at all, even though it would be in the risk-neutral case.

A further point is worth mentioning. The reason that has been discussed why some risk-averse parties bear risk is that it may turn out to be desirable for them to commit harmful acts and they will do so. However, as we know, there are other reasons for risk-bearing -- and thus for sanctions to be less than maximal -- notably, legal errors that result in the imposition of sanctions on innocent parties.

2.5 Comment on the misleading notion that sanctions are analogous to market prices -- that willingness to face sanctions for harmful acts implies that committing such acts is socially correct. It is commonly stated that if a party is willing to pay a sanction, or face an expected sanction, then it is not socially incorrect, indeed it is socially desirable, for him to commit an act, such as to pollute, since the willingness to bear the expected sanction signals that his benefit is higher than the expected sanction. The analogy to paying a price for a good is said to apply, whereby, if a party is willing to pay the price of a good, the purchase is inferred to be socially desirable, since the willingness to pay the price implies that the value that the party places on the good must exceed its production cost. This line of thinking is offered both as a criticism of the economic way of thinking by some, and as a point of interest, asserted to be correct, by economists.

However, this view represents an incorrect interpretation of economic analysis of optimal law enforcement. As has been explained above, *optimal law enforcement is characterized by underdeterrence -- and perhaps by substantial underdeterrence -- due to the costliness of enforcement effort and limits on sanctions*. For example, the probability and magnitude of sanctions against pollution may fall significantly short of discouraging as much pollution as would be ideal -- because of the costs of raising the likelihood of enforcement and because of limits on the magnitude of sanctions. Consider a firm that faces a maximum sanction equal to its assets of $100,000, that could take a precaution that costs $10,000 and would prevent pollution harm of $25,000, and that would be sanctioned for pollution only with a probability of 5 percent due to the high cost of detecting the source of the pollution. This firm might well find it in its private interest to pollute -- its savings from not taking the precaution of $10,000 is double the maximum possible expected sanction of 5% H $100,000 = $5,000. But the firm's failure to take the precaution would most definitely be socially undesirable -- pollution causes harm of $25,000 yet saves prevention costs of only $10,000. It is often the case that when parties choose to commit harmful acts and the likelihood of sanctions is low, it would be socially best that they do not commit the acts; they commit the acts only because the social cost of enforcement effort results in inadequate expected sanctions.

Note, however, that if enforcement is certain, the conclusion may be different. For example, if we imagine pollution taxes to be imposed with certainty in some context (because it is administratively easy to do so), then by setting the tax equal to the harm due to pollution, the privately induced behavior will also be socially desirable.[20] In such a setting, the behavior of the polluter is like that of a person who purchases a good on a market (where, note, the payment for the good is made with certainty).

2.6. Fault-based liability. The conclusions about the optimal probability and magnitude of sanctions under fault-based liability are similar to those I have discussed above for strict liability, but with some differences.

Optimal sanctions given the probability of their imposition. In this case, as under strict liability, it is optimal for the sanction to equal the harm multiplied by the inverse of the probability of its imposition, for that will result in an expected sanction equal to harm, and thus induce individuals not to act with fault.[21] However, unlike the outcome under strict liability, any higher sanction will also lead to desirable behavior, assuming that the fault system is error free. Higher sanctions only reinforce the incentive not to act with fault, but do not discourage desirable yet possibly harmful behavior -- for such behavior is not sanctioned. Also, unlike the outcome under strict liability, risk aversion does not reduce the optimal sanction, assuming again that the fault system is error free, as parties do not bear risk; parties who do harm will be those whose acts are not faulty and thus will not be sanctioned, and others will be discouraged from committing harmful acts.

Yet if the fault system is not error free, the optimal magnitude of sanction could, in general, be different from the harm multiplied by the inverse of the probability; the optimal sanction could be higher or lower depending on circumstances. The presence of error also means that risk aversion becomes relevant under the fault system, and thus lowers the sanction from what would otherwise be its optimal level.

Optimal sanctions and the optimal probability of their imposition. Here, as under strict liability, the optimal policy involves the maximal sanction and a low probability of its imposition if parties are risk neutral, for this policy conserves enforcement resources. If parties are risk averse, there is a lesser need to employ moderate sanctions than under strict liability because many of the parties who do harm are those who act without fault and thus do not bear risk. However, some risk will tend to be borne by parties if there is error in the fault determination. Also, it will often be the case that some parties will bear risk because of the general optimality of permitting underdeterrence in order to save enforcement costs.

3. Synopsis

The basic rules of liability to the state and optimal sanctions were first considered here under the assumption of certain enforcement. The main conclusions about liability rules were that both strict liability and the fault rule give rise to correct behavior, but strict liability requires less

[20] This will be so provided that the polluters can pay the tax. Polluters are more likely to be able to pay a tax equal to harm than the higher sanction that would be necessary to create an expected sanction equal to harm when sanctions are applied only with a probability. For example, the firm mentioned in the paragraph above would be able to pay a tax equal to the pollution harm of $25,000, as its assets are $100,000, but the firm would not able to pay $500,000, which is the sanction necessary to create an expected sanction of $25,000 when the probability of sanctions is 5 percent.

[21] Under the fault system a person is liable if and only if $g < h$. Thus, if $s = h/p$, then because expected liability for fault is h, no one will act with fault.

knowledge on the part of the state (only knowledge of harm). It was also noted that harm-based sanctions require the state to possess less information than act-based sanctions, but that act-based sanctions have the advantage that parties' assets need not be as high for liability rules to function well. The optimal magnitude of sanctions equals harm if parties are risk neutral, and is less than harm if parties are risk averse (and uninsured against sanctions).

Then it was assumed that parties face sanctions only with a probability, but the probability was regarded as fixed (which is sometimes realistic). The main point here was that the magnitude of sanctions should be raised to offset the probability of escaping sanctions. In particular, the optimal sanction equals the harm multiplied by the inverse of the probability of sanctions if parties are risk neutral, and is less than this if parties are risk averse.

Last, it was assumed that parties face sanctions with a probability that is optimally chosen. Here a crucial point was that there is a social advantage associated with a low probability-high sanction enforcement strategy: The low probability means that the state conserves enforcement resources, and the high magnitude of sanctions prevents dilution of desired deterrence. The optimal strategy involves maximal sanctions if parties are risk neutral, but lesser sanctions if parties are risk averse.

A second point of stress about optimal law enforcement is that it will tend to involve underdeterrence, for the costliness of enforcement effort will make it desirable to spend less than what would be needed to achieve perfect deterrence. Therefore, the fact that an individual chooses to commit an act and suffer the consequences does not imply that the act was desirable to commit -- the analogy to sanctions as prices that lead to socially desirable choices is misleading.

Note on the literature. The basic point that sanctions should be inflated to offset the probability of escaping liability, and in particular multiplied by the inverse of the probability of escaping liability, was emphasized by Bentham ([1789] 1973) in his treatment of law enforcement. Becker (1968) first considered the question of the optimal social choice of the probability of enforcement and stressed the advantage of the low probability-high sanction enforcement policy. Polinsky and Shavell (1979) initially considered risk aversion in enforcement policy and showed that it implied that optimal sanctions are not maximal.[22]

[22] For surveys of economic literature on enforcement, see Garoupa 1997, Mookherjee 1997, and Polinsky and Shavell 2000a.

Chapter 21

Deterrence with Nonmonetary Sanctions

In this chapter, I consider the deterrence of undesirable behavior by the state when the form of sanctions is nonmonetary. The important assumption that will be made about nonmonetary sanctions is that they are socially costly to impose, and the primary form of nonmonetary sanction that will be borne in mind is imprisonment. Imprisonment is clearly socially costly to employ: Prisons must be built and operated, production of individuals is forgone during their imprisonment, and individuals suffer disutility during imprisonment.

In the first section of the chapter, I consider enforcement assuming that nonmonetary sanctions are imposed with certainty, and in the second section, that they are imposed only with a probability determined by the enforcement effort of the state. Then, in sections three and four, I examine the question of when it is socially desirable to employ nonmonetary sanctions, rather than only monetary sanctions. In the last section, I consider types of nonmonetary sanctions apart from imprisonment.

The assumptions about individual behavior and social welfare that I make are similar to those of the last chapter. For simplicity, I focus on the assumption that individuals are risk neutral with respect to sanctions, but I will note other possibilities. Social welfare is assumed to equal the benefits that parties obtain from their acts, less the harm done by the acts, less the costs of enforcement, and less the costs associated with the imposition of sanctions.

1. Certain Enforcement with Nonmonetary Sanctions: Basic Theory of Liability[23]

1.1 Introduction. Here I initially consider strict liability and explain why it is generally a disadvantageous form of liability compared to fault-based liability when sanctions are nonmonetary. (This is in fundamental contrast to the conclusions reached when sanctions are monetary, as discussed in chapter 20.) I then discuss the optimal use of fault liability.

1.2 Strict liability for harm. Suppose that individuals are held strictly liable for causing harm. Then the sanction can generally be chosen so as to induce ideal behavior.[24] If, for instance, an act causes harm of 1,000 and there exists an imprisonment sanction[25] creating disutility equal to 1,000, then individuals will commit the act if and only if they obtain benefits exceeding 1,000, which constitutes ideal behavior under our assumptions.

Although optimal behavior can therefore be induced, this *socially desirable behavior will be accompanied by the imposition of socially costly sanctions on those who commit harmful acts.*

[23] The points made in this section are developed in Shavell 1985b, 1987a.

[24] The only reason that ideal behavior would not be achievable is that there may not exist a sanction high enough to offset the benefits to an individual. This possibility will be discussed below, but is not important to the argument to be made in this section on strict liability, so it will not be mentioned again here.

[25] I will speak of nonmonetary sanctions as imprisonment until section 5 below, where I explicitly consider other forms of nonmonetary sanctions.

If the social cost of imposing the sanction that creates disutility of 1,000 is, for instance, 1,500 (composed of the disutility of 1,000 suffered by a person who is sanctioned and the costs of operating the prisons), then each time a person commits the act (because the person obtains high benefits from so doing), social costs of 1,500 as well as the harm of 1,000 are generated. This makes strict liability a socially expensive way to induce behavior that would otherwise be desirable.[26]

Note too that because under strict liability social costs of imposing sanctions are incurred whenever individuals commit harmful acts, the optimal magnitude of the sanction will not be the magnitude that leads to ideal behavior; it will be such as to reduce the social costs of imposing sanctions.[27]

Comment: comparison to the case under monetary sanctions. When sanctions are assumed to be monetary and costless to impose, as in the last chapter, optimal behavor can be induced at no social cost by setting the sanction equal to the harm. Here, when sanctions are nonmonetary, the situation is altogether different, due to the cost of actually imposing the sanction. For example, consider a harmful act such as polluting. If sanctions are monetary, then strict liability induces optimal behavior at no social cost, for whenever an individual pollutes (because the benefits from doing so are higher than harm), he merely pays for harm, which causes no social cost, as his payment represents merely a transfer of command over resources. But if an individual is jailed for having polluted when the disutility of jail equals the harm from pollution, then although his polluting behavior is desirable (by assumption his benefits from so doing are higher than the harm generated), this form of sanction absorbs social resources.

1.3 Fault-based liability for harm. Under this rule, a person is subject to liability for harm if his act was undesirable, but is not held liable if his act was desirable. Hence, if the sanction for causing harm is sufficiently high, undesirable behavior will be deterred, whereas desirable behavior will not be discouraged because it will not result in punishment. An individual who would obtain a benefit of less than 1,000 from an act that causes harm of 1,000 --

[26] To amplify, let g be the gain from committing an act that causes certain harm of h, let $f(g)$ be the probability density of g in the population, let s be the sanction, let $d(s)$ be the disutility of s to individuals, and let ks be the additional social cost of imposing the sanction s, where $k > 0$. Under strict liability, ideal behavior -- commission of the act if and only if g is at least h -- can be induced if $s = h$. If so, social welfare equals

$$W = \int_h^m (g - h - (h + kh))f(g)dg,$$

where m is the maximum gain from committing the act. The ideal level of social welfare is not achieved because of the term $-(h + kh)$, which are the total costs associated with imposition of punishment.

[27] Using the notation of the previous note, the optimal magnitude of the sanction is the s that maximizes

$$W = \int_s^m (g - h - (s + ks))f(g)dg.$$

Setting the derivative of W with respect to s equal to 0 gives the first-order condition for the optimum, $(h + ks)f(s) = (1 - F(s))(1 + k)$, where F is the cumulative distribution of f. The interpretation of this condition is that the marginal net benefit from deterrence equals the marginal cost. From this condition, it is apparent that the optimal s could be above h (reflecting the fact that the harmful act involves social costs of not only h, but also $s + ks$, so exceeding h) or below h (reflecting the fact that social costs of punishment can be reduced by lowering s). The solution to this problem is discussed in Polinsky and Shavell 1984 and Kaplow 1990.

and thus for whom the act would be socially undesirable -- will not commit the act if the sanction is sufficiently high; but an individual who would obtain a benefit exceeding 1,000 from the act -- and thus for whom the act would be socially desirable -- would commit the act because he would not be held at fault and punished for so doing. Thus, ideal behavior is achieved under fault-based liability without the actual imposition of socially costly punishment.

A corollary point is that the optimal magnitude of the sanction for a socially undesirable act is any sanction sufficient to deter. It does not matter how high the sanction is, for because the threat of sanctions deters, sanctions are never applied, and hence higher sanctions do not result in higher social costs.

An important factor should be added: There is sometimes a possibility that an individual cannot be deterred from committing an undesirable act because his benefit exceeds even the maximal sanction (such as life imprisonment). In this situation, it is optimal not to impose any sanction on the individual even though his act is socially undesirable. For by hypothesis, all that imposing a sanction would create is a social cost; it would not accomplish deterrence of the individual, for that is by hypothesis impossible. For instance, suppose that the highest sanction that can be imposed on a person is 100 (because, say, imprisonment would not create such great disutility for him), and that he would obtain a benefit of 200 from committing the act causing harm of 1,000. Then, although his act is undesirable, it is optimal not to punish him.[28]

The conclusion is that, under optimal fault-based liability, sanctions are never imposed. They are not imposed if individuals act desirably, and their use is threatened when and only when that threat will successfully deter undesirable behavior. In sum, ideal behavior is achieved, except when deterrence is impossible, and it is achieved without the bearing of costs of actually imposing sanctions.[29]

This conclusion about optimally applied fault-based liability will be important to bear in mind in what follows. It should be emphasized that the point that sanctions are never imposed depends on the implicit presumption that the information of the social authority is perfect. In particular, the social authority has to know the benefits that individuals obtain not only to be able to determine which acts are desirable and which not, but also to be able to forecast when imposition of a sanction would deter and when not.

[28] In the previous chapter on monetary sanctions, I did not emphasize the point analogous to the one here -- that deterrence of undesirable acts might be impossible because the assets of a person might be limited. However, in the case of monetary sanctions, there is no advantage of relieving an impossible-to-deter person of liability, for the assumption is that imposing a monetary sanction does not involve a social cost. That is why the situation where individuals cannot be deterred was not a focus of discussion where sanctions are monetary, but it is significant here.

[29] Let me be precise about fault-based liability as discussed in this section. Under such liability, any act that is desirable -- such that $g \$ h$ -- is not sanctioned, and hence individuals commit such acts and are not punished. If an act is undesirable -- such that $g < h$ -- then, if there is an s exceeding g for the individual, he will be sanctioned for committing the act with such an s, and thus will be deterred. But if there does not exist an s for the person such that $s > g,$ then he cannot be deterred, so that it is optimal to set $s = 0$ for that person (otherwise he will commit the act and social costs will be $h + ks$ rather than just h). In sum, the formula for the optimal sanction under the fault rule, is as follows. Let $g,$ h, and m be the gain, harm, and maximal sanction that can be imposed on an individual. Then the optimal sanction $s = s(g,h,m)$ is apparent: if $g \$ h$, then $s = 0$; if $g < h$ and $m < g$, then $s = 0$; if $g < h$ and $m \$ g$, then $s \$ g$. Therefore, all individuals whose acts would be desirable commit them, all those who can be deterred from committing undesirable acts are deterred, and no one actually suffers punishment.

Fundamental advantage of fault-based liability over strict liability. Fault-based liability is different from strict liability because, under fault-based liability, deterrence of undesirable acts is achieved when it can be and without the imposition of sanctions on those who commit socially desirable acts. This feature of fault-based liability constitutes an advantage over strict liability when sanctions are socially costly to impose. As has been noted, however, use of fault-based liability does mean that the social authority needs greater information than it would to apply strict liability, which requires only information about the harm done.

1.4 Fault-based liability continued: when information of the social authority is imperfect. Now let us consider the situation in which the social authority's information is imperfect and the authority may err in assessing a person's benefits or how harmful his act was. There are several consequences of such errors.

First, some desirable acts may not be committed. This is because a person might fear that his desirable act would be erroneously seen as undesirable and that he would bear a sanction greater than his benefits. (A lost hiker might not enter an unoccupied cabin to phone for help because of fear that he would be sanctioned for breaking into a property.)

Second, some individuals who could have been deterred from committing undesirable acts will not be deterred. The social authority may believe that for a particular kind of act and person, a sanction of 500 would successfully deter, but in fact it does not, and thus the person commits the act even though a higher sanction would have deterred him.

Third, sanctions will actually be imposed and society will thus incur the costs associated with punishment. Sanctions will be imposed for a variety of reasons: Some who commit desirable acts will erroneously be sanctioned; some who commit undesirable acts and could have been deterred will be sanctioned by too low a sanction, as just discussed; and some who commit undesirable acts and could not have been deterred by any sanction will mistakenly be punished.

The optimal sanction will be chosen taking into account these various consequences of imperfect information. To understand the nature of the optimal sanction, consider an example.

Example 1. There are three types of parties: As who obtain a benefit of 500 from committing an act that causes harm of 100; Bs who obtain a benefit of 40 from committing the act; and Cs who obtain a benefit of 70 from committing the act. Assume that the maximum feasible sanction creates disutility of 50. Hence, the situation is that for As the harmful act is desirable (the benefit exceeds the harm); for Bs and Cs, the act is undesirable (the harm exceeds the benefit); and Bs can be deterred by a sanction of 40 or more, but Cs cannot be deterred by any feasible sanction.

If the social authority possesses perfect information, its optimal policy is clear: The authority will not sanction As, as their act is desirable; it will announce a sanction of at least 40 for Bs and thus deter them from acting (so it will not turn out to impose a sanction on them); and it will impose no sanction for Cs, as deterring them is impossible. Thus, As and Cs will commit harmful acts, Bs will be deterred, and no sanctions will actually be imposed.

Suppose, however, that the social authority has only imperfect information, in that it cannot distinguish between Bs and Cs.

What are optimal sanctions in this case? Clearly, As will not face a sanction and will commit the act, for As can be identified by the authority. However, because the authority is unable to distinguish Bs and Cs, they will necessarily face the same sanction.[30] If the sanction for them is 40, Bs will just be deterred, but Cs will not be deterred and will commit the act and suffer the sanction, resulting in the bearing of social costs. Any sanction above 40 will also deter the Bs and will result in the Cs bearing a higher sanction, so would be socially inferior to a sanction of 40. Any positive sanction below 40 will not deter either Bs or Cs, but will be imposed on both, so would be inferior to not imposing any sanction. Hence, what is optimal is either a sanction of 40, the minimum sanction that can deter the Bs, or no sanction at all. Which of these two possibilities is best depends on, among other things, the relative numbers of Bs and Cs. If Bs are sufficiently more numerous than Cs, the optimal sanction will be 40, as the deterrence of all the Bs will be worthwhile even though Cs will commit the act and suffer sanctions of 40; whereas if Cs are sufficiently numerous, a sanction of 0 will be best, as there are relatively few B's who can be deterred, and to deter them means imposing a sanction of 40 on all the undeterrable Cs.//

As is demonstrated in this example, because the use of sanctions does result in their actual imposition, the optimal level of sanctions is, in a rough sense, the lowest sanction that will achieve deterrence of the group who can be deterred, if that group is worth deterring in view of the sanctions that those who will not be deterred will then suffer.[31] With this in mind, let us examine further the determination of optimal sanctions.

Relationship of optimal sanctions to individual benefits from acts and to the magnitude of harm. The general nature of the determination of the optimal magnitude of sanctions, as discussed above, involves the social benefit of deterrence on one hand and the costs associated with imposition of sanctions (because of error relative to the ideal on the part of the social authority) on the other. What is the answer to the more specific question concerning the relationship between optimal sanctions, the benefits obtained by a person, and the harm? The answer is as follows.

The higher the benefits to a person contemplating a harmful act, the higher should be the sanction, for higher benefits require higher sanctions to deter. The person who would kill in

[30] If the state announced different sanctions for the two types (such as 40 for Bs and 0 for Cs), then individuals of the type who would suffer the higher sanction would claim to be of the other type (Bs would claim to be Cs), so in effect there would be just one sanction for both types, namely, the lesser of the two sanctions.

