Nobel Lecture: The Economic Way of Looking at Behavior

Gary S. Becker

University of Chicago and Hoover Institution

An important step in extending the traditional theory of individual rational choice to analyze social issues beyond those usually considered by economists is to incorporate into the theory a much richer class of attitudes, preferences, and calculations. While this approach to behavior builds on an expanded theory of individual choice, it is not mainly concerned with individuals. It uses theory at the micro level as a powerful tool to derive implications at the group or macro level. The lecture describes the approach and illustrates it with examples drawn from my past and current work.

I. The Economic Approach

My research uses the economic approach to analyze social issues that range beyond those usually considered by economists. This lecture will describe the approach and illustrate it with examples drawn from past and current work.

Unlike Marxian analysis, the economic approach I refer to does not assume that individuals are motivated solely by selfishness or material gain. It is a *method* of analysis, not an assumption about particular motivations. Along with others, I have tried to pry economists away from narrow assumptions about self-interest. Behavior is driven by a much richer set of values and preferences.

This is a slightly revised version of my Nobel Lecture, delivered December 9, 1992, in Stockholm, Sweden. It is dedicated to the memory of George J. Stigler, who died almost exactly 1 year before the lecture was delivered. Nobel laureate, outstanding economist, very close friend, and mentor, he would have been as happy as I was had he lived to see me deliver the 1992 Nobel Lecture in Economic Sciences. I have had valuable comments from James Coleman, Richard Posner, Sherwin Rosen, Raaj Sah, José Scheinkman, Richard Stern, and Stephen Stigler.

[*Journal of Political Economy*, 1993, vol. 101, no. 3]

The analysis assumes that individuals maximize welfare *as they conceive it,* whether they be selfish, altruistic, loyal, spiteful, or masochistic. Their behavior is forward-looking, and it is also assumed to be consistent over time. In particular, they try as best they can to anticipate the uncertain consequences of their actions. Forward-looking behavior, however, may still be rooted in the past, for the past can exert a long shadow on attitudes and values.

Actions are constrained by income, time, imperfect memory and calculating capacities, and other limited resources, and also by the opportunities available in the economy and elsewhere. These opportunities are largely determined by the private and collective actions of other individuals and organizations.

Different constraints are decisive for different situations, but the most fundamental constraint is limited time. Economic and medical progress have greatly increased length of life, but not the physical flow of time itself, which always restricts everyone to 24 hours per day. So while goods and services have expanded enormously in rich countries, the total time available to consume has not.

Thus wants remain unsatisfied in rich countries as well as in poor ones. For while the growing abundance of goods may reduce the value of additional goods, time becomes more valuable as goods become more abundant. The welfare of people cannot be improved in a utopia in which everyone's needs are fully satisfied, but the constant flow of time makes such a utopia impossible. These are some of the issues analyzed in the literature on time allocation (for two early studies, see Becker [1965] and Linder [1970]).

The following sections illustrate the economic approach with four very different subjects. To understand discrimination against minorities, it is necessary to widen preferences to accommodate prejudice and hatred of particular groups. The economic analysis of crime incorporates into rational behavior illegal and other antisocial actions. The human capital perspective considers how the productivity of people in market and nonmarket situations is changed by investments in education, skills, and knowledge. The economic approach to the family interprets marriage, divorce, fertility, and relations among family members through the lens of utility-maximizing, forward-looking behavior.

II. Discrimination against Minorities

Discrimination against outsiders has always existed, but with the exception of a few discussions of the employment of women (see Fawcett 1918; Edgeworth 1922), economists wrote little on this subject before the 1950s. I began to worry about racial, religious, and gender

discrimination while a graduate student, and I used the concept of discrimination coefficients to organize an approach to prejudice and hostility to members of particular groups.

Instead of making the common assumptions that employers consider only the productivity of employees, that workers ignore the characteristics of those with whom they work, and that customers care only about the qualities of the goods and services provided, discrimination coefficients incorporate the influence of race, gender, and other personal characteristics on tastes and attitudes. Employees may refuse to work under a woman or a black even when they are well paid to do so, or a customer may prefer not to deal with a black car salesman. It is only through *widening* of the usual assumptions that it is possible to begin to understand the obstacles to advancement encountered by minorities.