[31] This is only an approximate statement of the principle guiding the choice of the optimal sanction, because in reality the choice is more complicated than in the example. Among other things, there will not usually be a single, well-defined group who can be deterred -- the Bs in the example -- and another group -- the Cs in the example -- who cannot be deterred. For instance, suppose in the example that there is another group of Ds who obtain a benefit of 45 from committing the harmful act, so that the Ds as well as the Bs can be deterred. In this instance, it might well be desirable to employ a sanction of 40, and thus to deter the Bs but not to deter the Ds even though they can be deterred; that would be desirable if the B's were very numerous, the Ds very small in number, and the Cs modest in number. More generally, there will be a continuum of types of parties, and many different degrees of lack of information that the courts may suffer from. Nevertheless, the example captures the important compromise that the optimal sanction typically reflects -- the tradeoff between greater deterrence of some, and the greater suffering of sanctions by others, and thus the incurring of social costs in respect to them.

Formally, the choice of the optimal sanction under the fault rule is as follows. Given any set of observable characteristics of parties who come before it, the social authority can formulate a probability distribution of gains of parties. Then it can choose an optimal sanction given this probability distribution. The determination of that optimal sanction is essentially as described in note 5.

order to obtain a great deal of money may be harder to deter than the person who would kill to obtain a small amount of money or to satisfy a grudge. There is, however, a limit to this relationship: If the benefits become so great that deterrence may not be possible, then the sanction should fall (and to zero if deterrence is impossible for all persons to whom the sanction would apply).

Also, the higher the harm, the higher should be the sanction, for higher harm means that more is gained by deterring, so society should be willing to incur greater costs in actually imposing sanctions in order to achieve greater deterrence. Other things being equal, we should be willing to bear greater costs, associated with imprisoning people, to deter murder than assault, for murder is more harmful.

There is another reason why higher harm may sometimes call for higher sanctions: If the object of an individual is to harm someone, then higher harm will imply higher benefits. If higher harm is associated with higher benefits, it may be desirable to increase the sanction for higher harms because this means that deterrence of individuals is more difficult.[32]

Comment: comparison of optimal sanctions when sanctions are monetary. In the present setting, the magnitude of sanctions is chosen balancing deterrence benefits against the costs associated with actually imposing sanctions. Thus, an important theme has been that it is undesirable to impose sanctions that are higher than is likely to be needed to accomplish deterrence, so that those who are not deterred are punished as little as possible. When sanctions are monetary and assumed to be costless to impose (or, more realistically, are costly but significantly less so than imprisonment), the social need to limit the magnitude of sanctions is much lower, and the optimal sanction generally equals the harm. Thus, when sanctions are monetary, there is no need, in principle, to identify the strength of the benefit from an act or the motive for it, in order to determine the proper sanction since only harm need be measured.[33]

Comment: *realism of the assumption that the social authority's information is imperfect and that the actual imposition of sanctions is inevitable.* It is apparent that the case of imperfect information is the realistic case to consider, for social authorities cannot practically always know which acts are undesirable and which not, and who can be deterred by which sanctions and who cannot be deterred. It is evident as well that deterrence will frequently be impossible to achieve. There will often be individuals who cannot be deterred from committing an act no matter how high the sanction is, especially because, as will be discussed, the probability of the sanction will often be low. Moreover, if we depart momentarily from our model, in which individuals always calculate benefits against expected penalties, we know that individuals may suffer momentary lapses of control and not calculate, at least when the sanction will not be immediate and certain, so these individuals in these circumstances will be effectively undeterrable. When the social authority cannot determine who is undeterrable, these individuals will often bear sanctions,

[32] Additional reasons for sanctions to increase with harm are discussed in sections 2 and 5 of chapter 22, on marginal deterrence and general enforcement.

[33] This point is subject to the qualifications discussed earlier, such as that if parties are risk averse and not insured against sanctions, the proper sanction may be somewhat lower than harm, but the central point emphasized in the text remains true: That when sanctions are very cheap to impose socially, the optimal sanction tends to equal the harm, whereas when they are very expensive to impose, as they are when the sanction is imprisonment, the optimal sanction is quite different, and is limited to that necessary to deter those most important to deter.

creating social costs.

1.5 Act-based liability. The main points just discussed carry over to act-based liability. The strict form of such liability, under which sanctions would be imposed for committing an act, is clearly inferior to fault-based liability, under which sanctions would be imposed less often (and not at all if the social authority has perfect information). The main difference between act-based liability and harm-based liability is that the problem of inability to deter may be greater under harm-based liability. The reason, as was explained in the previous chapter, is that if harm from an act occurs only with a probability, then under harm-based liability, the sanction will be applied only with a probability. (A person who shoots at another will bear a sanction only if he hits his intended target.) As a consequence, the magnitude of the sanction necessary to deter will be larger under harm-based sanctions than under act-based sanctions (under which a person will bear a sanction for shooting, even if he misses). Thus, act-based liability may be superior to harm-based because it achieves deterrence with lower sanctions. On the other hand, the social authority may experience difficulty in assessing the expected harm from an act, whereas under harm-based liability, it at least knows that the act has generated the observed harm.[34]

1.6 Conclusion. What can be concluded from our discussion of the basic theory of liability when sanctions are nonmonetary? One of the main conclusions is that fault-based liability has appeal over strict liability, for under fault liability, deterrence of undesirable acts can be created with less actual imposition of sanctions, and in the ideal -- when the courts have perfect information -- with no imposition of sanctions. Another conclusion is that the theory of the determination of the optimal magnitude of the sanction under fault-based liability can be understood only by recognizing the lack of information of the social authority about parties' benefits from acts, the harmfulness of acts, and the possibility of deterrence. For only by taking the authority's lack of information into explicit account can it be explained why sanctions are ever imposed (and thus why there is a need to limit their magnitude). Additionally, these conclusions about the advantage of the fault system and the determination of optimal sanctions derive from the assumption that nonmonetary sanctions are costly to impose.

2. The Optimal Probability and Magnitude of Nonmonetary Sanctions[35]

2.1 Introduction. The theory here parallels that in the last chapter, so the analysis can be relatively brief. The question under consideration concerns, again, the choice of the probability and the magnitude of sanctions when account is taken of the cost of maintaining the probability of sanctions.

2.2 Behavior given the probability and magnitude of sanctions. A person may display risk-neutrality toward prison sentences and, for instance, be equally deterred by a certain one-year sentence and a 50 percent probability of a two-year sentence. This is the way a person will regard sanctions if the disutility of imprisonment is proportional to its length.[36]

[34] I will return to the issues discussed in this section when criminal attempts are considered in section 4 of chapter 24.

[35] The points made in this section are largely developed in Shavell 1985b, 1987a.

[36] Let s be the length of the prison sentence and $d(s)$ its disutility. Then the assumption of risk neutrality is that $d(s)$ = "s for some positive ", and for simplicity, I will often assume that " = 1.

Risk aversion. However, individuals may be risk averse with regard to imprisonment, and be more deterred by a 50 percent probability of a two-year sentence than by a certain one-year sentence. (In general, risk-averse individuals will be more deterred the greater the uncertainty in the sentence, its expected length held constant.) Individuals will be risk averse if the disutility of imprisonment rises more than in proportion to its length.[37] This could be so because of increasing yearning to join the functioning world or growing distaste for the prison environment as the time spent in prison increases.

Risk preference. Another possibility is that individuals are risk-preferrers, and would thus find a certain one-year sentence worse than a 50 percent chance of a two-year sentence. Individuals will prefer risk if the disutility of imprisonment to them rises less than in proportion to its length.[38] That would be so if, over time, imprisonment matters less as a person becomes accustomed to prison life and makes his adjustment, or if he discounts the future disutility of imprisonment. It would also be true if he experiences relatively large disutility from being in jail at all, due to humiliation and the stigmatizing effect of having been in prison for any length of time, or due to brutalization in the beginning of imprisonment.[39]

Probability versus magnitude of sanctions. The analogue of what was said in chapter 20 about the importance of the probability versus the magnitude of sanctions is true here. Namely, if a person is risk neutral regarding imprisonment, then a given percentage increase in either the probability or the magnitude of such sanctions will have the same effect on behavior. If a person is risk averse, then a given percentage increase in the magnitude of sanctions will have a greater effect than an equal percentage increase in the probability of sanctions. However, if a person is risk-preferring, a given percentage increase in the probability of sanctions will have a greater deterrent effect than the same percentage increase in the magnitude of sanctions.[40] It should also be noted that the general comments made about perceptions of sanctions and their likelihood in the previous chapter apply here; it is the perceived rather than the actual sanctions that determine deterrence.[41]

[37] Suppose, for example, that the disutility of the first year of imprisonment is 100 and that of the second is 200. Then the disutility of a certain one-year sentence is 100, and the expected disutility of a 50 percent chance of a two-year sentence is 50% H (100 + 200) = 150, so the individual will be more deterred by the latter. Formally, using the notation of the previous footnote, the assumption of risk-aversion is that $d\mathbb{N}(s) > d(s)/s$.

[38] Suppose, for instance, that the first year of imprisonment involves disutility of 100 and the second involves disutility of only 50. Then the disutility of a certain one-year sentence is 100, and the expected disutility of a 50 percent chance of a two-year sentence is 50% H (100 + 50) = 75, so the individual will be more deterred by the former. Formally, the assumption of risk-preference is that $d\mathbb{N}(s) < d(s)/s$.

[39] I did not discuss the possibility of risk preference with respect to monetary risks because it does not seem an important possibility, whereas risk preference with respect to imprisonment risks seems often to be descriptively accurate.

[40] To show what was just said in this paragraph, let p be the probability of the sanction, so $pd(s)$ is the expected disutility ED. We want to compare the elasticity of ED with respect to p -- namely, $(p/ED)(dED/dp)$ -- with its elasticity with respect to s, namely, $(s/ED)(dED/ds)$. In the risk-neutral case, $ED = pks$, so that $(p/ED)(dED/dp) = (1/ks)(ks) = 1 = (1/kp)(kp) = (s/ED)(dED/ds)$. In the risk-averse case, $(p/ED)(dED/dp) = (1/d(s))(d(s)) = 1$ and $(s/ED)(dED/ds) = (s/pd(s))(pd'(s)) = s(d'(s)/d(s))$. And $s(d'(s)/d(s)) > 1$, for the assumption of risk aversion is that $d'(s) > d(s)/s$. In the case of risk-preference, the argument is analogous to that in the risk-averse case.

[41] On perceptions of sanctions and their likelihood, see the references cited in note 10 of chapter 20.

2.3. Comment: observed behavior; deterrence in fact. What can be said about the actual behavior of individuals in the face of sanctions? A multitude of observations from everyday life suggests that individuals are discouraged from all manner of undesirable behavior when the likelihood and magnitude of sanctions is sufficiently high: Drivers slow down and tend to obey traffic rules when they see a police car; students' deportment improves under a teacher's gaze; criminals often refrain from acting when they would be easy to identify as responsible. Various events that result in gross changes in expected penalties have been noted to influence the incidence of violations of law; for example, police strikes have resulted in marked increases in crime, improvements in toxicology have led to declines in the incidence of poisoning, and increases in tax audit rates and sanctions have discouraged tax evasion.[42] In general, there is a great weight of empirical evidence demonstrating that increases in expected sanctions reduce violations.[43] However, some studies have questioned the interpretation of these results, and also have found relatively small effects of changes in the probability and magnitude of sanctions on behavior. Such findings may, in part, be due to inaccurate perceptions of expected sanctions, to violators' discounting of future imprisonment, and to subtle but important statistical problems.[44]

2.4 Optimal sanctions when the probability of their imposition is a given. What is the optimal magnitude of sanctions, given their probability? The rough answer is that whatever would be optimal if sanctions were certain should be inflated when sanctions are applied with a probability. If individuals are risk neutral, sanctions should be multiplied by the reciprocal of the probability of sanctions, so that the expected sanction is what it would be in a world with certain sanctions. Thus, where the optimal sanction would be two years imprisonment with certain sanctions, the optimal sanction would be six years if sanctions are applied with a probability of one third. If individuals are risk averse, optimal sanctions tend to be lower than otherwise, and if they are risk preferring, higher than otherwise.[45] However, it must be borne in mind that the

[42] On the effects of police strikes and of advancement in toxicology, see Andenaes 1966, 961-62; on the effects of tax auditing and penalties, see Andreoni, Erard, and Feinstein 1998.

[43] See, for example, the surveys Cook 1977, Ehrlich 1996, Eide 2000, and Glaeser 1998; and see also, for example, Andenaes 1975, Kessler and Levitt 1999, Levitt 1996, 1997, 1998a, 1998b, Viscusi 1986b, Wilson and Herrnstein 1985, chapter 15, and Witte 1980.

[44] On weak findings concerning deterrence, and methodological criticisms of studies of deterrence, see Andenaes 1975, Blumstein, Cohen, and Nagin 1978, Cook 1977, Ehrlich 1996, 56-63, and Eide 2000, 364-68. Of the methodological criticisms, two stand out. First, many studies do not take into account that imprisonment reduces crime due both to deterrence and incapacitation (thus, if crime falls due to an increase in imprisonment, the decline cannot be ascribed entirely to deterrence). Second, sanctions may not only influence crime, but be influenced by it, obscuring statistical findings (jurisdictions with high crime rates might raise levels of punishment to counter their problem; this would result in a positive correlation between high penalties and crime, but would not imply that high penalties fail to deter).

[45] It is possible that optimal sanctions would not change in the stated way, even though one would expect them to. For example, it is possible that risk aversion could increase the optimal sanction. Suppose that if individuals are risk neutral, the optimal sanction is zero, because the six-year sanction needed to deter those who can be deterred would result in excessive social costs from imposition of sanctions on those who cannot be deterred. However, if individuals are risk averse, the optimal sanction might be positive, because only a four-year sanction would be needed to deter those who can be deterred, and the social cost from actual imposition of sanctions on those who cannot be deterred is smaller and

appropriate probability-inflated sanctions may not be feasible to apply (a sentence of 100 years cannot be imposed), so in general it will not be possible to duplicate the deterrence that would be best to achieve were sanctions certain.

2.5 Optimal sanctions when the probability of their imposition is also optimally determined. As in the case of monetary sanctions, the basic insight applies here that, when the state chooses both the probability of imposing sanctions and their magnitude, a low probability-high sanction policy is often socially advantageous. The reason is again that a social savings can be achieved by conserving enforcement effort, while the magnitude of sanctions can be raised to offset the low probability and thereby to avoid dilution of deterrence.[46] Unlike the case with monetary sanctions, however, raising nonmonetary sanctions in itself raises social costs when the sanctions are imposed. Yet because sanctions are imposed less often, total costs of sanctions may not rise. This suggests what I will now elaborate upon, that when individuals are risk neutral, the optimal policy involves low probabilities and maximal sanctions (interpreted, perhaps, as life imprisonment). I will also explain that the conclusion is the same when individuals are risk averse, but when they are risk-preferring, less than maximal sanctions are often best.[47]

In the risk-neutral case, suppose that the sanction is less than maximal. For instance, suppose that the sanction is five years of imprisonment, that the maximal sanction is 20 years, and that the probability of sanctions is 40 percent. Now raise the sanction to 20 years and reduce the probability of sanctions to 10 percent. The reduction in the probability from 40 percent to 10 percent will save enforcement costs, a social benefit, and nothing else will change. First, the behavior of individuals will remain the same because the expected sanction will remain equal to two years -- 40% × 5 = 10% × 20. Second, the social costs of imposing sanctions will also be constant, for the expected sanction will remain equal to two years, that is, the number of person-years spent in jail will be unchanged. Although the expense of imposing a sanction on each person who is punished rises by a factor of four (the sentence rises from 5 years to 20 years), the number of people sanctioned falls by a factor of four (the probability falls from 40 percent to 10 percent).[48]

As to what probability is best, the answer to this question reflects two general

worthwhile for society to bear. Thus, the statement in text refers only to a general tendency.

[46] However, this basic point will be qualified when incapacitation is considered, for a too-low probability will undesirably lower incapacitation. See chapter 23, section 1.4(e).

[47] The general problem of choosing the probability p and magnitude of sanctions s optimally is to maximize

$$W(s) = \int_{pd(s)}^{m}[g - h - (pd(s) + pks)]t(g)dg - c(p).$$

Here $t(g)$ is the probability distribution of g conditional on the information that the social authority possesses about the act, under the relevant liability rule. Also, it is assumed that p can be independently chosen for this distribution $t(g)$. Note that the term $(pd(s) + pks)$ corresponds to the costs of imposing sanctions: the expected disutility experienced by a person who commits the harmful act, $pd(s)$, plus the expected public costs due to imprisonment, pks.

[48] Referring to note 25, the argument in this paragraph can be restated as follows. Assume s is less than the maximal sanction m, and raise s to m and lower p to p' so that $ps = p'm$. Then the lower limit of integration, $pd(s) = ps$ does not change, nor does the integrand, so that the integral is constant, but $c(p)$ falls to $c(p')$, raising W.

considerations. First, the higher the probability, the greater are enforcement costs, so that, as emphasized in the last chapter, it will generally be best for society to tolerate some degree of underenforcement in order to save enforcement resources. Second, the probability should be chosen so that the expected sanction leads to the appropriate tradeoff between actually imposing sanctions and achieving deterrence, as explained in section 1.4 above.

Risk-averse case.[49] If individuals are risk-averse regarding imprisonment sanctions, then the result that optimal sanctions are maximal is reinforced. In the example just discussed, if the imprisonment term is increased to 20 years, the probability of sanctions at which a risk-averse person would be equally deterred as before would be lower than 10 percent, such as 5 percent. This means not only that enforcement cost savings would be greater, but also that there would be a savings in the cost of imposing sanctions, as expected person-years in jail would fall from two years to one year.[50] (Note that this conclusion that risk aversion reinforces the result that optimal sanctions are maximal is opposite to the case in chapter 20, in which risk aversion regarding monetary sanctions leads to the optimality of less than maximal sanctions.[51])

Risk preference. If, however, individuals are risk-preferring with regard to imprisonment sanctions, the optimal sanction is not necessarily maximal. For the behavior of risk-preferrers to be the same when the magnitude of the sanction is raised from five years to 20, the probability of sanctions cannot fall to as low as 10 percent, it must be at a higher level, say 15 percent. But this means that enforcement cost savings are less than in the risk-neutral case and also that there is an increase in the cost of imposing sanctions (expected person-years in jail rise from two years to three years). Hence, it is possible that deterrence can be more cheaply achieved with a strategy of use of fairly probable sanctions that are not maximal. Another way to express this point is that raising the sanction from five to 20 years raises social costs of actually imposing sanctions fourfold for any person who is caught, but it may not increase the deterrence very much because the disutility of sanctions rises less than in proportion to their magnitude. (The person who dislikes imprisonment because of the brutalization and stigma that come from being imprisoned at all will not be four times as deterred by a 20 year sentence as by a five year sentence.) Hence, within some range, raising the magnitude of sanctions may be a less economical way of achieving deterrence than raising the likelihood of sanctions.

2.6 Comment on the false notion that willingness to face the sanction for an act implies that committing it is socially correct. In the present context of use of the nonmonetary sanction of imprisonment and fault-based liability, the idea that it is socially desirable for a person to commit a harmful act if he is willing to do so and bear the risk of being sanctioned -- if his gain exceeds the expected disutility of the sanction -- must be regarded as generally

[49] The conclusions to be stated about risk aversion and risk preference are presented in Polinsky and Shavell 1999.

[50] The general argument made is as follows. If $s < m$, then raise s to m and lower p to p' such that $pd(s) = p'd(m)$. Because of risk aversion, $p' < p(s/m)$. Now given p' and m, the lower limit of the integral in note 25 is unchanged, and the integrand falls because pks falls to $p'km$ (for $p'km < p(s/m)km = pks$). Moreover, $c(p)$ falls to $c(p')$. Hence W rises.

[51] The difference can be explained as due to a difference in the implication that expected sanctions fall when the probability is lowered so as to hold behavior constant: Here the decline in expected sanctions is socially desirable, as it reduces the costs of imposing sanctions, which can be interpreted as lowering taxes; in the last chapter, the decline in expected sanctions meant that sanction revenue fell and that taxes rose.

mistaken.

One reason is that emphasized in the previous chapter and noted again here: That because of the expense of catching violators of law, it will be best to save enforcement resources and to countenance underdeterrence relative to ideal deterrence. This factor of the optimality of underdeterrence is of greater significance where the cost of catching violators is high, and that may characterize the typical context of crimes, for which imprisonment sanctions are employed. Hence, to a person who says that he was willing to commit an act, such as stealing money, because his gain outweighed the expected sanction, the response could be that the expected sanction was not as high as society would have wished due to the social cost of raising it, especially through raising its likelihood, so that his decision was hardly in the social interest.