Presumably, the amount of observable discrimination against minorities in wages and employment depends not only on tastes for discrimination but also on other variables, such as the degree of competition and civil rights legislation. In the 1950s, a systematic analysis of how prejudice and other variables interact could begin with the important theory of compensating differentials originated by Adam Smith, and Gunnar Myrdal's pioneering *American Dilemma* (1944), but much remained to be done. I spent several years working out a theory of how actual discrimination in earnings and employment is determined by tastes for discrimination, along with the degree of competition in labor and product markets, the distribution of discrimination coefficients among members of the majority group, the access of minorities to education and training, the outcome of median voter and other voting mechanisms that determine whether legislation favors or is hostile to minorities, and other considerations. My advisors encouraged me to convert my doctoral dissertation into a book (Becker 1957). I have continued over my career to write books rather than only articles, a practice that has become uncommon in economics.

Actual discrimination in the marketplace against a minority group depends on the combined discrimination of employers, workers, consumers, schools, and governments. The analysis shows that sometimes the environment greatly softens, while at other times it magnifies, the impact of a given amount of prejudice. For example, the discrepancy in wages between equally productive blacks and whites, or women and men, would be much smaller than the degree of prejudice against blacks and women when many companies can efficiently specialize in employing mainly blacks or women.

Indeed, in a world with constant returns to scale in production, two segregated economies with the same distribution of skills would completely bypass discrimination, and they would have equal wages

and equal returns to other resources, regardless of the desire to discriminate against the segregated minorities. Therefore, discrimination by the majority in the marketplace is effective because minority members cannot provide various skills in sufficient quantities to companies that would specialize in using these workers.

When the majority is very large compared to the minority—in the United States whites are nine times as numerous as and have much more human and physical capital per capita than blacks—market discrimination by the majority hardly lowers its incomes, but may greatly reduce the incomes of the minority. However, when minority members are a sizable fraction of the total, discrimination by members of the majority injures them as well.

This proposition can be illustrated with an analysis of discrimination in South Africa, where blacks are some five times as numerous as whites. Discrimination against blacks has also significantly hurt whites, although some white groups have benefited (see Becker [1957] 1971, pp. 30–31; Hutt 1964; Lundahl 1992). Its sizable cost to whites helps explain why apartheid and other blatant forms of Afrikaner discrimination were never fully effective and eventually broke down.

Many economists have the impression that my analysis of prejudice implies that market discrimination disappears in the "long run" (Arrow [1972] seems to be the first to make this claim). This impression is erroneous because I had shown that whether employers who do not want to discriminate compete away all discriminating employers depends not only on the distribution of tastes for discrimination among potential employers, but critically also on the nature of firm production functions (see Becker [1957] 1971, pp. 43–45).

Of greater significance empirically is the long-run discrimination by employees and customers, who are far more important sources of market discrimination than employers. There is no reason to expect discrimination by these groups to be competed away unless it is possible to have enough efficient segregated firms and effectively segregated markets for goods (see Cain's [1986] good review of this and other issues regarding discrimination).

A novel theoretical development in recent years is the analysis of the consequences of stereotyped reasoning or statistical discrimination (see Phelps 1972; Arrow 1973). This analysis suggests that the *beliefs* of employers, teachers, and other influential groups that minority members are less productive *can* be self-fulfilling, for these beliefs may cause minorities to underinvest in education, training, and work skills, such as punctuality. The underinvestment does make them less productive (see a good recent analysis by Loury [1992]).

Evidence from many countries on the earnings, unemployment, and occupations of blacks, women, religious groups, immigrants, and

others has expanded enormously during the past 25 years. This evidence more fully documents the economic position of minorities and how that changes in different environments.

The economic theory of discrimination based on prejudice implies that actual discrimination by firms or workers is measured by how much profits or wages they forfeit to avoid hiring or working with members of a group that is disliked. Discrimination by consumers is measured by the higher prices they pay to avoid products or services produced by those members. Evidence on forgone profits, wages, or prices is typically not available, so discrimination against a group is usually measured by comparing the earnings of members of the group with earnings of the "majority" who have the same years of schooling, job experience, and other measurable characteristics. Since this indirect approach has obvious defects, these studies have not dispelled some of the controversies over the source of lower incomes of minorities.

Recent studies on whether banks discriminate in their mortgage lending against blacks and other minorities compare the likelihood of getting a loan for minority and white applicants who are similar in incomes, credit backgrounds, and other available characteristics. The conclusion typically has been that blacks but not Asian-Americans are rejected excessively compared to whites of similar characteristics.