A second mistake in the notion that it is socially desirable for a person to commit an act if he is willing to do so concerns the point that fault is the form of liability that we are assuming applies. Because liability is premised on fault, any liable act that is committed is, prima facie, socially undesirable. Hence, if a person commits an act for which he would be held liable, the most likely explanation for that having occurred, presuming that deterrence was possible, is that the social authority did not gauge properly the magnitude of sanction needed to deter, not that the act was desirable. If a person who faces a sanction of three years of imprisonment for theft proceeds to steal but could have been deterred by the threat of a six year sanction, the likely explanation is that the social authority did not realize that discouraging this type of individual from theft required a six year sanction. (Recall the discussion in section 1.4, which explained why, in the face of imperfect information, the social authority may choose a sanction that is not sufficient to deter some individuals even though they could have been deterred.) Only if the person was found at fault but in fact committed a socially desirable act (as in the case of the lost hiker breaking into a cabin) that the state did not properly evaluate would the proper interpretation of the person's willingness to commit an act for which he is sanctioned under fault-based liability be that it was socially good that he did so. For all the other reasons given here, the usually appropriate interpretation would be that the act was undesirable to commit.

3. When Nonmonetary Sanctions Are Optimal to Employ

3.1 In general. It has been supposed that monetary sanctions are socially costless to impose, whereas nonmonetary sanctions are socially costly to employ. Under these assumptions, nonmonetary sanctions are inferior to monetary sanctions and thus should not be used unless monetary sanctions alone cannot adequately deter. The latter in turn will be true when the expected sanctions that can be created with solely monetary sanctions are low relative to the harm that the sanctioned acts generate. In such circumstances, nonmonetary sanctions will be needed to create added deterrence and may be advantageous to use if the harm of the acts that would otherwise be committed is large enough.[52]

3.2 Factors bearing on the optimality of use of nonmonetary sanctions. Several factors are relevant to the desirability of utilizing nonmonetary sanctions. The first three to be mentioned bear on the likelihood that monetary sanctions will not be sufficient to deter and

[52] The point that the nonmonetary sanction of imprisonment is more expensive than monetary sanctions and thus, by implication, should not be employed unless monetary sanctions will not function adequately, was made by Bentham [1789] 1973 and emphasized by Becker 1968.

therefore that use of nonmonetary sanctions will be desirable.[53]

Level of assets. If the assets of parties are low relative to the magnitude of the sanction necessary to deter, then deterrence will tend to be insufficient if only monetary sanctions are employed. If a person's wealth is at most a few hundred dollars, then it would be difficult to deter him from committing acts that yield even modest benefits using solely monetary sanctions.

Probability of escaping sanctions. The greater is the likelihood of escaping sanctions, the greater is the magnitude of the sanction necessary to achieve deterrence, and thus the more likely this sanction is to exceed the assets of a person. Thus, even if a person's assets are not insubstantial, deterrence may become impossible to achieve if the probability of imposing the sanction is sufficiently low.

Level of private benefits obtained from an act. The larger are these benefits, the greater is the sanction needed to deter, and again the more likely it is that the necessary sanction will exceed a person's assets.

The expected harm due to the act committed. The larger the expected harm due to an act -- the higher the probability of harm and its potential magnitude -- the more important the act is to control, that is, the greater are the consequences of failure to deter it when that is desirable. Hence, other things being equal, the greater the expected harm from an act, the more likely it will be advantageous to use nonmonetary sanctions to deter.

*Comment***.** The factors just mentioned will be considered in chapter 24 to argue that in the core area of criminal law, monetary sanctions would be grossly inadequate for the purposes of deterrence, and thus that nonomonetary sanctions are warranted.

4. Joint Use of Nonmonetary and Monetary Sanctions

4.1 Nonmonetary sanctions should be used only as a supplement to maximal monetary sanctions. It was just explained that nonmonetary sanctions are needed to deter when monetary sanctions would not be adequate for that task. An aspect of this conclusion is that nonmonetary sanctions should not be employed unless monetary sanctions have been imposed to the greatest possible extent, which is to say, unless the monetary sanction equals the entire wealth of a party. Otherwise, the same level of deterrence could be accomplished at lower cost by increasing the monetary sanction. For example, suppose that a person faces a sanction of $20,000 and two years in prison for an act, but has $30,000 in assets. This cannot be optimal because society could construct a sanction that involves equivalent deterrence by increasing the monetary sanction to $30,000 and by decreasing the prison term somewhat; perhaps a $30,000 sanction and a term of one year would be equivalent to the $20,000 sanction and a term of two years. Such a sanction of $30,000 and one year, involving as it does less imprisonment, is cheaper for society. Thus, in general, it cannot be optimal for society to be using imprisonment unless society has already employed the monetary sanction to the utmost.[54]

[53] These and related factors are discussed in Shavell 1985b.

[54] See Polinsky and Shavell 1984.

4.2 Implications of the above conclusion that nonmonetary sanctions should only be used as a supplement to maximal monetary sanctions. There are several implications of the point just explained.[55]

Wealth and optimal sanctions. One implication concerns the specific nature of the relationship between a person's wealth and sanctions. If an individual's wealth is above the threshold at which deterrence with monetary sanctions will be adequate, the sanction should be entirely monetary. If his wealth is less than this threshold, the sanction should equal the person's entire wealth and should be accompanied by a nonmonetary sanction. Moreover, the lower is the level of a person's wealth below this threshold, the higher should be the nonmonetary sanction, so that the total sanction, reflecting the person's wealth plus the nonmonetary sanction, is maintained at the appropriate magnitude.

Harm and optimal sanctions. Another implication of the conclusion under discussion is that if the harmfulness of an act is below a certain threshold, then monetary sanctions alone will be enough to deter adequately. Once the expected harm surpasses this threshhold, however, it will be desirable for nonmonetary sanctions to accompany the monetary sanction, which will be maximal.

4.3 The above conclusion about nonmonetary sanctons and the possibility of costly monetary sanctions. As noted in the previous chapter, the imposition of monetary sanctions is not in fact socially costless, for assets need to be located and collected and the liable parties may hide assets or resist collection. This reduces the advantage of monetary over nonmonetary sanctions. Yet one presumes that the sanction of imprisonment would still usually be more expensive as a deterrent than money sanctions, in which case the conclusion that monetary sanctions should be imposed to their limit before imprisonment is imposed continues to apply. Nevertheless, that might not always be the case. Having to spend a day or two in jail might serve as a deterrent just as well as a fine of $10,000 for a fairly wealthy person, but be cheaper than collecting $10,000 from such a person.

[55] In addition to the two points to be discussed, another of note is that, when monetary and nonmonetary sanctions are employed together, it is not necessarily optimal to employ maximal nonmonetary sanctions. This can be understood by seeing why the type of argument given in section 2.5 for the optimality of maximal nonmonetary sanctions does not carry forward. For example, suppose that an individual faces a maximal monetary sanction of $10,000 and a prison term of five years, and that this will be imposed with probability 10 percent; thus the expected penalty is $1,000 plus .5 years of imprisonment Now double his prison sentence to 10 years and reduce the probability so that the expected penalty is the same. The new probability cannot be as low as five percent, for if it were five percent, the expected penalty would be $500 plus .5 years of imprisonment -- the reason being that when the probability falls, the expected monetary component of the penalty falls (because the monetary sanction cannot be raised above $10,000). Hence, the probability that maintains deterrence must be higher than five percent, such as seven percent. But this means that the expected number of person-years in jail will be more than .5 years, so that the social costs of imposition of imprisonment rise. For this reason, it does not follow that optimal imprisonment is maximal (although optimal monetary sanctions are maximal). For details, see Shavell 1991b.

5. Different Types of Nonmonetary Sanctions

5.1 The variety of nonmonetary sanctions. There are a variety of nonmonetary sanctions apart from imprisonment. There are, first of all, a number of nonmonetary sanctions that involve corporal punishment, notably, whipping, branding, and the death penalty. There are also various sanctions that restrict freedom, apart from imprisonment. These include requirements to live in half-way homes, restriction to one's residence, and other probationary restraints on conduct. Third, there are sanctions designed to humiliate and shame, such as publication of the names of individuals who have violated a law (for example, those who have hired prostitutes). Indeed, most nonmonetary sanctions may have a component of humiliation; criminal violations are usually matters of public record, so that the reputations of criminals are forever tainted.

5.2 Cost, disutility, and effectiveness of nonmonetary sanctions. The various sanctions differ in social cost and in the disutility they create. We might define the *effectiveness* of a sanction to be the disutility it generates per dollar of social cost. By this definition, some sanctions might be significantly more effective than imprisonment for certain people. For example, the humiliation of individuals who especially value their reputations that would be accomplished by publishing their names and their violations might serve as a significant deterrent; and because such humiliation is cheap for society (among other things, it does not remove individuals from the labor force), it might rank high in effectiveness. Similarly, confinement to one's residence might serve as a highly effective deterrent for some because it would prevent them from participating in much of free life but would not involve use of prisons and thus would be socially inexpensive relative to imprisonment. It should be emphasized in this connection that with the advance of technology, possibilities will increase for relatively cheap enforcement of sanctions involving restrictions in behavior through the use of remote electronic monitoring devices.[56]

5.3 Optimal choice among nonmonetary sanctions. Obviously, it is best for society to employ nonmonetary sanctions in the order of their effectiveness as deterrents. For example, if for a certain type of person humiliation combined with restrictions on freedom of movement would serve as a significantly more effective deterrent than imprisonment, such sanctions should be used before imprisonment is contemplated.[57] Further, if a nonmonetary sanction happens to be more effective than monetary sanctions, it should be used first (as was noted in section 4.3).

Note on the literature. The general point that, when nonmonetary sanctions are employed, the fault system enjoys a fundamental advantage over strict liability because sanctions are imposed less often, was first emphasized in Shavell (1985a). The theory of the optimal use of nonmonetary sanctions under the fault system is articulated in that article and in Shavell (1987c),

[56] Not only can devices such as television cameras, or unremovable wrist or ankle bracelets that send signals to a computer, be used to monitor movement, they could also be used to prevent the individuals from violating restrictions. For instance, bracelets could be designed to shock individuals or to inject them with an immobilizing drug by remote command.

[57] Consideration of incapacitation, which may favor imprisonment, is omitted here and will be discussed in chapter 23.

but has been adumbrated by Bentham ([1789] 1973) and others.[58] The general point that monetary sanctions enjoy an advantage over costly nonmonetary ones was also noted by Bentham and is emphasized by Becker (1968).[59]

[58] Beccaria [1767] 1995, chapters 2, 6, Montesquieu [1748]1989, book 6, chapters 9, 16, and Bentham [1789] 1973, 169-77 (who cites Beccaria and Montesquieu) all suggest that because nonmonetary sanctions are costly to impose, they should be used sparingly under the fault system (they do not consider strict liability) and, generally, only when likely to accomplish deterrence. However, they do not take into account the imperfect nature of the court's information. Only by doing so, as the reader knows (especially from sections 1.3 and 1.4 above), can one answer in an intellectually satisfactory way very basic questions about sanctions, including why sanctions are ever imposed and why extremely high sanctions should not be employed to help guarantee deterrence.

[59] The literature on the use of nonmonetary sanctions is discussed in the survey by Polinsky and Shavell 2000a.

Chapter 22

Extensions of the Theory of Deterrence

In this chapter, I consider various extensions of the theory of deterrence, most of which apply in their main aspects both when sanctions are monetary and when sanctions are nonmonetary, so that I will usually not distinguish these cases.

1. Individual Deterrence

1.1. Definition. In discussions of deterrence, the notion of *individual deterrence* (sometimes called *particular deterrence* or *special deterrence*) is usually mentioned.[60] Individual deterrence is the tendency of a person who has been penalized for committing an illegal act to be more deterred in the future from committing that act than he had been beforehand by the prospect of sanctions. For example, a person who has received a speeding ticket might be thought to be more deterred from speeding in the future by the possibility of sanctions than an otherwise identical person who has not received a speeding ticket. Individual deterrence is contrasted to *general deterrence,* the tendency of people who have not been sanctioned to be deterred by the prospect of sanctions for committing an illegal act.

1.2 Rationale for individual deterrence. The first point that should be made about individual deterrence is that it should not exist when calculating parties know the probability and the magnitude of sanctions for an illegal act. If a person realizes that he faces a probability of 30 percent of being ticketed for speeding on the highway and that the amount of the penalty is $100, it should not matter to him, when he is contemplating speeding, whether or not he himself had received a ticket in the past: In either case, he will face a 30 percent chance of bearing a $100 penalty if he now speeds.

Hence, for an individual to be more deterred as a consequence of having been penalized in the past -- for individual deterrence to exist -- it must be that the person does not know the probability or the magnitude of sanctions and, further, that his perception of the expected sanction must rise as a consequence of having been penalized.[61]

Will the perceived probability of sanctions increase as a result of having been sanctioned? The answer is yes, provided that a person does not know the precise probability of sanctions. In this situation, when a person is punished, he will rationally modify his estimate of

[60] See, for example, Andenaes 1983, LaFave 2000, 23, and Packer 1968, 45-48.

[61] An *actual* increase in the probability or in the magnitude of sanctions is, of course, possible as a result of an infraction. After an infraction, an individual could be watched more closely by enforcement authorities than otherwise, or the law could specify that the penalty for a second infraction (such as a second speeding ticket) is higher than that for a first infraction. But an increase in deterrence due to a literal increase in the probability or magnitude of sanctions is not what is meant by individual deterrence; individual deterrence is assumed to come about from the mere fact of having been sanctioned.

the likelihood of punishment, and always in an upward direction. If, for instance, a person thought the odds of receiving a speeding ticket were in the neighborhood of 30 percent and then actually was caught for speeding, he would rationally raise his estimate of the odds of a ticket, perhaps to 40 percent or 50 percent. That is, whatever his initial beliefs about the probability, being punished will lead a person to increase his estimate of the likelihood of being punished in the future, according to the laws of conditional probability.[62] Moreover, there is reason to believe that people often adjust their probabilistic beliefs upward as a result of being caught more than is justified by probability theory.[63] Hence, individual deterrence will come about on account of actual punishment influencing the perceived probability of future punishment, and this effect will be greater the more uncertainty parties have as to the odds of punishment.

With regard to the perceived magnitude of sanctions, the situation is different. If individuals have imprecise knowledge of the magnitude of sanctions, then there is no systematic reason to believe that they will raise their estimates of the magnitude of future sanctions as a result of being punished. If an individual had underestimated the magnitude of sanctions and learns that they are higher, he will be more deterred in the future; but if he had overestimated the magnitude of sanctions and learns that they are lower, he will be less deterred in the future.[64] Unless individuals underestimate actual sanctions more than they overestimate them, there is no reason to believe that being punished and thereby learning the true sanction would lead those who are punished to be more deterred in the future.

Finally, it may be mentioned that individual deterrence might arise for a reason apart from an increase in the perceived likelihood or magnitude of sanctions. The experience of punishment might trigger feelings of guilt, a realization that one has failed to act responsibly, and thus cause some individuals not to repeat their violations (on such guilt, see sections 2 and 3 of chapter 26).[65]

[62] Suppose that a person believes the probability of being caught and sanctioned is either small, p_s, or high, p_h, where $p_s < p_h$. Further, he believes that the likelihood that p_s is the probability is q and that the likelihood that p_h is the probability is $1 - q$. Then the person's likelihood now of being caught is $qp_s + (1 - q)p_h$. If the person commits the act and is caught, he will revise upward the probability of p_h and downward the probability of p_s. In particular, his probability of p_h conditional on being caught will rise from $1 - q$ to $1 - q' = (1 - q)p_h/[qp_s + (1 - q)p_h]$ (which equals $(1 - q)/[qp_s/p_h + (1 - q)] > (1 - q)$); and his probability of p_s conditional on being caught will fall from q to $q' = qp_s/[qp_s + (1 - q)p_h]$ (which equals $q/[q + (1 - q)p_h/p_s < q$). Hence, his probability of being caught will rise to $q'p_s + (1 - q')p_h$.

[63] See, for example, Tversky and Kahneman 1974.

[64] Suppose that the person does not know the magnitude of the sanction; he believes it is either small t_s, or large t_h, where q is the likelihood of t_s and $1 - q$ that of t_h. Suppose for simplicity as well that the probability of sanctions is known and equals p. Then the expected sanction ex ante is $p[qt_s + (1 - q)t_h]$. Suppose also that if a person is punished, he will learn the true sanction, either t_s or t_h. Then, if he is punished and the true sanction is t_s, the expected sanction will be pt_s; and if he learns that the true sanction is t_h, the expected sanction will be pt_h. Hence, the expected sanction after a person is caught is $q[pt_s] + (1 - q)[pt_h]$. But this equals the expected sanction ex ante, $p[qt_s + (1 - q)t_h]$. Thus, there is no individual deterrence due to being sanctioned when there is uncertainty over the magnitude of the sanction.

[65] To amplify, this explanation rests on the assumption that, after being punished, the person will view the act in question differently, and will anticipate that if he commits it again, he will feel more guilty about it than he had anticipated he would beforehand; so an element of his calculus, namely, the internal sanction of guilt for committing the act, will change as a result of punishment.

1.3 Significance of individual deterrence. From the above discussion, it appears that that individual deterrence is potentially important only when there is substantial uncertainty about the likelihood of sanctions or when, for some reason, parties systematically underestimate the magnitude of sanctions or experience unanticipated feelings of guilt. Otherwise, when actors have reasonably good knowledge of the likelihood of sanctions, individual deterrence does not seem of much relevance. Notably, one suspects that for firms, individual deterrence often does not come about because firms tend to apprise themselves of the risk of sanctions for violations of law. In all, it seems that individual deterrence is usually of secondary significance.[66]

2. Marginal Deterrence

2.1 Definition. It has so far been assumed that an individual chooses whether or not to commit a single harmful act, so deterrence has been an either-or phenomenon. But an individual might choose 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 only to kidnap a person or also to kill him. In such contexts, the threat of sanctions plays a role in addition to the usual one of deterring individuals from committing harmful acts altogether: For individuals who are not deterred altogether, expected sanctions still influence which harmful acts these individuals choose to commit. These 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 expected sanction exceeds that for a less harmful act is sometimes referred to as *marginal deterrence*.[67]

2.2 Enforcement policy and marginal deterrence. Other things being equal, it is socially desirable that enforcement policy creates marginal deterrence, so that those who are not deterred from committing harmful acts have a motive to moderate the amount of harm that they cause. This suggests that sanctions should rise with the magnitude of harm (and, therefore, that all but the most harmful acts should be punished with less than maximal sanctions). However, fostering marginal deterrence may conflict with achieving deterrence generally: 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 appropriately deterred from committing such acts.[68]

Two additional observations should be made 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 more deterred from killing their victims if greater police resources are devoted to apprehending kidnappers who murder their victims than to

[66] There has been substantial study of individual deterrence from imprisonment, and the general finding is that imprisonment does not have much effect on criminality after release. See, for example, Lab and Whithead 1988 and Wright 1994, 25-36.

[67] The notion of marginal deterrence was remarked upon in some of the earliest writing on enforcement; see Beccaria [1767] 1995, 21, and Bentham [1789] 1973, 171. The term "marginal deterrence" apparently was first used by Stigler 1970.

[68] For formal treatments of marginal deterrence, see Friedman and Sjostrom 1993, Mookherjee and Png 1994, Shavell 1992, and Wilde 1992.

apprehending those who do not. (Note, though, that in circumstances in which enforcement is general -- see section 5 below -- the probability of detection cannot be independently altered for acts that cause different degrees of harm.)

Second, marginal deterrence is naturally and automatically accomplished if the expected sanction equals harm for all levels of harm; for if a person is paying for harm done, whatever its level, he will have to pay 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 incentive not to pollute will be correct.[69]

3. Costs of Imposing Monetary Sanctions.

3.1 Principal conclusion: Cost should be added to sanction. Although the imposition of monetary sanctions was presumed to be costless in chapter 20, that is not in fact the case; legal proceedings, locating the assets of a person, and forcing him to disgorge assets all involve expenses.

The main difference that the presence of such costs makes is that the cost of imposing a sanction should be added to the sanction that would otherwise be optimal. The essential reason is that the effective social harm caused by a harmful act is the direct harm *plus* the indirect harm comprised of the expected cost of imposing sanctions. For example, suppose that a person's act causes direct harm of $100, that the person will suffer a sanction with certainty, and that the cost of imposing a sanction is $5. Then the situation is virtually the same as it would be if the person's act caused $105 of direct harm and there were no cost of imposing sanctions, for in either situation society bears $105 of costs. Hence, the optimal penalty for the harmful act that causes $100 of harm and costs $5 to penalize is $105, not $100; society wants the person to refrain from committing the harmful act unless the benefit to him is at least $105, rather than at least $100.

The conclusion that the cost of imposing the sanction should be added to the otherwise-optimal sanction also holds when there is only a probability of catching violators. Suppose that the likelihood of catching individuals who cause harm of $100 is 50 percent. Therefore, as explained in chapter 20, the optimal sanction would be $200 in the absence of consideration of the cost of imposing sanctions. If, however, it costs $5 to impose the sanction, then the claim is that the optimal penalty is $205. The reason is that, when a person commits the harmful act, the expected cost of imposing sanctions is 50% H $5 = $2.50, so that the expected sanction should be $102.50. And if the amount paid when the person is caught is $205, the expected sanction will be $102.50. Notice here that although we multiply the harm of $100 by a factor of two to reflect the chance of escaping sanctions, the basic rule for calculation of the optimal penalty is simply to add the cost of imposing the sanction to its otherwise optimal level.[70]

[69] As emphasized in section 2.4 of chapter 20, however, it often is desirable for society to tolerate some underdeterrence in order to save enforcement costs, in which case expected sanctions will be less than harm. Then consideration of marginal deterrence alters the structure of sanctions that would otherwise be best.

[70] To be precise, we know from general arguments along the lines of chapters 8 and 20 that (risk-neutral) parties will be induced to behave socially correctly provided that their expected liability equals the expected social harm due to their acts. If the direct harm due to an act is h, the probability of a monetary sanction is p, and the social cost of imposing

3.2 Comments. To the basic rule that the cost of imposing a sanction should be added to the otherwise optimal sanction, a number of qualifications and additions are worth making.

(a) *Marginal versus fixed costs of imposing sanctions.* It was taken for granted above that the costs of imposing sanctions are marginal in the sense that they are borne when and only when an additional person is sanctioned, but often there are also fixed costs of imposing sanctions, that is, costs that do not vary with the number of individuals sanctioned. For instance, the expense of a computer system for purposes of enforcement may have to be incurred regardless of the number of individuals sanctioned. As these costs do not increase if another person is sanctioned, there should be no addition to the sanction on their account. (It would be wrong, for example, to "allocate" these costs, charging each person the average amount.) Such fixed costs might, however, have an effect on the sanctioning policy. The fixed costs might influence the optimal probability of catching individuals (the fixed costs might well rise with the probability of enforcement, even though they are not affected, by the number of individuals sanctioned). If large enough, the fixed costs might make it undesirable to sanction parties at all, for then the fixed costs would be avoided. But the point here is that the fixed costs do not affect the optimal magnitude of the sanction given the probability of catching and sanctioning parties.

(b) *Costs increase with the magnitude of sanctions.* The cost of imposing sanctions may increase with the magnitude of the sanction because of greater resistance to sanctions as their amount increases. In this case, it can be shown that the optimal sanction should be the harm plus an amount somewhat lower than the actual cost of imposing sanctions, for this sanctioning policy reduces the incentive of parties to spend in resisting sanctions.

(c) *Costs borne by sanctioned parties.* Some of the costs of imposing sanctions are borne by the sanctioned parties themselves, in their own time and effort and in hiring legal counsel. Such costs do not affect the optimal sanction, for the parties automatically take them into account as an implicit sanction that they bear.

(d) *The optimal probability.* The appeal of the use of low probabilities of sanctions increases when imposition of sanctions is costly, for then low probabilities mean a savings in costs of imposing sanctions as well as a savings in enforcement expenses.

(e) *The form of liability.* There is an underlying advantage of fault-based liability when there are costs of imposing sanctions. As the reader knows, under a perfectly functioning fault-based rule, all parties will be deterred from acting undesirably and thus no sanctions will ever be imposed, so no costs of imposing sanctions will be borne. Of course, as has also been discussed, various sources of error mean that parties will be found liable under the fault system, and thus the advantage of this form of liability in reducing costs of imposing sanctions is diminished.[71]

the sanction is k, then the expected social harm due to the act is $h + pk$. If the sanction when the person is caught is, as claimed to be optimal, $(h/p) + k$, then the expected sanction is $p[(h/p) + k] = h + pk$, so that incentives will indeed be correct.

[71] The points made in this section are developed in Polinsky and Shavell 1992, although Becker 1968, 192, recognized that sanctions should reflect enforcement costs.

4. Self-Reporting of Violations

4.1 Definition. In the consideration of law enforcement, the assumption to this point has been that individuals are sanctioned only if they are detected by an enforcement agent. But in reality parties sometimes disclose their own violations to enforcement authorities. For example, firms often report violations of environmental and safety regulations, individuals frequently notify police of their involvement in traffic accidents, and even criminals sometimes admit their illegal acts and turn themselves in to the police. Such behavior will be called *self-reporting*.

4.2 Inducement of self-reporting and its social desirability. How, precisely, can individuals be led to report their own violations, and why might it be socially desirable for the structure of enforcement to be such as to encourage self-reporting? Self-reporting can be induced by the state's lowering the sanction for individuals who disclose their own infractions. Moreover, the reduction in the sanction for self-reporting can be made small enough that deterrence is only negligibly reduced -- thus, self-reporting can be accomplished with only a slight effect on deterrence. To illustrate, consider a situation in which risk-neutral violators of a law face, say, a 50 percent probability of being caught and of having to pay a sanction of $100, so that the expected sanction is $50. If there is no reduction in the sanction for self-reporting, no one will rationally report on himself; for it would not make sense to pay $100 for sure rather than to bear an expected sanction of only $50 if one does not self-report. But suppose that if a person self-reports, he only has to pay a sanction of $49.99. Under this scheme, every violator will in principle decide to come forward since $49.99 is less than the expected sanction of $50 that he would otherwise face.[72] Note as well that, because the penalty is $49.99 instead of $50 in expectation, the penalty for a violation has barely fallen, so that deterrence of the violation will be essentially the same under the self-reporting scheme as it would be in the absence of any reduction in the sanction for self-reporting.[73]

Why is self-reporting socially advantageous? One reason is that self-reporting tends to lower enforcement costs because, when it occurs, the enforcement authority does not have to identify and prove who the violator was. For instance, environmental enforcers do not need to spend as much effort trying to detect pollution and establishing its source if firms that pollute report that fact, and police do not have to continue their investigation of a robbery if the robber comes forward and confesses.[74]

Second, self-reporting reduces risk for potential violators, and thus is advantageous if

[72] More realistically, the self-reporting scheme would have to involve greater than a $.01 advantage for violators to be led to report on themselves.

[73] To state the argument of this paragraph formally, let p be the probability of being caught for a violation and s the sanction then imposed, so that the expected sanction is ps if the person does not self-report. Let s' be the sanction if a violator self-reports, and set $s' = ps$! ,, where , > 0 is arbitrarily small. A violator will therefore want to self-report because s' is less than ps, but the deterrent effect of the sanction will be (approximately) the same as if he did not self-report.

[74] In some contexts, however, self-reporting will not save enforcement costs. For example, suppose that a police officer waits by the roadside to spot speeders. Then, were a driver to report that he had sped, this would not reduce policing costs, presuming that the officer still needs to be stationed at the roadside to watch for other speeders. Usually, though, there would be some cost savings as a result of self-reporting (for example, the police officer would not have to chase as many speeders).

potential violators are risk averse.[75] Drivers bear less risk because they know that if they cause an accident, they will be led to report this to the police and suffer a modest, certain sanction, rather than face the probability of a substantially higher sanction imposed if they are caught for having caused an accident (such as being caught for a hit-and-run driving accident).

Third, self-reporting sometimes allows harm to be mitigated because it may mean that harm is reported without undue delay. Early identification of a toxic leak will facilitate its containment and clean-up, and the reporting of a traffic accident may result in the victim receiving medical attention that otherwise would not have come until later.[76]

5. General Enforcement

5.1 Definition. In many settings, law enforcement may be said to be *general* in the sense that several different types of violations may 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 a driver who goes through a red light or who speeds, or a tax auditor may detect a variety of infractions when he examines a tax return. To investigate such situations, I will suppose for simplicity below that a single probability of detection applies to all harmful acts, regardless of the magnitude of the harm.[77] The contrasting assumption to the present one is that law enforcement is *specific* to the harmful act, meaning that the state selects the probability of sanction independently for each type of harmful act. This is the assumption that was made previously.

5.2 Optimal enforcement policy. The main point that I want to make is that in contexts in which enforcement is general, the strategy of employing very high sanctions accompanied by very low probabilities of detection, in order to save enforcement costs, is no longer as appealing as had been argued earlier (see especially section 2 of chapter 20). Further, when enforcement is general, it is optimal to employ maximal sanctions only for the most harmful acts; otherwise, it is best to impose lower sanctions the less harmful the act.

To explain why a high sanction and low probability of enforcement does not always tend to be a desirable enforcement policy, consider the case of risk-neutral parties and deterrence of a relatively small infraction, such as double-parking. Before, it was explained that if the sanction for that infraction was less than maximal, it would typically be beneficial to raise the sanction and lower the probability of apprehension so as to save enforcement expenses while maintaining deterrence of the act. In the context of general enforcement, this scheme is no longer necessarily

[75] The argument of note 14 that self-reporting can be induced without lowering deterrence applies with minor modification when individuals are risk averse. Let U be the utility of a person's wealth and y his initial wealth. Then, in the absence of self-reporting, the expected utility of a violator is $(1\ !\ p)U(y) + pU(y\ !\ s)$. Let c be such that $U(y\ !\ c) = (1\ !\ p)U(y) + pU(y\ !\ s)$. (That is, c is the so-called certainty equivalent of the sanction s.) Then any sanction for self-reporting of $c\ !\ ,$, where $, > 0$ is small, will lead to self-reporting, with only negligible effect on deterrence.

[76] The basic theory of self-reporting in law enforcement is developed in Kaplow and Shavell 1994b, but see also Malik 1993 and Innes 1999.

[77] It will be clear that the main point developed in this section does not depend on the assumption that the same probability of enforcement applies to all acts. The only requirement is that the probabilities for different acts are linked, all a function of the same enforcement expenditure.

beneficial, however. If the likelihood of catching double-parking violations is lowered by reducing the number of police, the likelihood of detecting other, perhaps more serious violations, will *also* be lowered due to there being fewer police. And that may be socially undesirable, for it may not be possible to raise the sanctions for these other violations enough to maintain deterrence, as they may already be punished by very high sanctions. Indeed, if a more serious act (say intentionally running someone over with one's car) is already punished by the maximal sanction, deterrence of that act will be reduced if the likelihood of sanctions falls because there are fewer police on duty.

Let me now sketch more of the argument about optimal enforcement policy. Consider the class of very harmful violations. To deter them adequately, society needs a sufficiently high probability of apprehension, meaning a certain number of enforcement agents, even though it can and will impose the greatest sanctions for these serious violations. Now given that society uses the number of enforcement agents that it needs to control adequately the very harmful acts, these enforcement agents will, as a byproduct, produce a sufficiently high probability of sanctions for less serious acts that, for those acts, moderate sanctions can be used to deter them appropriately. As a consequence, the optimal sanctions for the less serious acts may well be in proportion to their harmfulness. For example, suppose that the probability of catching violations must be one-third in order to control properly the most serious offenses. Then the optimal sanction for a violation is three times the harm, so that, for the range of harms below one-third of an individual's wealth, the individual will be able to pay the optimal sanction, and in that range the sanction will be higher the higher is the harm.[78]

To conclude, it may be worthwhile pointing out that, were enforcement specific to the act, it would be optimal for the sanction for each act, regardless of its harmfulness, to be maximal, and for the probability of enforcement to be lower for less serious acts. The difference when enforcement is general is that the probability of enforcement of different acts is not independent.

[78] The formal argument about optimal enforcement policy of this section may be described roughly as follows in the case of monetary sanctions (the case of nonmonetary sanctions is similar). Let $s(h)$ be the sanction given harm h. Then, for any general probability of detection p, the optimal sanction schedule is $s^*(h) = h/p$, provided that h/p does not exceed the level of wealth of individuals w, which is the maximal feasible sanction; if h/p is not feasible, the optimal sanction is w. In particular, this schedule is obviously optimal given p because it implies that the expected sanction equals harm, thereby inducing ideal behavior, whenever that is possible, and the expected sanction is as high as feasible otherwise. The question remains whether it would be desirable to lower p and raise sanctions to the maximal level for the low-harm acts for which $s^*(h)$ is less than maximal. The answer is that if p is reduced for the relatively low-harm acts (and the sanction raised for them), then p -- being general -- is also reduced for the high-harm acts for which the sanction 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. The optimal lowering of p reflects a compromise between saving enforcement costs and diluting deterrence of relatively high harm acts. This argument, and the distinction between general and specific deterrence, is introduced in Shavell 1991b; see also Mookherjee and Png 1992 for a closely related analysis.

6. Insurance against Sanctions

The possibility of insurance against sanctions has not yet been mentioned, and I have assumed implicitly above that parties do not carry such insurance. As a general matter, this is in keeping with reality: Insurers are not permitted to offer coverage against most criminal fines and some civil penalties.[79]

The chief issue of interest to us is whether the observed policy against sanction insurance is socially desirable from a theoretical perspective. The relevant issues here are similar to those discussed in relation to the social desirability of liability insurance (see sections 4 and 7 of chapter 11), so I can be brief. If sanction insurance is available, risk-averse parties who might violate the law will tend to wish to purchase the insurance. Thus, the availability of sanction insurance will reduce the bearing of risk by individuals who violate the law, which is in itself socially desirable. However, the ownership of sanction insurance will tend to dilute the deterrent effect of sanctions, for violators will be less afraid of sanctions owing to the insurance. Whether allowing the purchase of liability insurance is socially undesirable or desirable depends on the importance of these two effects.

Some reflection about the context of law enforcement suggests that the social advantage of reducing risk for potential violators is outweighed by the dilution of deterrence factor, making prohibition of sanction insurance socially desirable. First, it seems that, for many acts that society seeks to control through public enforcement of law, the potential violator has a clear ability to commit or not to commit the act giving rise to sanctions; a person generally knows whether he is committing a crime, and many other types of violation. If this is the case, then a person can avoid risk by deciding to obey the law. He does not much need sanction insurance to avoid the risk of penalty for beating someone up, committing fraud, or intentionally cheating on his taxes.[80]

Second, in the context of law enforcement, we have emphasized that it is generally desirable for society to conserve enforcement expenses by maintaining a relatively low probability of sanctions, and to countenance underenforcement as a consequence.[81] The fact that, in reality, there is substantial underdeterrence of many undesirable acts is consistent with this point. Given that there is a problem of underdeterrence because of society's desire to save enforcement expenses, it would only compound the problem of underdeterrence to allow individuals to obtain sanction insurance. To put the point differently, were we to allow individuals to hold sanction insurance, society would have to increase its expenditure on enforcement in order to achieve the level of deterrence that we enjoy when the insurance is forbidden.[82]

[79] See, for example, Jerry 1996, 471-77, Keeton 1971, 285-305, Keeton et al. 1984, 586, and McNeely 1941.

[80] In contrast, in the typical tort setting, a person may find himself liable through some sort of accident. As discussed in sections 1 and 2 of chapter 10, individuals may be found negligent by mistake, and may not have complete control over their behavior. Thus, the value of liability insurance in reducing risk in the tort context seems, as a general matter, much greater than in the law enforcement context.

[81] See section 2 of chapter 20.

[82] Again, the contrast with the tort setting is instructive. Society does not face a general problem of

The foregoing is not meant to deny the possibility that sanction insurance is socially desirable in some situations. Suppose that individuals are able to control only probabilistically behavior that may result in sanctions (say they cannot necessarily prevent oil from leaking from a boat into a lake), and there is not a real problem of underdeterrence because the magnitude of harm is not great in relation to individuals' assets and the likelihood of detection is substantial (enforcement agents can easily ascertain when spills occur). Then the value of insurance in reducing risk may be substantial, and the ownership of insurance will not be problematic for incentives (no more so than in the usual tort context). In such circumstances, insurance against sanctions may be desirable.

7. Sanctions for Repeat Offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. For example, under the U.S. Sentencing Commission's guidelines for punishment of federal crimes, both imprisonment terms and criminal fines are enhanced if a defendant has a prior record; civil money penalties also sometimes depend on whether the defendant has a record of prior offenses.[83] I will attempt to explain here why such policies may be socially desirable.

Note first that sanctioning repeat offenders more severely cannot be socially advantageous if deterrence always induces ideal behavior. 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 sanction because of his having a record of sanctions would overdeter him now; it would not be socially desirable to raise sanctions on account of past infractions.

Accordingly, only if deterrence is inadequate is it possibly desirable to make sanctions depend on offense history in order to increase deterrence. But deterrence often will be inadequate because, as I have stressed, 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 offense history may be beneficial for two reasons. First, the use of offense history may create an additional incentive not

underdeterrence in the tort context, at least one not comparable to that in the domain of public enforcement, for harmful events in the area of tort, such as car accidents, will generally result in suit or settlement if injurers are liable. Hence, if liability insurance reduces somewhat the incentive to take proper care, this does not matter as much in the tort area as it does in the enforcement area. Moreover, if insurers can observe the level of care, incentives will be appropriate in the usual tort situation. In the context of enforcement, however, that is not necessarily so; if insurers can observe whether individuals violate the law, that will not lead individuals to refrain from violations if the expected sanction is less than the harm.

[83] See United States Sentencing Commission (1995, §4A1.1, chapter 5 part A, and §5E1.2). Regarding civil penalties, see, for example, 8 U.S.C. §1324a(e)(4)-(5)(1997), imposing minimum fines of $250 for a first offense, $2,000 for a second offense, and $3,000 for subsequent offenses concerning hiring, recruiting, and referral behavior under the Immigration Reform and Control Act; and see 29 U.S.C. §666(a)-(c) (1997), stating that the maximum fine is $7,000 for certain violations of the Occupational Safety and Health Act that are not repeated, but that the maximum fine rises to $70,000 if the violations are repeated.

to violate the law: If detection of a violation implies not only an immediate sanction, but also a higher sanction for a future violation, an individual will be deterred more from committing a violation presently.[84] Second, making sanctions depend on offense history allows society to take advantage of implicit information about the dangerousness of individuals and the need to deter them. Individuals with offense histories may well be more likely than average to commit future violations, which might make it desirable for purposes of deterrence to impose higher sanctions on them.[85]

There is also an obvious incapacitation-based reason for making sanctions depend on offense 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.

[84] There is a subtlety in demonstrating the optimality of punishing repeat offenses more severely. Namely, if there is a problem of underdeterrence, one might wonder why it would not be optimal to raise the sanction for a first offense, rather than, instead, to enhance deterrence by punishing repeat offenses more severely. See Polinsky and Shavell 1998a on the possible optimality of making sanctions depend on offense history because of the additional deterrence that such a policy creates.

[85] Note that this reason for making sanctions depend on offense history is different from the first reason: The second reason involves the assumption that offenders are different from one another and that the optimal sanction for some offenders is higher than for others; the first reason applies even if individuals are identical. On the second, information-based, reason for making sanctions depend on offense history, see Chu, Hu, and Huang 2000, Polinsky and Rubinfeld 1991, and Rubinstein 1979.

Chapter 23

Incapacitation, Rehabilitation, and Retribution

In this chapter, I discuss briefly several functions of sanctions apart from deterrence, namely, incapacitation, rehabilitation, and retribution.

1. Incapacitation

1.1 Definition of incapacitation. The most familiar form of incapacitation is imprisonment, which prevents individuals from engaging in undesirable acts in free society by removing them from it**.** More generally, incapacitation can be defined to be prevention of a class of undesirable acts by barring a party from engaging in an activity that would allow the party to commit the acts. For example, a person could be prevented from causing accidents when driving by preventing him from driving by voiding his driver's license, or a restaurant could be prevented from causing harm from serving spoiled food by being forced to close.

1.2 Incapacitation distinguished from deterrence. Preventing a party from engaging in an activity in which he could do harm is quite different from deterrence, that is, dissuading the party from committing an undesirable act through the threat to impose sanctions if he commits it. Deterrence works only when the party knows about and considers the possibility of sanctions, and only when the sanctions can actually be applied. (If the person is judgment proof and the sanction is monetary, the sanction cannot be applied; if the person is old or dying of a disease, the imprisonment term cannot be long.) Incapacitation functions independently of these factors.

1.3 Basic model of enforcement and incapacitation.[86] To focus on incapacitation, let us assume that individuals cannot be deterred and, initially, that each individual has an unchanging propensity to commit harmful acts (measured by the expected harm) per time period. Let us further suppose that society incurs expenses in raising the probability of apprehending individuals who do harm and who will be considered for incapacitation, and that society bears certain costs per period of incapacitation.

Under these assumptions, what is the optimal length of incapacitation of a person who does harm and who has been apprehended? The answer is simply that if the person's propensity to do harm each period exceeds the cost of incapacitation per period, he should be incapacitated each period, that is, forever; otherwise, he should not be incapacitated at all.