Unfortunately, these studies do not use the correct procedure for assessing whether banks discriminate, which is to determine whether loans are more profitable to blacks (and other minorities) than to whites. This requires examining the default and other payback experiences of loans, the interest rates charged, and so forth. If banks discriminate against minority applicants, they should earn *greater* profits on the loans actually made to them than on those to whites. The reason is that discriminating banks would be willing to accept marginally profitable white applicants who would be turned down if they were black.

III. Crime and Punishment

I began to think about crime in the 1960s after driving to Columbia University for an oral examination of a student in economic theory. I was late and had to decide quickly whether to put the car in a parking lot or risk getting a ticket for parking illegally on the street. I calculated the likelihood of getting a ticket, the size of the penalty, and the cost of putting the car in a lot. I decided it paid to take the risk and park on the street. (I did not get a ticket.)

As I walked the few blocks to the examination room, it occurred

to me that the city authorities had probably gone through a similar analysis. The frequency of their inspection of parked vehicles and the size of the penalty imposed on violators should depend on their estimates of the type of calculations potential violators like me would make. Of course, the first question I put to the hapless student was to work out the optimal behavior of both the offenders and the police, something I had not yet done.

In the 1950s and 1960s, intellectual discussions of crime were dominated by the opinion that criminal behavior was caused by mental illness and social oppression, and that criminals were helpless "victims." A book by a well-known psychiatrist was entitled *The Crime of Punishment* (see Menninger 1966). Such attitudes began to exert a major influence on social policy, as laws changed to expand criminals' rights. These changes reduced the apprehension and conviction of criminals and provided less protection to the law-abiding population.

I was not sympathetic to the assumption that criminals had radically different motivations from everyone else. I explored instead the theoretical and empirical implications of the assumption that criminal behavior is rational (see the early pioneering work by Bentham [1931] and Beccaria [(1797) 1986]), but again "rationality" did not imply narrow materialism. It recognized that many people were constrained by moral and ethical considerations, and they did not commit crimes even when these were profitable and there was no danger of detection. However, police and jails would be unnecessary if such attitudes always prevailed. Rationality implied that some individuals become criminals because of the financial and other rewards from crime compared to legal work, taking account of the likelihood of apprehension and conviction, and the severity of punishment.

The amount of crime is determined not only by the rationality and preferences of would-be criminals but also by the economic and social environment created by public policies, including expenditures on police, punishments for different crimes, and opportunities for employment, schooling, and training programs. Clearly, the types of legal jobs available as well as law, order, and punishment are an integral part of the economic approach to crime.

Total public spending on fighting crime can be reduced, while keeping the mathematically expected punishment unchanged, by offsetting a cut in expenditures on catching criminals with a sufficient increase in the punishment to those convicted. However, risk-preferring individuals are more deterred from crime by a higher probability of conviction than by severe punishments. Therefore, optimal behavior by the state would balance the reduced spending on police and courts from lowering the probability of conviction against

the preference of risk-preferring criminals for a lesser certainty of punishment. The state should also consider the likelihood of punishing innocent persons.

In the early stages of my work on crime, I was puzzled by why theft is socially harmful since it appears merely to redistribute resources, usually from wealthier to poorer individuals. I resolved the puzzle (Becker 1968, p. 171, n. 3) by pointing out that criminals spend on weapons and on the value of the time in planning and carrying out their crimes, and that such spending is socially unproductive—it is what is now called "rent seeking"—because it does not create wealth, only forcibly redistributes it. I approximated the social cost of theft by the dollars stolen since rational criminals would be willing to spend up to that amount on their crimes. I should have added the resources spent by potential victims protecting themselves against crime.

One reason why the economic approach to crime became so influential is that the same analytic apparatus can be used to study enforcement of all laws, including minimum wage legislation, clean air acts, insider trader and other violations of security laws, and income tax evasions. Since few laws are self-enforcing, they require expenditures on conviction and punishment to deter violators. The U.S. Sentencing Commission (1992) has explicitly used the economic analysis of crime to develop rules to be followed by judges in punishing violators of federal statutes.

Studies of crime that use the economic approach have become common during the past quarter century. These include analysis of the optimal marginal punishments to deter increases in the severity of crimes—for example, to deter a kidnapper from killing his victim (the modern literature starts with Stigler [1970])—and the relation between private and public enforcement of laws (see Becker and Stigler 1974; Landes and Posner 1975).