The optimal probability of apprehension will reflect the tradeoff between the cost of raising this probability and the benefit in terms of reduced harms through incapacitating more individuals.

1.4 Extensions and comments. (a) *Assumption that propensity to do harm is constant*

[86] For a formal model of incapacitation, see Shavell 1987b; for theoretically oriented discussions of incapacitation, see, for example, Blumstein 1983 and Packer 1968, 48-53. For extensive, and still relevant critical discussion of the literature on incapacitation, see Blumstein, Cohen, and Nagin 1978, and for a recent review and assessment, see Spelman 2000.

over time. If the propensity of an individual to do harm diminishes over time, then it becomes optimal to end incapacitation as soon as the propensity to do harm per period falls below the per period cost of incapacitation. This is a significant point, because the evidence is that the propensity to commit many types of crimes declines with age.[87]

(b) *Optimal sanction unrelated to probability of imposition.* It should be noted that the optimal length of incapacitation depends only on the propensity of individuals to do harm, *not* on the likelihood with which they are apprehended. In particular, from the standpoint of incapacitation, there is no reason to impose a higher sanction if the probability of detection is low. This contrasts with the situation under deterrence, where, as was emphasized in the last two chapters, lower probabilities of apprehension call for higher sanctions.

(c) *Relevance of the commission of a harmful act to the imposition of sanctions.* The optimal sanction depends only on the propensity to do harm, that is, the estimated future dangerousness of a person. There is, then, no intrinsic reason to require that a person actually have committed an undesirable act or that he actually have done harm for him to be incapacitated. However, the commission of a harmful act does often constitute evidence about the propensity to do harm, and for that reason a requirement of commission of a harmful act might be socially rational to impose for incapacitation. (In addition, departing from the model, the danger of state abuse of its ability to sanction would be lessened, one supposes, if there is a requirement that a party actually have committed a harmful act for him to be penalized.) According to the theory of deterrence, note, the requirement that there be a harmful act for there to be punishment is fundamental; deterrence can work only if a person knows that he will be punished if, but only if, he commits a harmful act.

(d) *Incapacitation and deterrence.* Suppose that individuals can be deterred as well as incapacitated. Specifically, consider again the model examined in the previous chapters on deterrence, but now assume that sanctions incapacitate as well as deter. Then, having two useful functions, sanctions will be optimal to employ more often than would otherwise be the case. Thus, where imprisonment would not be justified by its ability to deter, imprisonment might be warranted when account is taken also of its value in incapacitation. Similarly, where imprisonment would not be justified by its ability to incapacitate (suppose an embezzler of funds is discovered after he has retired and will have no future opportunity to embezzle), consideration of deterrence might call for imprisonment (potential embezzlers might be discouraged from acting due to the prospect of sanctions).

(e) *Optimal probability of incapacitation and optimal sanctions for deterrence.* To accomplish a desirable degree of incapacitation, the probability of sanctions must not be too low. This in turn may imply that the magnitude of sanctions needed for purposes of deterrence should not be too high. Hence, among other things, the argument (see section 2.5 of chapter 21) for very low probabilities of apprehension and for maximal sanctions might not apply. For instance, to achieve an appropriate degree of incapacitation of those who rob, it might be necessary to ensure that at least, say, 20 percent of robbers are apprehended. This might mean that the optimal sanction for deterrence purposes should be significantly lower than the maximum possible

[87] See,for example, Greenberg 1983, Wilson and Herrnstein 1985, 126-47, and U.S. Department of Justice 2001b, 362-63.

imprisonment term.

1.5 Actual importance of imprisonment as a form of incapacitation. The number of people who are presently imprisoned in the United States is 1.9 million, representing about 3 percent of the adult population, and the percentage of the population who will be incarcerated at some time during their lives is approximately 5 percent.[88] The annual cost of imprisonment is on the order of $47 billion, or about $24,000 per incarcerated person.[89] The annual incapacitive benefit of imprisonment -- its direct effect in reducing crime by keeping those who would otherwise commit crimes imprisoned -- has been calculated by some analysts to be in the neighborhood of 20 percent of the present level of crime.[90] The following calculation is also informative. If one estimates that the average prisoner would have committed 10 crimes per year were he not incarcerated, then the incapacitative benefit is that in the absence of incarceration, crimes would increase by about 19 million annually, or by about 90 percent.[91] Even if the incapacitative benefit is only 20 percent of the present level of crime, it would save society at least $100 billion, outweighing the $47 billion cost of imprisonment.[92] Accordingly, we can see that incapacitation is a very important and apparently well-justified function of imprisonment in this country.

2. Rehabilitation

2.1 Definition of rehabilitation. By rehabilitation is meant an induced reduction in a person's propensity to commit undesirable acts. This change may come about through direct effort of the state, notably through educational programs (such as those provided in prison) or as a byproduct of imposition of sanctions, when a person reflects on his behavior and decides to behave in a socially more responsible manner in the future.

2.2 Basic model of enforcement and rehabilitation. Assume that the sole function of

[88] See U.S. Department of Justice 2001b, p. 488, presenting an estimate of 1.933 million individuals for the year 2000. For an estimate of the fraction of the population who will be in prison at some time in their lives, see Bonczar and Beck, 1997.

[89] Annual expenditures in 1997 were $43.511 billion; see U.S. Department of Justice 2001b, 3. In terms of the consumer price index in 2000, the expenditures equal $46.656 billion; see Statistical Abstract of the United States: 2001, 451. Since there were 1.933 million persons imprisoned in 2000, the annual cost per person is $24,137.

[90] See Spelman 1994, 227, and the studies cited by Wright 1994, 116-17.

[91] The estimate of 10 crimes per year is actually somewhat conservative; see the discussion of literature on incapacitation in Wright 1994, 114-18. See also, for example, DiIulio and Piehl 1991, who find that the average annual number of violations per prisoner would be 141 and that the median would be 12. Using the estimate of 10 crimes per person per year and the fact that there were 1.933 million persons imprisoned in 2000, it follows that had these prisoners been free, they would have committed 19.333 million crimes annually. The actual number of crimes committed in 1999 was about 21.84 million; see Statistical Abstract of the United States: 2001, 182, so that an increase of 19.333 million crimes would represent an 88.5 percent increase in the overall level of crime.

[92] Anderson 1999, 625, estimates the annual cost of crime-related injury and death to be about $574.395 billion, and 20 percent of this amount is over $100 billion. This cost, note, is an incomplete measure of the social cost of crime, for it does not take into account, among other factors, the efforts made to avoid being a victim of crime and the efforts made to undertake crime.

sanctions is to rehabilitate. Then it is optimal to impose sanctions if and only if the rehabilitative benefit -- reduced future harm -- exceeds the cost of imposing the rehabilitative sanction. Thus, it is optimal for a person caught for drunk driving to be put in a class on driver responsibility if and only if the benefit, in terms of a reduction in expected accident losses, exceeds the cost of the class. The optimal probability of apprehending individuals who may be subject to sanctions is governed by the rehabilitative benefits that this brings about, assuming optimal imposition of rehabilitative sanctions.

2.3 Extensions and comments. (a) *Characteristics of optimal rehabilitative sanctions are similar to those of optimal incapacitative sanctions.* The optimal rehabilitative sanction, like the optimal incapacitative sanction, does not depend on the probability of apprehension. In addition, the actual commission of a harmful act is not intrinsically important to the rehabilitative sanction; in principle it would be desirable to rehabilitate any person who is known to need rehabilitation and can be improved at sufficiently low cost. For instance, someone who is known to get drunk and to be irresponsible, and thus to be likely to drive when drunk, might profit from a class on driver responsibility even if he has not committed any driving infraction. However, as stated before, the commission of a harmful act (like drunk driving) as a prerequisite for punishment may serve a valuable informational purpose and make governmental abuse of its authority less likely.

(b) *Rehabilitation and incapacitation.* If sanctions both rehabilitate and incapacitate, then the optimal length of sanction will, of course, reflect these functions. A notable implication is that a person whom society chooses to incapacitate would tend to receive a shorter sanction as a consequence of rehabilitation than incapacitation alone would call for.[93] This is because rehabilitation will hasten the time by which the person is sufficiently less dangerous that release is socially cheaper than continued incapacitation.

(c) *Rehabilitation and deterrence.* To some degree, rehabilitation may dilute deterrence. If a person believes that he will change in positive ways, for instance, that he will learn valuable skills in prison, the sting of the sanction may be lessened. This effect can be counteracted, but at a cost, by increasing the length of the sanction.

2.4 Actual importance of rehabilitation. Today, there is much skepticism about rehabilitation because there is substantial recidivism and little evidence supporting the notion that, in the United States, anyway, those who go to prison are less dangerous when released (except due to the effect of age on criminality).[94] Indeed, it is sometimes asserted that the opposite happens in today's prisons, that people who are in prison learn bad habits and ways of criminal life, so that they will do more harm, rather than less, as a result of imprisonment. However, one supposes that the failure of rehabilitation is more a function of present conditions than of intrinsic factors, and that rehabilitation might be of substantial importance in the future.

[93] There is, in principle, a possibility that rehabilitation would lengthen the stay of a person who would suffer a positive incapacitative sanction. It could be that, although the date at which he would become less dangerous than it costs to incapacitate comes earlier, it would still be beneficial to incarcerate him longer in order to further reduce his harmfulness.

[94] See, for example, Andenaes 1975, 339, Cook 1977, 165-66, Packer 1968, 53-58, Schwartz 1983, and Wright 1994, 25-36.

3. Retribution

3.1 Definition of retribution. The retributive motive is the desire of individuals to see wrongdoers punished. That is, individuals may derive utility from the knowledge that wrongdoers are punished. Such utility may depend on the proportionality of the punishment to the wrongdoing and may be greater the more serious the act of the wrongdoer.[95] Additionally, retributive utility may be more significant for victims of wrongdoing, or for those associated with them, than for the population at large.

3.2 Comments on the retributive desire. (a) *Criticism of the desire.* Some commentators suggest that retributive satisfaction should not be credited in the social calculus because the satisfaction is associated with the suffering of another. This view, that certain types of satisfaction should not be counted in social welfare, is problematic and leads to anomalies, as will be generally discussed later.[96]

(b) *Sociobiological origin.* It has been observed that the desire for retribution serves a helpful sociobiological purpose. The presence of the desire means that those who are attacked will be likely to fight back. This discourages attack, which is a good thing because it means that people will not need to devote as much time to protecting what they have nor be as likely to become involved in destructive and wasteful conflict. Hence, one would predict that the retributive urge, at least if not too strong (in which case even slights would trigger conflict), would win out in evolutionary competition, as it apparently has in other animals as well as humans.[97]

(c) *Effect on the probability of sanctions.* The retributive urge also serves a purpose in present day society, which is to give people a motive to ward off transgression and thus to deter it, as well as to report on transgressors to social authorities so that they can be punished. Pure self-interest would often lead individuals not to respond directly, nor would it usually lead individuals to report transgressors to enforcement agents, as that takes effort and may invite retaliation. Hence, the retributive urge may be a significant factor in maintaining the probability of apprehension at its level; in the absence of the desire for retribution, many more enforcement agents would be needed to maintain the probability of apprehension.[98]

[95] A natural formalization of retributive utility is that it is a function $r(s,w)$ where s is the sanction, w is the degree of wrongdoing, and r is single-peaked in s and maximized at $s(w)$. Here $s(w)$ is the appropriate sanction given w, and $s(w)$ is increasing in w.

[96] It is explained in section 5.5 of chapter 26 that any measure of social welfare that is not based on utilities of individuals sometimes will reduce the well-being of *all* individuals. Therefore, that is true of a measure of social welfare that excludes certain sources of utility. A mundane example of this possibility is that all individuals might sometimes play practical jokes on others, sometimes themselves be the butt of practical jokes, and derive more utility from playing these jokes and enjoying them when others play them, than they suffer disutility as victims of the jokes. Therefore, all individuals might prefer a world in which practical jokes are permitted than one in which they are barred. If, however, utility that is derived from the displeasure of others (the victims of jokes) is not credited in the social calculus, practical jokes might be disallowed, making all worse off.

[97] On the biological origins of retribution, see, for example, Daly and Wilson 1988, chapters 10, 11, Frank 1988, chapters 3,4, Hirshleifer 1978, 334, Hirshleifer 1987, and Trivers 1971, 49.

[98] This point is stressed by Posner 1980.

3.3 Basic model of enforcement and retribution. If the only purpose of punishment were retribution, the optimal magnitude of sanction would be that which maximized the pleasure from satisfying the retributive desire minus the costs of imposing punishment. The optimal probability of apprehension would reflect this retributive gain net of costs as the benefit from capturing a person.

3.4 Comments. (a) *Retribution and deterrence.* As observed above, retribution enhances enforcement by increasing the motive of individuals to report what they know to social authorities, so it generally contributes to enforcement. With regard to the magnitude of sanctions, however, the effect of retribution is unclear. On one hand, the optimal sanction from the perspective of deterrence will often exceed that demanded by the retributive goal: The low probability of sanctions, which is best from the viewpoint of deterrence because it saves enforcement costs, raises the sanction needed to deter, yet the desire for retribution is not affected by the low probability of sanctions. (From the deterrence perspective, for example, we may want to impose a ten year prison sentence on a car thief because the odds of finding him are quite low, but the demand for retribution against him may well limit the sentence to a lesser level.) On the other hand, the retributive desire could exceed the proper punishment from the deterrence perspective, as where a person could not have been deterred (suppose a person killed another when in an insane rage).[99]

(b) *Retribution and incapacitation.* The optimal sanction from the perspective of incapacitation does not seem to be related in a clear way to what is demanded by retribution. The optimal incapacitative sanction would be lower than that needed to satisfy retributive desires if the wrongdoer would be unlikely to do harm in the future (suppose a person murders another in unique circumstances). Conversely, the optimal incapacitive sanction would be higher than that appropriate for retribution if a person did little harm yet would be likely to do great harm in the future.

(c) *Retribution and rehabilitation.* The goal of rehabilitation appears to conflict with that of retribution, supposing that rehabilitation reduces the disutility associated with punishment. However, as remarked earlier, this problem can be mitigated at a social cost by increasing the magnitude of the sanction.

[99] On enforcement policy in the light of retribution and deterrence, see Polinsky and Shavell 2000b.

Chapter 24

Criminal Law

1. Description of Criminal Law

1.1 Domain of criminal law. Most legal systems designate an area of law as *criminal,* that is, label certain acts as criminal and punish them in ways that are in some respects unique to criminal law. Although there is no simple, overarching definition of criminal acts, the following categories of criminal acts will help to describe the domain of criminal law.

(a) *Acts that are intended to do substantial harm.* The major category of criminal acts are those in which an individual intends to do significant harm. For example, murder, rape, robbery, counterfeiting, and treason are considered to be criminal acts. Notice that these acts ordinarily have the character that the person carrying them out intends harm in the sense that he wants harm to occur: The object of the murderer is usually to kill his victim, that of the rapist to rape, and that of the robber to take what is not his.[100] If harm does not actually come about, the act is normally still treated as criminal; thus, if a person attempts murder or rape or robbery but does not succeed, his act is still criminal. If harm is not intended, then even if it comes about as a result of an act, the act is not usually considered criminal. Thus, if a person shoots his gun while hunting and happens to hit another hunter whom he had no reason to notice, his act would not be criminal, or if he takes a suitcase that is someone else's but that he thought was his own, this will not be theft, because he did not intend to take something that was not his. Also, if harm is intended but is small in magnitude, then the act will not usually be considered criminal. Thus, if a person intentionally disturbs another person by, say, speaking loudly, his act will not be criminal even though his purpose may have been to do harm.[101]

(b) *Acts that are concealed, even if substantial harm was not intended.* Another category of act that is often considered criminal is an act that is harmful, or potentially harmful, and for which the actor has attempted to conceal or evade his responsibility. For instance, if a person flees the scene of a car accident, his act will usually be treated as criminal, or if a firm covers up the violation of a safety regulation, its act will often be characterized as criminal. This is a separate category of criminal act from that described in the previous paragraph, as the act that is evaded does not have to be intentional or to have created substantial risk or caused substantial harm for criminal liability to result.

(c) *Certain other acts.* In addition, there are diverse, particular acts that are categorized as criminal, even if substantial harm is not intended and even if concealment or evasion are not necessarily an issue. Falling into this category of criminal acts are, for example, a restaurant serving liquor to minors, and speeding.

[100] Intent will be more carefully defined later, where the definition will be expanded to include acting in a way that is felt to be extremely likely to cause harm, even if the harm is not itself desired.

[101] A minor qualification to this paragraph is that if a person is forced to act in one of several harmful ways, and chooses the least harmful, then he will not be held criminally liable, even though his act is intended to do substantial harm. See section 4.1 below for further discussion of this point.

1.2 Criminal sanctions. When an act is criminal, the sanctions that apply may include imprisonment, various other nonmonetary sanctions, money penalties, and social sanctions associated with being labeled a criminal.

(a) *Imprisonment and other constraints on freedom.* Imprisonment is a sanction that is unique to criminal law, as are certain other constraints on freedom, such as confinement to one's home enforced by electronic monitoring, probation, or required community service.[102] Such sanctions are typically imposed for acts in the central area of crime (a) above, and sometimes for acts in the second category (b), but usually not for crimes in the third category (c).

(b) *Other nonmonetary punishments.* In addition to constraints on freedom there are such sanctions as whipping, amputation of limbs, and also banishment and the death penalty. Today, use of some of these punishments is restricted or nonexistent in many Western countries, but in the past they have been important, and are employed contemporaneously in many areas of the world to one degree or another, mainly for acts in the major area of crime.[103] An additional form of nonmonetary sanction is punishment primarily intended to shame or humiliate; historical examples include the pillory, and today we see such practices as publishing the name of an offender in a newspaper or requiring him to post a sign on his property or a bumper sticker on his automobile.[104]

(c) *Money penalties.* Criminal acts may also result in money penalties, that is, in criminal fines. Fines are sometimes imposed for acts in the core area of crime, but are not usually the only sanction for those acts; whereas they may be the only sanction for acts in the second category (b), and they are often the only sanction for acts in the third category (c). Criminal fines differ from civil sanctions, such as tort judgments and civil penalties, in two respects that usually make them more effective. First, parties generally cannot purchase liability insurance against criminal fines,[105] although they can and usually do purchase coverage against civil sanctions. Second, parties cannot deduct criminal fines as business expenses and thereby reduce their income taxes on that account.[106]

(d) *Labeling and reputational penalties.* When an individual is convicted of a criminal offense, he is often said to be labeled as a criminal. Sometimes, however, a convicted criminal may conceal his past, and it may be difficult for others to determine what it was. Efforts to label individuals with criminal records are occasionally made. Historically, labeling was accomplished, among other ways, by branding individuals.[107] The effects of labeling include shame and humiliation as well as an implicit monetary sanction to the extent that labeling

[102] There is however the possibility of civil institutionalization of people who are found to be insane. This is similar to imprisonment in that the consequence of certain acts or behavior is a state-enforced restraint on conduct.

[103] For example, in certain countries governed by Islamic law, whipping or stoning may occur as punishment for unlawful intercourse, and limbs may be amputated as punishment for theft. See Forte 1983.

[104] On the pillory, see Beattie 1986, 464-68, and on humiliations today, see, for example, Hoffman 1997.

[105] See, for example, Jerry 1996, 400-13.

[106] Internal Revenue Code, Section 162(f).

[107] See, for example, Baker 2002, 515.

compromises a person's earning ability.[108]

2. Explanation for Criminal Law

2.1 Question to be addressed. Having described criminal law, the question arises, why does it exist? By this I mean, why should society want to designate a certain set of acts as falling under a special head, that of criminal law, and then use imprisonment and other sanctions as punishments for commission of these acts?

2.2 Answer in outline. The answer is at root simple: *Society requires criminal law in order to constrain certain behavior that could not otherwise adequately be controlled.* Specifically, I will suggest that acts in the core area of crime (acts intended to do substantial harm, category (a)) cannot be appropriately discouraged by the threat of monetary sanctions alone, so that the additional sanction of imprisonment (and/or other severe punishments) becomes socially desirable as a deterrent, and also as a means of preventing the future commission of undesirable acts by means of incapacitation. I will attempt to explain along similar lines that acts in the other two areas of crime (concealed acts, category (b), and certain additional acts, category (c)) need to be made criminal in order to deter them properly. At the end of this section, I will comment on the relationship of this thesis to the thesis that criminal law is intended to punish acts with especially bad moral qualities.

2.3 Acts in the major category of crime would be inadequately deterred by monetary sanctions alone. The hypothesis is that acts in the major category of crime, namely, rape, murder, theft, robbery, and other acts traditionally punished by imprisonment, would be inadequately deterred it they were punished solely with monetary sanctions. To this end, consider the factors noted in section 3.2 of chapter 21 that bear on the need for nonmonetary sanctions.

(a) *Level of assets.* Statistics show that individuals who commit crimes tend to have low wealth.[109] Moreover, this association is not unexpected: Those with little wealth have greater reason to commit economically-motivated crimes such as theft than do others, and low wealth is correlated with general characteristics that are linked to criminality, including substandard education, drug and alcohol abuse, and social alienation. To the extent that those who tend to commit crimes have low levels of wealth, the use of monetary sanctions alone would not be likely to provide an effective deterrent against crime because violators would not be able to pay the sanctions needed to accomplish deterrence.

(b) *Probability of escaping sanctions.* The probability of escaping sanctions for crimes is substantial; according to recent data, for example, the rate of incarceration for reported larceny-theft is in the neighborhood of just 8 percent, that for reported rape about 25 percent, and even

[108] It is occasionally observed that the reputational effect of being labeled as a criminal becomes diluted as the class of criminal acts broadens to include acts that are not viewed as especially bad. Note that this view rests on the assumption that a person who is labeled as a criminal is regarded as having committed some act, perhaps the average act, in the general category of criminal acts. If, by contrast, the specific nature of a criminal's violation becomes known, whether it is rape or income tax evasion, for example, his reputational loss would be determined by what he did, and the breadth of the class of criminal acts would not dull the reputational loss associated with committing this or that criminal act.

[109] Notably, the inmate population is composed of people who have very little income prior to arrest. For example, U. S. Department of Justice 1988, 35, reports that "the average inmate was at the poverty level before entering jail."

that for murder, only approximately 42 percent.[110] The reason that the probability of escaping sanctions is significant for crimes is, of course, that these acts are often planned and executed by individuals just so that they will be able to avoid identification or apprehension and thus escape sanctions.[111] To the degree that those who commit crimes can escape punishment, the monetary sanctions necessary to deter are raised, and along with them the likelihood that these sanctions would exceed violators' wealth and not successfully deter them.

(c) *Level of private benefits.* The benefits that individuals obtain from committing the acts that are in the core area of crime are frequently large. Those who steal significant amounts, who murder, and who rape, are committing acts for which the private benefits are substantial. This factor raises the monetary sanction needed to deter and reduces the ability to deter with monetary sanctions alone.[112]

(d) *Expected harmfulness of acts.* The expected harm caused by acts in the core area of crime appears to be substantial. First, the actual magnitude of harm associated with the acts in question tends to be high (certainly this is true when a person is murdered or raped, for example). Second, the likelihood of harm from these acts is generally high. This is both apparent as a matter of observation about the criminal acts (when a person sets out to rob or to murder, he is often going to succeed), and it is also something that follows from the frequently intentional character of the criminal acts, that they are such that the person is usually trying to cause harm.[113, 114] If, for these two reasons, the expected harm caused by criminal acts in the core area is high, the acts are more important to deter than others. Thus, society's willingness to bear the

[110] For larceny-theft, the fraction of reported cases leading to arrest and prosecution is about 19 percent, of which about 71 percent result in convictions, of which about 63 percent result in incarceration, impling that the frequency of incarceration is about 8 percent; see U.S. Department of Justice 2001b, 383, 458, 463. The corresponding statistics for rape are 49 percent, 63 percent, and 79 percent, and for murder they are 69 percent, 64 percent, and 95 percent; see the same source and pages. Note that because most crimes (other than murder) are under-reported, the true rates of incarceration for larceny-theft and rape must really be lower than the numbers calculated here. For instance, the likelihood that rape is not reported is estimated to be about 48 percent; see U.S. Department of Justice 2001b, 189, 383. On this basis, the likelihood of incarceration for rape would be not 25 percent but roughly 13 percent.

[111] In contrast, note, the typical tort arising from an accident occurs (as the word "accident" suggests) at an unpredictable time and place, and thus only by chance such that the responsible party can easily avoid being identified.

[112] The situation is different in the context of the typical tort, in which the private benefit that an actor usually obtains from acting improperly is only the avoidance of the cost of a precaution, like saving the effort of removing oily rags that could cause a fire. It requires a much smaller penalty to induce a person to give up this sort of gain than it does to induce a person to give up the likely gains from most crimes.

[113] Further, acts are sometimes made criminal just because they produce an extremely high likelihood of harm (even though harm is not desired by the actor). Suppose, for instance, that a person knowingly leaves a live wire exposed where children are playing and a child is electrocuted; this might constitute manslaughter owing to criminal negligence. See, for example, LaFave 2000, 246-57, 721-28.

[114] The expected harm associated with negligent acts that result in torts seems to be lower than that caused by criminal acts. The magnitude of the harm caused by negligence may, of course, be as high as that caused by a criminal act -- for instance, a tort may result in a person's death. But the likelihood of harm resulting from negligent acts appears to be much less than that from criminal acts (compare the likelihood of death from negligent driving to the likelihood of death from attempted murder by shooting at someone).

cost of employing imprisonment as a sanction in order to enhance deterrence should be greater for acts in the core area of crime than for other acts.

(e) *Illustration: murder.* Consider the crime of murder and ask whether, in view of what has just been said, murder could be controlled tolerably well through the use of money sanctions alone. It appears not. For example, even the median level of wealth of individuals below 35 years of age is less than $9,000,[115] so that the median expected monetary sanction for individuals in this cohort would be (see paragraph (b)) at most 42% H $9,000 or $3,780. Given this low penalty, one supposes that the murder rate would mushroom; the number of situations in which the value of murder to a potential murderer would exceed $3,780 is probably great. Because of the substantial social harm due to a much higher murder rate, it thus seems that if society were ever to employ only monetary sanctions to control murder, society would quickly realize that it would be rational to incur the costs of imprisonment in order to reduce the murder rate to a more acceptable level.

2.4 Use of imprisonment for acts in the major category of crime increases deterrence. The use of imprisonment increases deterrence of the major criminal acts from the inadequate level that would result from the threat of monetary sanctions alone. A person whose assets are too low to be deterred from theft or murder or treason may well be deterred by the prospect of imprisonment. This increased deterrence may well justify the use of imprisonment, despite the costs associated with its imposition.

2.5 But use of imprisonment sometimes fails to deter acts in the major category of crime, lending appeal to incapacitation. It is often true that even use of imprisonment is not enough to deter people from committing acts in the core area of crime; levels of crime are distinctly positive, and at some times in some places have been quite high.[116] This is not surprising. The likelihood of capture may be small, or at least perceived to be small, making the expected sanction less than the benefit; moreover, people may suffer lapses in their ability to weigh benefits against expected sanctions.

The individuals who commit criminal acts despite the threat of sanctions are individuals whom society may want to incapacitate. This factor adds to the appeal of imprisonment.

2.6 Other types of criminal acts -- categories (b) and (c) -- would be inadequately deterred if not labeled as criminal. Criminal acts that are concealed even if substantial harm was not intended (category (b)) are by definition relatively hard to deter. A hit-and-run driver is more difficult to deter than a driver who stays at the scene; a firm that pollutes and then tries to evade responsibility by destroying evidence is more difficult to deter than a firm that pollutes and does not conceal its act. Hence, higher sanctions are called for when concealment occurs, and possibly imprisonment. Further, even if imprisonment is not justified, the labeling of the acts

[115] See Statistical Abstract of the United States 2001, 447, which gives $9,000 as the median net worth of families with heads less than age 35.

[116] For example, in 2000, the urban robbery rate (all rates in this note are per 100,000 population) was 621 in Washington, D.C., and 224 in New York, as compared to 52 in Iowa and 27 in Vermont; the urban violent crime rate was 1,508 in Washington, D.C., and 831 in Florida, as compared to 107 in North Dakota and 180 in New Hampshire; see Morgan and Morgan 2002, 407, 425. Also, for another example, in 1997, the homicide rate was 57 in Washington, D.C., and 43 in Pretoria, as compared to about 2 each in Oslo, Lisbon, and London; see International Comparisons of Criminal Justice Statistics 1999 (May 2001), http://www.homeoffice.gov.uk/rds/pdfs/hosb601.pdf, table 1.2.

as criminal may be desirable because that augments deterrence due to the associated social sanctions, and the imposition of criminal fines may be useful because that too raises deterrence relative to civil sanctions (see section 1.2 above).

With regard to the residual category (c) of criminal acts, it seems that, as with category (b), the likelihood of sanctions is often low. Consider the crime of serving liquor to minors. The likelihood of sanctions for this act may be low, not necessarily due to active concealment by violators, but rather because of difficulty in determining when a violation has occurred (an establishment may serve liquor to a minor and not know this). Acts in the residual category (c) frequently seem to have the characteristic that the chance of imposing sanctions is significantly less than one hundered percent, and/or that the acts are either harmful in themselves or are likely to lead to harm. Whether solely monetary sanctions or nonmonetary sanctions as well will be called for will depend on other factors, as discussed generally earlier.

2.7 Different explanation of criminal law -- based on moral quality of acts -- versus present, functional explanation. Perhaps the major alternative explanation of criminal law is that it allows society to demarcate and to punish in a special way those acts that are deemed particularly morally offensive. I will not attempt to define here the moral quality of an act, but will rely on the reader's intuition as a guide. I will briefly suggest that the moral character of acts is unsatisfactory as a unitary explanation of criminal law, and is inferior to the functional explanation advanced above (although the true explanation of criminal law is undoubtedly not unitary).

(a) *Moral theory of criminal law is unable to explain why some criminal acts are less bad morally than some noncriminal acts.* It is evident that some criminal acts are morally less bad than some acts that are not criminal. Compare the crime of the theft of $100 worth of food from a supermarket by a hungry person, or the crime of a bartender unintentionally serving liquor to a minor, to a tort such as the calculated omission by a corporate officer of a warning about a product hazard that results in multiple deaths. Such comparisons are problematic for the moral-theoretical explanation of criminal law. Yet these comparisons are resolved by the functional theory, in that the sanctions of criminal law are needed to control theft and the serving of liquor to minors, but are not generally needed to control corporate torts such as failure to warn of product hazards (because corporations ordinarily have the assets to be deterred tolerably well by solely monetary, civil sanctions).

(b) *Moral theory fails to explain important characteristics of criminal law.* The moral theory does not explain certain significant features of criminal law, whereas the functional theory does. I will illustrate with two examples. First, consider that under criminal law, an attempt, such as shooting at someone but missing, is punished. It is incumbent on the moral theorist to say why such attempts should be sanctioned even though they do not result in harm, whereas under tort law, a very dangerous wrongful act, such as negligently leaving a live wire exposed at a playground, will not be sanctioned if it does not result in harm. As I will explain below in section 4.2, there is a functional explanation for this difference between criminal and tort law (based on the need to enhance deterrence in the criminal context, but not in the tort context, by means of punishing dangerous acts that do not result in harm). Second, consider that a basic feature of criminal law is that a victim and an offender are not allowed to settle their differences privately, whereas in civil disputes private settlement is permitted (indeed, encouraged). Why this difference should exist is not clear on moral grounds, whereas a straightforward functional explanation is offered for it in section 4.12 (based on the dilution of deterrence that settlement would engender in the criminal context).

(c) *Moral theory does not address the fact that many present-day crimes were punished primarily by fines or equivalents in the past, that tort and criminal law were not distinct.* Historically, tort and criminal law were not separate. Rather, penalties denominated in terms of wealth (in money or goods), paid to victims, often according to a schedule, were employed for undesirable acts, including those that today would be considered criminal, such as murder and rape; prisons were not used.[117] That for a long time societies did not formally distinguish crime from torts requires explanation. The moral theory does not offer obvious possibilities, assuming, as I do, that basic attitudes about what acts are especially bad were similar in the past to what they are today. The functional theory, however, does provide possible explanations for why a system based on money and wealth sanctions could have worked reasonably well in the past to prevent what we today call crimes. One conjecture is that the likelihood of escape for bad acts was much smaller in the past than in our modern, anonymous society, so that the magnitude of wealth penalties necessary to deter may have been smaller. Another conjecture is that, because wealth penalties were often imposed on kinship groups that had the capacity to pay (and which could exert pressure on the particular offender), the problem of the judgment-proof offender that I have emphasized as the major explanation of the need for criminal law now may not have been severe then. In addition, one supposes that informal social sanctions operated with greater effect than in present day society, lessening the need for criminal sanctions. Moreover, the institution of prisons may have been excessively costly for societies in the past, as they were generally much less wealthy; building and operating prisons, and taking people out of the labor force, might have been close to an unthinkable economic burden for most societies over the course of history. In sum, the lesser social need to develop a separate criminal law and the relatively high cost of establishing a system of imprisonment may have been such that it was socially rational not to distinguish in a formal and self-conscious way tort law from criminal law as we know it.[118] Although frankly speculative, this line of reasoning illustrates the ability of the functional theory to explain why criminal law and tort law were not separate in the past.

(d) *Conclusion.* The moral-theoretical explanation for criminal law seems inferior to the functional explanation. This is hardly to deny, however, that there is a general congruence between criminal acts and morally offensive acts, nor that there exist important relationships between criminal law and our system of morality (on which see chapters 26 and 27).

3. Optimal Use of Imprisonment Reviewed

[117] See, for example, Berman 1983, 53-56, and Pollock and Maitland 1911, volume 2, chapter 8, section 1. The payments were apparently often enforced by subjecting a violator who refused to pay to blood-feud, or by declaring him an outlaw; see, for instance, the section cited in Pollock and Maitland.

[118] To be clear, I am not saying that, for example, an accidental killing would have been viewed in the same way as murder in former times; I suppose that the two acts would have been seen as quite different. (For example, as Pollock and Maitland 1911, 2:450-52, suggest, whether a slayer would have the option to pay for the death, rather than be subject to blood-feud or outlawry if the victim's kin wanted that, may at times have depended on the nature of the killing.) What I am saying is that there was no need for a distinctly different legal treatment of torts and crimes, no need comparable to ours, in the former period. The institution of criminal law is a product of our times, not an intrinsic feature of the legal system.

Having attempted to explain why criminal law exists and why the domain of criminal acts is what it is, I want to review here the nature of the theoretically desirable use of the sanction of imprisonment from the point of view of both deterrence and incapacitation. This will be referred to in the next section, where I examine important doctrines of criminal law.

3.1 Optimal deterrence and imprisonment. The point developed in chapter 21 was that imprisonment, being costly to impose, should be employed so that it accomplishes deterrence at a low cost. This implies that the socially desirable sanction for an act is that which would be just sufficient to deter most of those who would tend to commit the act; thus, the sanction should be such that the expected sanction should just outweigh the expected benefit that most potential offenders would obtain from the act. (A higher sanction would not, by hypothesis, be needed to deter most in the group, but some in the group could not or would not be deterred, and would commit the act; for them, imposing a greater sanction would mean that society would bear a larger cost.) Further implications about the optimal use of imprisonment to deter socially undesirable acts are as follows.

(a) The sanction should be higher the greater the probability or the magnitude of harm due to the act, for the greater the expected harm, the more socially worthwhile it will be to increase deterrence despite the higher social cost of using sanctions.

(b) The sanction should be higher the greater is the private benefit the actor obtains, as long as the benefit appears to be within the range that allows for the possibility of deterrence. This follows because the sanction should be just high enough to deter.

(c) The sanction should be zero, or small, if the actor appears to be impossible to deter.

(d) The sanction should be higher the lower is the probability of apprehending the actor, provided that there is a possibility of deterrence. This follows because, in order to create an expected sanction necessary to deter, the actual sanction must rise if the probability of its imposition falls.

3.2 Optimal incapacitation and imprisonment. The point emphasized about incapacitation in chapter 23 was that a person who is apprehended should be imprisoned and thereby incapacitated as long as the expected harm he would do per period if free exceeds the cost per period of imprisonment. This implies that the character of the act a person committed is relevant in so far as it provides information about a person's future propensity to do harm. If the harmfulness of acts that a person committed is predictive of the person's future harmfulness, then if the committed act exceeds a threshold in seriousness, he should be imprisoned.

It should be noted that, from the perspective of incapacitation, the probability of apprehension is irrelevant to the optimal sanction. Also, the ability to deter the actor is irrelevant; if someone could not have been deterred, the person should still be imprisoned if his future dangerousness is sufficiently high.

4. Principles of Criminal Law

In this part, I will describe important principles and doctrines of criminal law, and analyze them in light of the theory concerning optimal deterrence and optimal incapacitation. The major principles and doctrines of concern will be intent, attempt, causation, and a variety of defenses to criminal liability.

4.1 Intent. A central feature of the criminal law is the emphasis it places on intent. To analyze intent, it is best to begin by making several definitions. Let us say that a party "desires" a result if it would either directly or indirectly raise his utility.[119] Let us also say that a party "intends" a result if he (a) desires the result and (b) acts in a way that he believes will raise the probability of the result.[120] This definition of intent seems to comport with its ordinary meaning.[121] According to the definition, we would say that X intended that Y die if he desired Y's death and shot at Y and killed him. If, however, X shot at Y but instead struck Z whose death he did not desire, we would not say that he intended that Z die. Also, we would not say that X intended that Y die if X desired Y's death, played golf with Q, and Y happened to be killed in an automobile accident (since the round of golf with Q did not increase the probability of Y's death).

In criminal law, the role of intent, as I have defined the term, may be summarized by several statements.[122] First, intent to do harm is ordinarily a principal factor in determining liability and the severity of punishment. Second, the effect of intent on liability and punishment is generally the same whether an intended harmful result is directly desired or indirectly desired. Whether X shot Y because Y was his enemy or only because Y stood in the way of an inheritance will not ordinarily affect the punishment of X under the law. Third, whether a harmful result different from the desired result occurs does not usually influence a party's legal treatment. When X aims at Y but shoots Z instead, it is murder just as if X aims at Z.

These features describing the role of intent in criminal law are roughly consistent with the purposes of deterrence, for intent appears to be linked to the factors that, according to theory, call for, or increase, the level of, sanctions.[123] Intent is, first of all, positively related to the probability of harm, for when a party intends to do harm, he acts so as to raise the probability of harm. This factor is particularly significant when the courts' direct evidence about the probability of harm is limited, because courts can often make inferences about the probability of harm from knowledge of a party's desires. For instance, when a court has little evidence about X's shooting of Y but knows that X had the purpose of killing Y, the court might infer that X carefully drew a bead on Y.[124] Second, intent may be correlated with the likely magnitude of harm, because a

[119] In the language of utility theory, a result is desired (a) if the result is an argument in the individual's utility function and thus would raise his utility in a direct manner, or (b) if the result would lead to an increase in his expected utility because it is correlated or associated with a change in an argument in his utility function.

[120] The significance of erroneous beliefs is discussed below. For now, I assume that the party's beliefs about the probability are correct.

[121] The traditional definition of intent in the criminal law is broader: A party "intends" a result even where he does not desire it if he acts in a way that makes it highly probable (rather than only more probable).

[122] See, for example, LaFave 2000, 229-41.

[123] Oliver Wendell Holmes, Jr., was one of the first writers to try to establish a connection between intent and factors that ought to increase the sanction appropriate for deterrence. His discussion focused on the relationship between intent and the probability of doing harm. See Holmes [1881] 1963, 52-62.

[124] It is worth developing this example in more detail. Suppose that X and Y were hunting together when X shot Y. X claims that he fired at a deer running between him and Y and unfortunately did not see Y. A witness who was standing at some distance away confirms that there was a deer running between X and Y, but he is not able to say whether X noticed Y or aimed at him rather than at the deer. With only this very imperfect knowledge of X's act, a court could highly value

party who desires a harmful result is prone to do greater damage than one who does not. X will be more likely to shoot Y in a vital spot than in an arm or a leg if X desires to harm Y. Intent is also closely associated with the private benefits that parties expect to derive from their acts. By definition, the utility of parties who intend harm is raised by the occurrence of harm, and as just indicated, both the probability and magnitude of such desired harm tend to be higher when there is intent. Thus, parties who intend to do harm will be more difficult to deter.[125] Finally, intent may be linked to the probability that a party will escape a sanction, since a party who intends to commit a harmful act is more likely to choose a particular place and time to commit the act so as to avoid identification and arrest, or to take steps thereafter to do so.[126]

These arguments suggest why intent, though mainly a mental factor, ought to influence liability and punishment according to deterrence theory. Moreover, it should be noted that the arguments do not depend on whether the intended harm is directly or indirectly desired. In either case the probability and magnitude of harm, the expected private benefits, and the likelihood of escaping sanctions are likely to be higher than for unintentional conduct. Further, the arguments are largely unaffected by whether the actual result was the same as the desired result. It therefore makes sense that such distinctions usually do not affect a party's punishment.