Fines are preferable to imprisonment and other types of punishment because they can deter crimes effectively if criminals have sufficient financial resources—if they are not "judgment proof," to use legal jargon. Moreover, fines are more efficient than other methods because the cost to offenders is also revenue to the state. My discussion of the relations between fines and other punishments has been clarified and considerably improved (see, e.g., Polinsky and Shavell 1984; Posner 1986).

Empirical assessments of the effects on crime rates of prison terms, conviction rates, unemployment levels, income inequality, and other variables have become more numerous and more accurate (the pioneering work is by Ehrlich [1973], and the subsequent literature is extensive). The greatest controversies surround the question of

whether capital punishment deters murders, a controversy that arouses much emotion but is far from being resolved (see, e.g., Ehrlich 1975; National Research Council 1978).

IV. Human Capital

Until the 1950s economists generally assumed that labor power was given and not augmentable. The sophisticated analyses of investments in education and other training by Adam Smith, Alfred Marshall, and Milton Friedman were not integrated into discussions of productivity. Then Theodore W. Schultz and others began to pioneer the exploration of the implications of human capital investments for economic growth and related economic questions.

Human capital analysis starts with the assumption that individuals decide on their education, training, medical care, and other additions to knowledge and health by weighing the benefits and costs. Benefits include cultural and other nonmonetary gains along with improvement in earnings and occupations, whereas costs usually depend mainly on the forgone value of the time spent on these investments. The concept of human capital also covers accumulated work and other habits, even including harmful addictions such as smoking and drug use. Human capital in the form of good work habits or addictions to heavy drinking has major positive or negative effects on productivity in both market and nonmarket sectors.

The various kinds of behavior included under the rubric of human capital help explain why the concept is so powerful and useful. It also means that the process of investing or disinvesting in human capital often alters the very nature of a person: training may change a lifestyle from one with perennial unemployment to one with stable and good earnings, or accumulated drinking may destroy a career, health, and even the capacity to think straight.

Human capital is so uncontroversial nowadays that it may be difficult to appreciate the hostility in the 1950s and 1960s toward the approach that went with the term. The very concept of *human* capital was alleged to be demeaning because it treated people as machines. To approach schooling as an investment rather than a cultural experience was considered unfeeling and extremely narrow. As a result, I hesitated a long time before deciding to call my book *Human Capital* (1964) and hedged the risk by using a long subtitle that I no longer remember. Only gradually did economists, let alone others, accept the concept of human capital as a valuable tool in the analysis of various economic and social issues.

My work on human capital began with an effort to calculate both private and social rates of return to men, women, blacks, and other groups from investments in different levels of education. After a

while it became clear that the analysis of human capital can help explain many regularities in labor markets and the economy at large. It seemed possible to develop a more general theory of human capital that includes firms as well as individuals and that could consider its macroeconomic implications.

The empirical analysis tried to correct data on the higher earnings of more educated persons for the fact that they are abler: they have higher IQs and score better on other aptitude tests. It also considered the effects on rates of return to education of mortality, income taxes, forgone earnings, and economic growth. Ability corrections did not seem very important, but large changes in adult mortality and sizable rates of economic growth did have big effects. Meltzer (1992) recently has argued that the high death rates, especially from AIDS, of young males in many parts of Africa greatly discourage investments in human capital there.

The empirical study of investments in human capital received a major boost from Mincer's (1974) classic work. He extended a simple regression analysis that related earnings to years of schooling (Becker and Chiswick 1966) to include a crude but very useful measure of on-the-job training and experience: years after finishing school; he used numerous individual observations rather than grouped data, and he carefully analyzed the properties of residuals from earnings-generating equations. There are now numerous estimated rates of return to education and training for many countries (for a summary of some of this literature, see Psacharopoulos [1985]); indeed, the earnings equation is probably the most common empirical regression in microeconomics.

The accumulating evidence on the economic benefits of schooling and training also promoted the importance of human capital in policy discussions. This new faith in human capital has reshaped the way governments approach the problem of stimulating growth and productivity, as was shown by the emphasis on human capital in the recent presidential election in the United States.

One of the most influential theoretical concepts in human capital analysis is the distinction between general and specific training or knowledge (see Becker 1962; Oi 1962). By definition, firm-specific knowledge is useful only in the firms providing it, whereas general knowledge is useful also in other firms. Teaching someone to operate an IBM-compatible personal computer is general training, whereas learning the authority structure and the talents of employees in a particular company is specific knowledge. This distinction helps explain why workers with highly specific skills are less likely to quit their jobs and are the last to be laid off during business downturns. It also explains why most promotions are made from within a firm rather than through hiring—workers need time to learn about a firm's struc-

ture and "culture"—and why better accounting methods would include the specific human capital of employees among the principal asset of most companies.