From the standpoint of optimal incapacitation, intent is significant in so far as it provides information about the future dangerousness of a person. Is a person who commits an act in which he intends harm going to be dangerous in the future? As a general matter, and as a crude approximation to the truth, the answer seems to be in the affirmative; we infer something about the character of a person when we learn that he intended to cause harm, and this leads us to increase our estimate of the probability that the person will do harm in the future. However, the importance of this factor depends on the particulars of the case. If husband X murders his wife in order to be free to marry his lover, and the circumstances leading to this act are unlikely to repeat themselves, then his intent per se would not seem to signify much about future dangerousness. On the other hand, if Y, who has never had a full time job, intends to and does carry out a robbery, we would surmise that his intentional behavior does suggest future danger, because the circumstances that gave rise to his actions are likely to apply again; he is likely to want more money in the future and to be able to steal to obtain it. In each case, intent tells us something about the character of the individual that is relevant to predicting future behavior, but what it

information about X's intent (for example, evidence that he would profit from Y's death and planned to kill Y at a good opportunity) or lack thereof (for example, evidence that X had nothing to gain from Y's death). Thus, knowledge of X's intent may alter a court's assessment of the probability of harm due to X's act. If however, the court's direct knowledge of X's act were complete (for example, suppose the court possessed a close-up video-recording of his behavior), it would not need to know anything about intent in order to assess the probability of harm. But, as will be seen, the court might well find knowledge of intent valuable for other reasons.

[125] If X intends to kill Y, it will be difficult to deter him, because he wants Y dead and because shooting at Y makes this result likely. By contrast, if X is a true friend of Y, to deter a negligent or reckless shot will not require a substantial sanction (if it requires any sanction at all).

[126] In some cases, however, the factor of intent could increase the probability of sanctions because a person's motives might be discoverable and lead police to investigate him. The importance of this consideration depends on the type of crime and the particular case. It might be significant in some cases of murder, for instance, but would probably not be in most cases of theft and robbery.

tells us is very partial in nature. Thus, we may conclude that intent has relevance for the need to incapacitate, but it does not seem that we can explain the importance given to intent mainly through appeal to incapacitation.

Consideration of situations in which a party is not liable despite his intent to do harm sheds further light on intent with respect both to deterrence and to incapacitation.[127] A party may intend to do harm but escape liability because circumstances make his act socially desirable. For example, a party forced to choose between two harmful acts may invoke the defense of necessity if he chooses the less harmful act. In addition, a party may act under duress and escape liability; here deterrence is difficult or impossible and there is no reason to incapacitate, so imposing sanctions would not be socially worthwhile.

Conversely, parties are sometimes punished despite their lack of intent to do harm. When a party does not desire a harmful result but acts in such a way that serious harm becomes very likely (suppose that a drunk person drives at 90 miles per hour through a school zone and runs over a child), he may be punished under criminal law. Imposition of sanctions here may be justified because the expected harm is high; the fact that the party does not desire the harm does not make his behavior less dangerous. Similarly, when a party does not desire harm but commits a strict liability crime, his punishment may be justified in principle if the courts find it very difficult to differentiate between desirable and undesirable acts.

4.2 Attempt. The criminal law punishes attempts to do harm. If a person shoots at another but misses, if he picks a pocket that turns out to be empty, if he is found with a forged check but has not yet cashed it, he is guilty of a crime of attempt even though he has not done harm.[128]

The punishment of attempts serves to enhance deterrence, because it effectively raises the probability of sanctions for potentially harmful acts: A person who commits a potentially harmful act faces the prospect of sanctions not only if his act turns out to cause harm, but also if it does not and constitutes only an attempt.[129] Moreover, the punishment of attempts is a socially inexpensive means of increasing the probability of sanctions, for opportunities to punish attempts often arise as a byproduct of society's investment to apprehend parties who actually do cause harm.[130] Hence, it can be argued that punishment of attempts is socially desirable from the

[127] That there should be such situations is not surprising a priori, for intent was only said to be linked with factors leading to the optimality of liability.

[128] See, for example, LaFave 2000, 535-67. As mentioned in section 2.7, this is in contrast to the situation in tort law, where there is no liability unless harm is done. See Keeton et. al. 1984, section 30.

[129] In the tort context, punishing the analogue of attempts -- negligent acts that do not result in harm -- would also raise the probability of sanctions and enhance deterrence. But in the tort context, the need to enhance deterrence is much less than that in the criminal context, for in the tort context parties typically do not escape suit with high probability. Hence, in the tort context, making parties pay money damages only when they actually do harm, and in an amount equal to the harm, should tend to create adequate incentives to reduce harm, as is emphasized in chapter 8.

[130] Given that the police stand ready to apprehend those who do harm (by giving chase, investigating suspicious behavior, and the like), apprehending individuals who commit unsuccessful attempts may not involve substantial marginal cost. At least the added cost of raising the probability of sanctions by apprehending those who commit unsuccessful attempts should be much less than the added cost of raising comparably the probability of sanctions by apprehending only more of those parties who succeed in causing harm.

standpoint of the theory of deterrence.[131]

The force of this argument for sanctioning attempts clearly increases with the likelihood that a party will be apprehended for an attempt. When an act takes a long time to execute (especially when it requires preparations) or when it has a substantial chance of not succeeding (for example, shooting from a distance), the probability of being caught for an attempt will be higher than otherwise. Therefore, the deterrent value of punishing attempts will also be higher.

With regard to the theory of incapacitation, it is evident that, to the degree that attempts signify future dangerousness, attempts call for punishment.

The possible desirability of punishing attempts according to the theory of deterrence or of incapacitation does not imply that there is an advantage in punishing attempts in the way that criminal law does, namely, less severely than acts that actually result in harm. In discussing this feature of criminal law, it is useful to consider separately two types of attempts that do not cause harm: interrupted attempts -- acts discovered before they could have succeeded -- and completed attempts -- acts that were brought to completion and thus that might have succeeded.

With respect to interrupted attempts and deterrence, the following argument is sometimes made.[132] If the sanction for an attempt is lower than that for doing harm, a party who begins an attempt might be induced to reevaluate and abandon it, since he then will be punished less. If, however, the sanction for the attempt is the same as for doing harm, he may as well continue. As stated, this argument fails to recognize the possibility of treating the abandoned attempt leniently, while imposing a full sanction on attempts that are not abandoned but only interrupted by others. Suppose, for instance, that no sanctions are imposed for abandoned attempts and that the sanction for an interrupted attempt is the same as the sanction for an act that causes harm. Then the party who sets out to commit a harmful act will certainly have reason to abandon it; not only will he escape sanctions, but he will otherwise face a sanction for a later interrupted attempt equal to the sanction for doing harm.[133]

Nevertheless, punishing interrupted attempts less severely than acts that result in harm may be advantageous under both deterrence and incapacitation theory. Because interrupted attempts may later be abandoned or fail, there is less evidence of the dangerousness of interrupted attempts and thus less reason for sanctioning them than acts that do result in harm.[134] The significance of this argument plainly depends on the character of the attempt and the point at which it is interrupted. If an attempt is nearly complete and is likely to succeed, the argument does not carry much weight. That would be so, for instance, where a person had already dropped

[131] This thesis, and the arguments sketched below, are developed in Shavell 1990. Among other things, that article explains why raising the probability of sanctions by punishing attempts is advantageous, given the apparent alternative of a strategy of imposing a higher level of sanctions but imposing them only when harm is done. The essence of the explanation is that raising the magnitude of sanctions may have various disadvantageous effects (such as distorting marginal deterrence) and may not be workable because of the upper limit on sanctions.

[132] See, notably, Beccaria [1767] 1995, 95.

[133] However, this argument presumes that courts are able to distinguish between abandoned and interrupted attempts.

[134] This the reader will recognize as a version of the general argument advanced earlier that the actual harm done might influence the sanction because of the court's incomplete information about the dangerousness of an act.

a lethal dose of poison into his intended victim's drink. An attempt interrupted further from completion might properly be sanctioned less severely, however. Indeed, an attempt might reasonably escape a sanction altogether if it is interrupted so early that there is great doubt whether and in what manner it would have been continued. Thus, if a person was apprehended merely when leaving a drugstore with poison, it might be unclear whether he would have used the poison, and unclear too whether his behavior would satisfy the definition of attempt in criminal law.

Two arguments analogous to those just discussed are often advanced to justify imposition of   lower sanctions for unsuccessful completed attempts than for acts that succeed in causing harm. First, it is asserted that if the sanction for an unsuccessful completed attempt is equal to the sanction for a successful attempt, a party whose initial attempt fails will have nothing to lose by trying again. This argument overlooks the fact that the sanction for an initial unsuccessful attempt may equal the sanction for an initial successful attempt, and yet the party will have something to lose by trying a second time as long as the sanction for a second successful attempt is higher. For example, if the sanction for an initial attempt is a sentence of five years whether or not it succeeds, but the sanction for causing harm on a second attempt is ten years, a party who at first fails will clearly have reason not to try a second time.[135]

The other argument for punishing unsuccessful completed attempts less severely than those resulting in harm is that the failure of an attempt may constitute evidence that it was a less dangerous act. As with interrupted attempts, the strength of this evidentiary rationale depends on the nature of the attempt. And while one can think of situations in which the rationale would be important, in many that come to mind, it does not seem so.[136]

Finally, it is interesting to consider attempts that cannot possibly succeed.[137] There are two types of such attempts. The first, for which it is often said liability should not be imposed, is exemplified by the case of a person who sticks pins in a voodoo doll, intending to kill his enemy.[138] Here an objective observer might say that the type of act committed never causes harm, so that there is no reason to deter the act or to incapacitate the actor.[139] The second kind of attempt, for which there would be liability, is illustrated by the case of a person who shoots a

[135] The argument in this paragraph presumes that courts can determine if a party repeats an attempt.

[136] For instance, if a person puts poison in his intended victim's drink but the victim fails to succumb, it is true that this is some evidence that the act was less dangerous than one that produced a death -- perhaps because the dosage of poison was too low. But this might not constitute enough evidence to lower the sanction significantly. In any event, there is less reason to lower the sanction than if the person had been interrupted before he completed the attempt, for then there would have been doubt about whether the attempt would have been completed, as well as whether it would have been successful if completed.

[137] See, for example, LaFave 2000, 552-60.

[138] Note that because this person believes he is raising the probability of his enemy's death, we would say he intends his death under the definition of intent used in this chapter.

[139] This presumes that the person who failed with the voodoo doll would not have tried other ways of killing his enemy, such as shooting at him. If there is evidence that the person would have turned to other methods, then his act would appropriately be defined as "trying to kill an enemy by some means" rather than "trying to kill an enemy using a voodoo doll," and there would be a reason to punish him.

bullet into a dummy that he thinks is his enemy. In this instance, an objective observer would say that the type of act committed creates positive expected harm, for shooting things that appear to be human beings will usually result in harm. Hence, according to both deterrence and incapacitation theory, the act should be punished.

4.3 Causation. When a party's act is followed by harm, two causal issues may arise.[140] The first concerns the question of whether the act was the "necessary" cause of the harm, that is, whether the harm would not have occurred but for the act. Thus, if X poisons Y's drink and Y then dies, yet an autopsy reveals that Y coincidentally died of a heart attack before he could have succumbed to the poison, X's act would not be the necessary cause of Y's death. The criminal law ordinarily treats an act that was not the necessary cause of harm as if it were an attempt: The party is punished for the act, but less so than if the act were the necessary cause of the harm.[141]

This outcome makes sense -- though only partial sense -- according to deterrence theory. It makes sense that acts that are not the necessary causes of harm are punished, for this enhances deterrence in the same way that punishment of attempts does, namely, by increasing the probability of sanctions.[142] The reason for the imposition of lesser sanctions, however, is not apparent. That acts sufficient to cause harm turn out not to be the necessary causes of harm is happenstance; it does not mean that the expected harmfulness of the acts was any lower. Therefore, the sanctions for such acts ought not to be diminished, according to deterrence theory.[143]

If a party's act was the necessary cause of harm, the legal issue arises of whether his act was the "proximate cause" of harm. Generally, acts said to be the proximate cause of harm can be recognized as acts that substantially increased the probability or magnitude of the type of harm that occurred. To illustrate, if X severely beats Y and Y later dies in the hospital from internal injuries, it would probably be said that X's actions were the proximate cause of Y's death. If, however, Y dies in an automobile accident while being taken to the hospital after the beating, it would probably not be said that X's actions were the proximate cause of Y's death.[144] In determining punishment, the criminal law usually takes into account whether or not the harm was proximately caused. X might be held liable for murder when Y dies from his internal injuries, but not when Y dies in the automobile accident.

Punishing a party for harm that he proximately caused -- and not just for the act he

[140] See, for example, LaFave 2000, 292-320.

[141] This situation is different under tort law, where a party usually escapes liability if his act was not the necessary cause of harm. See Keeton et al. 1984, 265.

[142] In the usual model of torts, there is no reason to enhance deterrence by use of sanctions when a party's act is not the necessary cause of harm. The threat of liability only when their acts are necessary causes of harm is enough to induce parties to take adequate care, assuming that suit will be brought when parties are liable. See chapter 5 of Shavell 1987a.

[143] The fact that Y died of a heart attack does not cast doubt on the potency of the poison. Note that this is in contrast to a failed or interrupted attempt to murder Y by poisoning.

[144] The probability of dying in an automobile accident on a single trip (even if by ambulance) is small, and is therefore not much increased by X's beating of Y. Indeed, Y might have been going somewhere else by automobile if he had not been going to the hospital, in which case Y's chance of dying in an automobile accident would not have been raised by X's beating him.

committed, as the act is otherwise understood by the court -- might sometimes be justified in view of the evidentiary value of the actual harm done for the assessment of the act by the court. Y's death from internal injuries would be indicative of the severity of the beating he received, whereas his death in an automobile accident would not convey such information.[145]

The implications of incapacitation theory are similar to those just discussed. Namely, an act that turns out not to be the necessary cause of harm should be punished as much as a similar act that does cause harm. For when X poisons Y's drink and Y happens to die of a heart attack, we generally have as much information about the dangerousness of the act and the actor as we would have if Y had not died of a heart attack. However, to the degree that proximate causation implicitly supplies us with information about the expected harmfulness of the act, such causation is relevant to proper punishment for purposes of incapacitation.

4.4 Responsibility. Under criminal law, the imposition of sanctions depends on whether a person who commits a harmful act is deemed "responsible," in whole or in part, for his behavior. Major reasons why a person may not be held responsible are insanity, automatism, involuntary intoxication, or youth. If these reasons apply, the person's liability may be diminished or eliminated.

This aspect of criminal law has an obvious potential justification according to deterrence theory, as the conditions that reduce or relieve one's responsibility for otherwise criminal acts make it unlikely that the use of sanctions would accomplish significant deterrence.[146] An insane or involuntarily intoxicated person, for example, cannot be deterred from committing certain acts by the threat of punishment, so that the elimination of punishment might be appropriate. However, two general reasons for restricting such escape from liability suggest themselves. First, individuals may often be able to feign successfully the conditions that limit their responsibility. This possibility may be significant with respect to some of the conditions (for example, insanity), but not for all (youth is difficult or impossible to pretend). Second, individuals may sometimes choose to act in ways that make their (true) conditions especially dangerous. An epileptic might drive an automobile, or a person subject to insane rages might decide to purchase a gun. The imposition of liability could induce these individuals to act differently and thereby to reduce dangers over which they later would have no control.

With regard to incapacitation theory, it is apparent that, in many cases where a person is not legally responsible for his act, he will be no less dangerous to society than if he were responsible. A person who has an uncontrollable urge to set fires, or who is subject to insane, violent rages, is dangerous to society even though he cannot help himself. Therefore, his lack of responsibility does not diminish the need to incapacitate him. Hence, contrary to the implication of deterrence theory, sanctions are called for from the point of view of incapacitation. (Of course, the form of incapacitation need not be incarceration in a prison; it could be confinement to a facility for the criminally insane.)

[145] That the logic of this paragraph applies generally and is not a feature of my example can be appreciated from the characterization of proximately caused harms as those whose probability or severity was increased in a substantial way by a party's act. It is exactly when this is true that the occurrence of the outcome may convey useful information about how much the party's act increased the expected harm.

[146] This point is stressed by Bentham [1789] 1973, 164.

In some cases, lack of responsibility might not imply future dangerousness, or at least not dangerousness sufficient to warrant punishment. If youth is the reason for lack of responsibility, it might be felt that with time and maturation, the person would be unlikely to commit a similar act, and certainly with involuntary intoxication that would be the case.

4.5 Ignorance of the law.[147] If a person claims that he was unaware that his act was unlawful, he will ordinarily be found liable anyway. However, he may sometimes escape liability if he had little opportunity to learn about the law (as with an unpublished or little-known ordinance) or if he was acting in reliance on a mistaken interpretation of the law made by a court or an appropriate government officer.[148]

Such an approach is consonant with the theory of deterrence.[149] If a person is held liable for violating well-appreciated laws or laws that can be learned through reasonable effort, whether or not he admits that he knows such laws, he will not be able to escape liability by pretending not to understand the laws and will have an incentive to learn the laws and adhere to them. If, however, a reasonable effort is insufficient to learn a legal rule, it is best to permit parties to escape liability, since a party can be deterred by possible sanctions only if he knows which acts will lead to the application of sanctions.

4.6 Mistake. A person may commit an act that he believes to be innocent although it is actually harmful. In a classic case, a person takes an umbrella from a restaurant assuming that it is his, when it really belongs to someone else. There is no criminal liability in such instances.[150] This feature of criminal law makes sense, for people cannot be deterred from committing acts that they are unaware are harmful. Moreover, assuming that acts believed to be harmless usually are harmless, the expected harm associated with the acts is too low to warrant use of sanctions.[151]

A related type of mistake arises when a party knowingly commits an undesirable act but believes it to be either more or less harmful than it actually is. For example, an individual might steal a valuable piece of jewelry, thinking it a mere bauble, or he might shoot to kill a "person" who turns out to be a dummy (as I mentioned in the section on attempt). The legal principles employed in these situations are not uniform; although sanctions are frequently based on what a party did in fact, many times they are affected by what he thought he was doing.

There are two reasons why the harmfulness of the act the party thought he was committing should influence the sanction under deterrence theory. First, the benefits an

[147] In this and the next several sections , I will discuss various justifications and excuses for committing harmful acts that lead to escape from criminal liability. They are considered separately from responsibility, and from each other, since they present different issues.

[148] See, for example, LaFave 2000, 432-34, 440-49.

[149] See Bentham [1789] 1973, 164.

[150] See, for example, LaFave 2000, 432-37.

[151] Note that the situation under discussion in this paragraph, of not knowing that an act is harmful because of lack of knowledge of some circumstance (who owns the umbrella), is analytically indistinguishable from that of the last section, of lack of knowledge of the law. In this vein, it should be noted that a person who could easily have determined that his act was not innocent (say the umbrella he took was obviously not his -- it had another person's monogram on it) might not be able to escape liability.

individual expects to derive from committing an act, and thus the ability to deter him, depend on what he thinks he is doing, not on what he is in fact doing.[152] Second, the expected harm associated with an act may be more closely related to what the party thinks it is than to what it turns out to be in the particular instance. The act of taking what one thinks is a bauble might usually mean that only a bauble is missing; the act of shooting at what one thinks is a person will usually be very harmful. This second reason is also applicable according to incapacitation theory; if the future dangerousness of a person is more closely associated with the act the party thought he was committing (shooting at what he thought was a person), then that should guide the sanction, and not the harm actually done.

Nevertheless, one important factor suggests that the actual harm should influence the sanction. Individuals may be able to convince the courts that they thought they were doing little harm when in truth they knew they were doing greater harm. If so, and if the sanction is based on the courts' erroneous assessment of parties' beliefs, sanctions will be too low and diminished deterrence will result.[153] Hence, there is some reason to raise the sanction when the actual harm exceeds what the wrongdoer claims to have thought it would be. But note that there is no corresponding argument for lowering the sanction when the actual harm turns out to be less than what the individual thought it would be, since he will have no incentive to exaggerate the harm he thought he was doing.

4.7 Entrapment. A person may raise the defense of entrapment if a law enforcement official induces him to commit a criminal act that he would not ordinarily commit.[154] When, for instance, a game warden induces a hunter to shoot at bald eagles and the hunter would not otherwise have done this, the hunter can assert the defense of entrapment.

The argument for this defense on grounds of deterrence theory is that if persons would not ordinarily commit criminal acts, there is no behavior that needs to be deterred. Similarly, according to incapacitation theory, if individuals do not have a general tendency to cause harm, but cause it only in the restricted and unusual circumstances of entrapment, they do not represent a future danger to society. Thus, under either theory, punishment of the individuals, and effort devoted to their entrapment, must be considered a social waste;[155] moreover, their entrapment might also result in the actual doing of harm.[156] Hence, it is best not to punish the parties, and to discourage entrapment activity. The former is directly accomplished by allowing the entrapment defense; and the latter is indirectly accomplished by allowing the defense, since enforcement officials will not then derive the benefit of securing additional criminal convictions.