Firm-specific investments produce rents that must be shared between employers and employees, a sharing process that is vulnerable to "opportunistic" behavior because each side may try to extract most of the rent after investments are in place. Rents and opportunism due to specific investments play a crucial role in the modern economic theory of how organizations function (see Williamson 1985) and in many discussions of principal-agent problems (see, e.g., Grossman and Hart 1983). The implications of specific capital for sharing and turnover have also been used in analyzing marriage "markets" to explain divorce rates and bargaining within a marriage (see Becker, Landes, and Michael 1977; McElroy and Horney 1981) and in analyzing political "markets" to explain the low turnover of politicians (see Cain, Ferejohn, and Fiorina 1987).

The theory of human capital investment relates inequality in earnings to differences in talents, family background, and bequests and other assets (see Becker and Tomes 1986). Many empirical studies of inequality also rely on human capital concepts, especially differences in schooling and training (see Mincer 1974). The sizable growth in earnings inequality in the United States during the 1980s that has excited so much political discussion is largely explained by higher returns to the more educated and better trained (see, e.g., Murphy and Welch 1992).

Human capital theory gives a provocative interpretation of the so-called gender gap in earnings. Traditionally, women have been far more likely than men to work part-time and intermittently partly because they usually withdrew from the labor force for a while after having children. As a result, they had fewer incentives to invest in education and training that improved earnings and job skills.

During the past 25 years all this changed. The decline in family size, the growth in divorce rates, the rapid expansion of the service sector (where most women are employed), the continuing economic development that raised the earnings of women along with those of men, and civil rights legislation encouraged greater labor force participation by women and, hence, greater investment in market-oriented skills. In practically all rich countries, these forces significantly improved both the occupations and relative earnings of women.

The United States' experience is especially well documented. The gender gap in earnings among full-time men and women remained at about 35 percent from the midfifties to the midseventies. Then women began a steady economic advance, which is still continuing;

it narrowed the gap to under 25 percent (see, e.g., O'Neill 1985; Goldin 1990). Women are flocking to business, law, and medical schools, and they are working at skilled jobs that they formerly shunned or were excluded from.

Schultz and others (see, e.g., Schultz 1963; Denison 1962) early on emphasized that investments in human capital are a major contributor to economic growth. But after a while the relation of human capital to growth was neglected, as economists became discouraged about whether the available growth theory gave many insights into the progress of different countries. The revival of more formal models of endogenous growth has brought human capital once again to the forefront of the discussions (see, e.g., Romer 1986; Lucas 1988; Becker, Murphy, and Tamura 1990; Barro and Sala-i-Martin 1992).

V. Formation, Dissolution, and Structure of Families

The rational choice analysis of family behavior builds on maximizing behavior, investments in human capital, the allocation of time, and discrimination against women and other groups. The rest of the lecture focuses on this analysis since it is still quite controversial, and I can discuss some of my current research.

Writing *A Treatise on the Family* (1981) is the most difficult sustained intellectual effort I have undertaken. The family is arguably the most fundamental and oldest of institutions: some authors trace its origin to more than 40,000 years ago (Soffer 1990). The *Treatise* tries to analyze not only modern Western families but those in other cultures and changes in family structure during the past several centuries.

Trying to cover this broad subject required a degree of mental commitment over more than 6 years, during many nighttime as well as daytime hours, that left me intellectually and emotionally exhausted. In his autobiography, Bertrand Russell says that writing the *Principia Mathematica* used up so much of his mental powers that he was never again fit for really hard intellectual work. It took about 2 years after finishing the *Treatise* to regain my intellectual zest.

The analysis of fertility has a long and honorable history in economics, but until recent years marriage and divorce, and the relations between husbands, wives, parents, and children, had been largely neglected by economists (although see the important study by Mincer [1962]). The point of departure of my work on the family is the assumption that when men and women decide to marry, or have children, or divorce, they attempt to raise their welfare by comparing benefits and costs. So they marry when they expect to be better off

than if they remained single, and they divorce if that is expected to increase their welfare.