[152] The individual who thinks he is stealing a bauble, and thus not obtaining much of value, might be easier to deter than the individual who thinks he is stealing valuable jewelry.

[153] Individuals may not be properly deterred from stealing valuable jewelry if they know they can convince the courts that they thought the jewelry was only a bauble.

[154] See, for example, LaFave 2000, 449-66.

[155] To clarify this point, consider a situation in which a person would never commit a criminal act if not entrapped. Here, plainly, punishing the person and devoting effort to entrap him is wasteful, since otherwise the person would never cause harm. (It is irrelevant under deterrence theory that the person might be thought bad because he could be induced to commit a criminal act in certain constructed circumstances.)

[156] For example, the game warden might not be able to take the hunter into custody before he shoots a bald eagle.

However, the defense of entrapment may not be justifiable when individuals would often commit by themselves the criminal acts that they are led to commit by an enforcement agent. In such cases, it is by hypothesis desirable to deter the individuals or to incapacitate them. Therefore, it may be useful to employ certain law enforcement activity, including deception and subterfuge leading to the inducement of criminal acts, in order to raise adequately the probability of sanctioning the individuals.

4.8 Duress. A person will not be held liable for a harmful act if he committed the act only because of duress -- a threat of serious injury or of death. To invoke the defense of duress, the threat to a person must have been both imminent and credible, and the person must not have killed someone (although the sanction may still be mitigated in that case). Whether or not the defense is available, the threatening party will be liable for the act he induced.[157]

The defense of duress is obviously desirable if the threatened party truly cannot be deterred by the prospect of a legal sanction for committing the act.[158] Hence, the law's insistence on the imminence and credibility of the threat, and on its being one of serious injury or of death, seems understandable. But its refusal to allow the defense when the threatened party has killed someone does not seem rational according to deterrence theory, as it is quite possible that the threatened party could not have been deterred from killing; after all, he will often be comparing an immediate threat to a sanction that will not be immediate, if it is imposed at all.

According to incapacitation theory, the defense of duress is also warranted, for the party who is forced by a threat to commit a harmful act probably does not represent a future threat to society.

Moreover, it is desirable that an individual who makes a threat be held liable for crimes committed as a result of that threat, for this will be necessary to deter and/or incapacitate such individuals.

4.9 Necessity. The defense of necessity may be asserted when an individual, forced by circumstances to choose between two harmful acts, chooses the less harmful act.[159] This makes clear sense according to deterrence theory, as it is socially desirable for a party to minimize harm. Furthermore, the defense is rational according to incapacitation theory for similar reasons; the individual who would choose to do the lesser of two harms does not pose a danger to society, quite the opposite.

4.10 Defense of self, of another, or of property. The law regarding self-defense and protection of others and of property is, roughly, that one may use the amount of force apparently necessary to ward off an aggressor whose threat one believes is unlawful and immediate, and who cannot be stopped by police intervention.[160] Plainly, allowing the use of force will enhance deterrence of aggression. Limiting the justified use of force makes sense under the presumption that the courts are better able than threatened parties to decide on sanctions.

[157] See, for example, LaFave 2000, 467-76.

[158] This is pointed out by Bentham [1789] 1973, 165.

[159] See, for example, LaFave 2000, 476-86.

[160] See, for example, LaFave 2000, 491-508.

4.11 Consent. A person may sometimes escape criminal liability if the individual affected by his act had consented to its commission. The defense of consent, however, is not available when serious bodily injury is done.[161] The justification for the defense, under deterrence or incapacitation theory, centers on the concept of harm: If consent is taken to mean that there is no harm, then there is no reason to deter acts to which someone has consented.[162] According to this reasoning, a person's consent even to serious bodily injury apparently ought to be allowed as a defense. Yet consideration must be given to the counterarguments that the person who consents may not properly evaluate his situation, that his family and friends may be affected by the contemplated act, and that the injuring party may deceive the courts about the victim's consent.

4.12 Condonation and settlement. The fact that a person who has suffered harm may later condone, or settle with, the individual who is responsible for the harm[163] may not be used as a defense against criminal liability.[164] If a person is robbed, for example, and he then discovers the robber and forgives him, the robber will still be subject to punishment.[165] According to deterrence and incapacitation theory, the major reason for not allowing condonation as a defense to criminal liability is that deterrence would be diluted and incapacitation negated. Were the defense allowed, the "sanctions" imposed by victims -- usually some form of apology, the return of property, or a payment, but never imprisonment, would be less than the sanctions the courts would otherwise impose.[166] Moreover, victims might many times wish to condone injuring parties, for there is no reason to believe that a victim's personal interest in punishing an injuring party would generally correspond to the social interest in deterrence or in incapacitation. (This point is closely related to that about the divergence between the private and social incentives to litigate discussed in chapter 17.) Finally, were the defense allowed, there might be a real danger that victims would be coerced into "condoning" injuring parties.[167]

Note on the literature. Economic analysis of the criminal law began with Beccaria

[161] See, for example, LaFave 2000, 516-19.

[162] See Bentham [1789] 1973, 163.

[163] Condonation is distinct from giving the party prior consent; it occurs after the harm is done.

[164] See, for example, LaFave 2000, 521-23.

[165] As a practical matter, however, prosecution may be difficult if the victim is reluctant to provide testimony about the crime.

[166] It should be observed that in tort law, where of course settlement is allowed, payments made in settlement would often approximate the expected court-determined monetary sanctions, since otherwise victims would tend not to want to settle.

[167] The argument of this paragraph, that allowing settlements between injurers and victims would compromise public law enforcement, did not clearly apply before the development of effective mechanisms of public law enforcement. For an interesting illustration of this point, see Klerman 2001, who emphasizes that in thirteenth century England, during which private prosecutions of crime were usually necessary to bring wrongdoers to justice, it was found that when courts frowned on settlements, private prosecutions declined, and many wrongdoers undesirably escaped sanction; thus courts were led to respect settlements between injurers and victims for a period.

([1767] 1995) and, especially, Bentham ([1789] 1973), who succinctly made general, major points about restricting use of sanctions to situations where they would accomplish deterrence. Holmes ([1881] 1963) contains a chapter on criminal law with insightful remarks about deterrence and, especially, attempt and intent. Posner (1985a) and Shavell (1985a) contain brief examinations of the doctrines of criminal law from an economic perspective.[168]

[168] See also, for example, Ben-Shahar 1998 and Shavell 1990 on attempt, and Cohen 1989, Fischel and Sykes 1996, Khanna 1996, and Lott 2000 on corporate crime.

References

Andenaes, Johannes. 1966. The General Preventive Effects of Punishment. *University of Pennsylvania Law Review* 114:949-83.

----- 1975. General Prevention Revisited: Research and Policy Implications. *Journal of Criminal Law and Criminology* 66:338-65.

----- 1983. Deterrence. In *Encyclopedia of Crime and Justice,* edited by Sanford H. Kadish, 2:591-97. New York: Free Press.

Anderson, David A. 1999. The Aggregate Burden of Crime. *Journal of Law and Economics* 42:611-42.

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. Tax Compliance. *Journal of Economic Literature* 36:818-60.

Baker, John. H. 2002. *An Introduction to English Legal History*. Fourth edition. London: Butterworths.

Beattie, J. M. 1986. *Crime and the Courts in England, 1660-1800*. Princeton: Princeton University Press.

Bebchuk, Lucian A., and Louis Kaplow. 1992. Optimal Sanctions When Individuals Are Imperfectly Informed about the Probability of Apprehension. *Journal of Legal Studies* 21:365-70.

Beccaria, Cesare. [1767] 1995. *On Crimes and Punishments, and Other Writings*. Edited by Richard Bellamy, translated by Richard Davies, with Virginia Cox and Richard Bellamy. New York: Cambridge University Press.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76:169-217.

Ben-Shahar, Omri. 1998. Criminal Attempts. In *The New Palgrave Dictionary of Economics and the Law*, edited by Peter Newman, 1:546-50. London: Macmillan.

Bentham, Jeremy. [1789] 1973. *An Introduction to the Principles of Morals and Legislation*. In *The Utilitarians*. Reprint of 1823 edition. Garden City, N.Y.: Anchor Books.

Berman, Harold J. 1983. *Law and Revolution: The Formation of the Western Legal Tradition*. Cambridge, Mass: Harvard University Press.

Blumstein, Alfred. 1983. Incapacitation. In *Encyclopedia of Crime and Justice*, edited by Sanford H. Kadish, 3:873-80. New York: Free Press.

Blumstein, Alfred, Jacqueline Cohen, and Daniel S. Nagin, editors. 1978. *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates.* Washington, D.C.: National Academy of Sciences.

Bonczar, Thomas P., and Allen J. Beck. 1997. Lifetime Likelihood of Going to State or Federal Prison. Bureau of Justice Statistics Special Report, March, 1997, NCJ-160092. Washington D.C.: U.S. Department of Justice.

Carr-Hill, Roy A., and Nicholas H. Stern. 1979. *Crime, the Police, and Criminal Statistics: An Analysis of Official Statistics for England and Wales Using Econometric Methods*. London: Academic Press.

Chu, C. Y. Cyrus, Sheng-cheng Hu, and Ting-yuan Huang. 2000. Punishing Repeat Offenders More Severely. *International Review of Law and Economics* 20:127-40.

Cohen, Mark A. 1989. Corporate Crime and Punishment: A Study of Social Harm and Sentencing Practice in the Federal Courts, 1984-1987. *American Criminal Law Review* 26:605-60.

Cook, Philip J. 1977. Punishment and Crime: A Critique of Current Findings Concerning the Preventive Effects of Punishment. *Law and Contemporary Problems* 41:164-204.

Daly, Martin, and Margo Wilson. 1988. *Homicide.* New York: A. de Gruyter.

DiIulio, John J., Jr., and Anne Morrison Piehl. 1991. Does Prison Pay? *Brookings Review*, 9 (no. 4):28-35.

Ehrlich, Issac. 1996. Crime, Punishment, and the Market for Offenses. *Journal of Economic Perspectives* 10 (no. 1):43-67.

Eide, Erling. 2000. Economics of Criminal Behavior. In *Encyclopedia of Law and Economics*, edited by Boudewijn Bouckaert and Gerrit De Geest, 5:345-89. Cheltenham: Edward Elgar.

Fischel, Daniel R., and Alan O. Sykes. 1996. Corporate Crime. *Journal of Legal Studies* 25:319-49.

Forte, David F. 1983. Comparative Criminal Law and Enforcement: Islam. In *Encyclopedia of Crime and Justice*, edited by Sanford H. Kadish, 1:193-200. New York: Free Press.

Frank, Robert H. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton.

Friedman, David, and William Sjostrom. 1993. Hanged for a Sheep -- The Economics of Marginal Deterrence. *Journal of Legal Studies* 22:345-66.

Garoupa, Nuno. 1997. The Theory of Optimal Law Enforcement. *Journal of Economic Surveys* 11:267-95.

----- 1999. Optimal Law Enforcement with Dissemination of Information. *European Journal of Law and Economics* 7:183-96.

Glaeser, Edward L. 1998. Economic Approach to Crime and Punishment. In *The New Palgrave Dictionary of Economics and the Law*, edited by Peter Newman, 2:1-6. London: Macmillan.

Greenberg, David F. 1983. Age and Crime. In *Encyclopedia of Crime and Justice,* edited by Sanford H. Kadish, 1:30-35. New York: Free Press.

Hirshleifer, Jack. 1978. Natural Economy versus Political Economy. *Journal of Social and Biological Structures* 1:319-37.

----- 1987. On the Emotions as Guarantors of Threats and Promises. In *The Latest on the Best: Essays on Evolution and Optimality*, edited by John Dupré, 307-26. Cambridge, Mass.: MIT Press.

Hoffman, Jan. Crime and Punishment: Shame Gains Popularity. *New York Times*, January 16, 1997, p. A1.

Holmes, Oliver Wendell, Jr. [1881] 1963. *The Common Law*. Mark DeWolfe Howe, editor. Boston: Little Brown.

Innes, Robert. 1999. Remediation and Self-Reporting in Optimal Law Enforcement. *Journal of Public Economics* 72:379-93.

Jerry, Robert H. 1996. *Understanding Insurance Law.* Second edition. New York: Matthew Bender.

Kaplow, Louis. 1990. A Note on the Optimal Use of Nonmonetary Sanctions. *Journal of Public Economics* 42:245-47.

----- 1992. The Optimal Probability and Magnitude of Fines for Acts That Definitely Are Undesirable. *International Review of Law and Economics* 12:3-11.

Kaplow, Louis, and Steven Shavell. 1994b. Optimal Law Enforcement with Self-Reporting of Behavior. *Journal of Political Economy* 102:583-606.

-----2002b. *Fairness versus Welfare*. Cambridge, Mass.: Harvard University Press. (Also published in *Harvard Law Review* 114:961-1388.)

Keeton, Robert E. 1971. *Basic Text on Insurance Law.* St. Paul, Minn.: West Publishing Co.

Keeton, W. Page, Dan Dobbs, Robert Keeton, and David Owen. 1984. *Prosser and Keeton on the Law of Torts*. Fifth edition. St. Paul, Minn.: West Publishing Co.

Kessler, Daniel, and Steven D. Levitt. 1999. Using Sentence Enhancements to Distinguish Between Deterrence and Incapacitation. *Journal of Law and Economics* 42:343-63.

Khanna, Vikramaditya S. 1996. Corporate Criminal Liability: What Purpose Does It Serve? *Harvard Law Review* 109:1477-1534.

Klerman, Daniel. 2001. Settlement and the Decline of Private Prosecution in Thirteenth-Century England. *Law and History Review* 19:1-65.

Lab, Steven P., and John T. Whitehead. 1988. An Analysis of Juvenile Correctional Treatment. *Crime and Delinquency* 34:60-83.

LaFave, Wayne R. 2000. *Criminal Law.* Third edition. St. Paul, Minn.: West Group.

Levitt, Steven D. 1996. The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation. *Quarterly Journal of Economics* 111:319-51.

----- 1997. Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review* 87:270-90.

----- 1998a. Juvenile Crime and Punishment. *Journal of Political Economy* 106:1156-85.

----- 1998b. Why Do Increased Arrest Rates Appear to Reduce Crime: Deterrence, Incapacitation, or Measurement Error? *Economic Inquiry* 36:353-72.

Lott, John R., Jr. 2000. Corporate Criminal Liability. In *Encyclopedia of Law and Economics*, edited by Boudewijn Bouckaert and Gerrit De Geest, 5:492-501. Cheltenham: Edward Elgar.

Malik, Arun S. 1990. Avoidance, Screening, and Optimum Enforcement. *Rand Journal of Economics* 21:341-53.

----- 1993. Self-Reporting and the Design of Policies for Regulating Stochastic Pollution. *Journal of Environmental Economics and Management* 24:241-57.

McNeely, Mary C. 1941. Illegality as a Factor in Liability Insurance. *Columbia Law Review* 41:26-60.

Montesquieu, Charles-Louis de Secondat, baron. 1989. *The Spirit of the Laws*. Translated and edited by Anne M. Cohler, Basia Carolyn Miller, and Harold Samuel Stone. Cambridge: Cambridge University Press.

Mookherjee, Dilip. 1997. The Economics of Enforcement. In *Issues in Economic Theory and Public Policy: Essays in Honour of Professor Tapas Majumdar,* edited by Amitava Bose, Mihir Rakshit, and Anup Sinha, 202-49. Oxford: Oxford University Press.

Mookherjee, Dilip, and Ivan P. L. Png. 1992. Monitoring vis-à-vis Investigation in Enforcement of Law. *American Economic Review* 82:556-65.

----- 1994. Marginal Deterrence in Enforcement of Law. *Journal of Political Economy* 102:1039-66.

Packer, Herbert L. 1968. *The Limits of the Criminal Sanction*. Stanford: Stanford University Press.

Polinsky, A. Mitchell, and Daniel L. Rubinfeld. 1991. A Model of Optimal Fines for Repeat Offenders. *Journal of Public Economics* 46:291-306.

Polinsky, A. Mitchell, and Steven Shavell. 1979. The Optimal Tradeoff Between the Probability and Magnitude of Fines. *American Economic Review* 69:880-91.

----- 1984. The Optimal Use of Fines and Imprisonment. *Journal of Public Economics* 24:89-99.

----- 1992. Enforcement Costs and the Optimal Magnitude and Probability of Fines. *Journal of Law and Economics* 35:133-48.

----- 1994. Should Liability Be Based on the Harm to the Victim or the Gain to the Injurer? *Journal of Law, Economics, and Organization* 10:427-37.

----- 1998a. On Offense History and the Theory of Deterrence. *International Review of Law and Economics* 18:305-24.

----- 1999. On the Disutility and Discounting of Imprisonment and the Theory of Deterrence. *Journal of Legal Studies* 28:1-16.

----- 2000a. The Economic Theory of Public Enforcement of Law. *Journal of Economic Literature* 38:45-76.

----- 2000b. The Fairness of Sanctions: Some Implications for Optimal Enforcement Policy. *American Law and Economics Review* 2:223-37.

Pollock, Frederick, and Frederic William Maitland. 1911. *The History of English Law Before the Time of Edward I.* Second edition. Cambridge: Cambridge University Press.

Posner, Richard A. 1980. Retribution and Related Concepts of Punishment. *Journal of Legal Studies* 9:71-92.

----- 1985a. An Economic Theory of the Criminal Law. *Columbia Law Review* 85:1193-1231.

Rubinstein, Ariel. 1979. An Optimal Conviction Policy for Offenses That May Have Been Committed by Accident. In *Applied Game Theory*, edited by Stephen J. Brams, Andrew Schotter, and G. Schwodiauer, 406-13. Wurzburg: Physica-Verlag.

Sah, Raaj K. 1991. Social Osmosis and Patterns of Crime. *Journal of Political Economy* 99:1272-95.

Schwartz, Richard D. 1983. Rehabilitation. In *Encyclopedia of Crime and Justice*, edited by Sanford H. Kadish, 4:1364-74. New York: Free Press.

Shavell, Steven. 1985a. Criminal Law and the Optimal Use of Nonmonetary Sanctions as a

Deterrent. *Columbia Law Review* 85:1232-62.

----- 1985b. Uncertainty Over Causation and the Determination of Civil Liability. *Journal of Law and Economics* 28:587-609.

----- 1987a. *Economic Analysis of Accident Law*. Cambridge, Mass.: Harvard University Press.

----- 1987b. A Model of Optimal Incapacitation. *American Economic Review: Papers and Proceedings* 77:107-10.

----- 1987c. The Optimal Use of Nonmonetary Sanctions as a Deterrent. *American Economic Review* 77:584-92.

----- 1990. Deterrence and the Punishment of Attempts. *Journal of Legal Studies* 19:435-66.

----- 1991b. Specific versus General Enforcement of Law. *Journal of Political Economy* 99:1088-1108.

----- 1992. A Note on Marginal Deterrence. *International Review of Law and Economics* 12:345-55.

Spelman, William. 1994. *Criminal Incapacitation*. New York: Plenum Press.

----- 2000. The Limited Importance of Prison Expansion. In *The Crime Drop in America*, edited by Alfred Blumstein and Joel Wallman, 97-129. Cambridge: Cambridge University Press.

Statistical Abstract of the United States 2001. 2001. Economics and Statistics Administration, U.S. Census Bureau. Washington, D.C.: U.S. Department of Commerce.

Stigler, George J. 1970. The Optimum Enforcement of Laws. *Journal of Political Economy* 78:526-36.

Trivers, Robert L. 1971. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology* 46:35-57.

Tversky, Amos, and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185:1124-31.

U. S. Department of Justice. 1988. *Report to the Nation on Crime and Justice, Technical Appendix.* Bureau of Justice Statistics, second edition. NCJ-112011. Washington, D.C.: U.S. Department of Justice.

----- 2001b. *Sourcebook of Criminal Justice Statistics 2000.* Bureau of Justice Statistics, NCJ-190251. Washington D.C.: U.S. Department of Justice.

U.S. Sentencing Commission. 1995. *Federal Sentencing Guidelines Manual*. 1995 edition. St. Paul, Minn.: West.

Viscusi, W. Kip. 1986b. The Risks and Rewards of Criminal Activity: A Comprehensive Test of Criminal Deterrence. *Journal of Labor Economics* 4:317-40.

Wilde, Louis L. 1992. Criminal Choice, Nonmonetary Sanctions, and Marginal Deterrence: A Normative Analysis. *International Review of Law and Economics* 12:333-44.

Wilson, James Q., and Richard J. Herrnstein. 1985. *Crime and Human Nature*. New York: Simon & Schuster.

Witte, Ann Dryden. 1980. Estimating the Economic Model of Crime with Individual Data. *Quarterly Journal of Economics* 94:57-84.

Wright, Richard A. 1994. *In Defense of Prisons.* Contributions in Criminology and Penology, number 43. Westport, Conn.: Greenwood Press.