People who are not intellectuals are often surprised when told that this approach is controversial since it seems obvious to them that individuals try to improve their welfare by marriage and divorce. The rational choice approach to marriage and other behavior is in fact often consistent with the instinctive economics "of the common person" (Farrell and Mandel 1992).

Still, intuitive assumptions about behavior are only the *starting point* of systematic analysis, for alone they do not yield many interesting implications. Marquis of Deffand said, when commenting on the story that St. Denis walked two leagues while carrying his head in his hands, that "the distance is nothing; it is only the first step that is difficult." The first one in new research is also important, but it is of little value without second, third, and several additional steps (I owe this reference to the marquis and the comparison with research to Richard Posner). The rational choice approach takes further steps by using a framework that combines maximizing behavior with the analysis of marriage and divorce markets, specialization and the division of labor, old-age support, investments in children, and legislation that affects families. The implications of the full model are often not so obvious and sometimes run sharply counter to received opinion.

For example, contrary to a common belief about divorce among the rich, the economic analysis of family decisions shows that wealthier couples are *less* likely to divorce than poorer couples. According to this theory, richer couples tend to gain a lot from remaining married, whereas many poorer couples do not. A poor woman may well doubt whether it is worth staying married to someone who is chronically unemployed. Empirical studies for many countries do indicate that marriages of richer couples are much more stable (see, e.g., Becker, Landes, and Michael 1977; Hernandez 1992).

Efficient bargaining between husbands and wives implies that the trend in Europe and the United States toward no-fault divorce during the past two decades did not raise divorce rates and, therefore, contrary to many claims, that it could not be responsible for the rapid rise in these rates. However, the theory does indicate that no-fault divorce hurts women with children whose marriages are broken up by their husbands. Feminists initially supported no-fault divorce, but some now have second thoughts about whether it has favorable effects on divorced women.

Economic models of behavior have been used to study fertility ever since Thomas Malthus's classic essay; the great Swedish economist, Knut Wicksell, was attracted to economics by his belief in the Malthusian predictions of overpopulation. But Malthus's conclusion that fer-

tility would rise and fall as incomes increased and decreased was contradicted by the large decline in birth rates after some countries became industrialized during the latter part of the nineteenth century and the early part of this century.

The failure of Malthus's simple model of fertility persuaded economists that family size decisions lay beyond economic calculus. The neoclassical growth model reflects this belief, for in most versions it takes population growth as exogenous and given (see, e.g., Cass 1965; Arrow and Kurz 1970).

However, the trouble with the Malthusian approach is not its use of economics per se, but an economics inappropriate for modern life. It neglects that the time spent on child care becomes more expensive when countries are more productive. The higher value of time raises the cost of children and thereby reduces the demand for large families. It also fails to consider that the greater importance of education and training in industrialized economies encourages parents to invest more in the skills of their children, which also raises the cost of large families. The growing value of time and the increased emphasis on schooling and other human capital explain the decline in fertility as countries develop, and many other features of birth rates in modern economies.

In almost all societies, married women have specialized in bearing and rearing children and in certain agricultural activities, whereas married men have done most of the fighting and market work. It should not be controversial to recognize that the explanation is a combination of biological differences between men and women—especially differences in their innate capacities to bear and rear children—and legal and other discrimination against women in market activities, partly through cultural conditioning. However, large and highly emotional differences of opinion exist over the relative importance of biology and discrimination in generating the traditional division of labor in marriages.

Contrary to allegations in many attacks on the economic approach to the gender division of labor (see, e.g., Boserup 1987), this analysis does not try to weight the relative importance of biology and discrimination. Its main contribution is to show how sensitive the division of labor is to *small* differences in either. Since the return from investing in a skill is greater when more time is spent utilizing the skill, a married couple could gain much from a sharp division of labor because the husband would specialize in some types of human capital and the wife in others. Given such a large gain from specialization within a marriage, only a *little* discrimination against women or *small* biological differences in child-rearing skills would cause the division of labor between household and market tasks to be strongly and sys-

tematically related to gender. The sensitivity to small differences explains why the empirical evidence cannot readily choose between biological and "cultural" interpretations. This theory also explains why many women entered the labor force as families became smaller, divorce became more common, and earning opportunities for women improved.

Relations among family members differ radically from those among employees of firms and members of other organizations. The interactions among husbands, wives, parents, and children are more likely to be motivated by love, obligation, guilt, and a sense of duty than by self-interest narrowly interpreted.

It was demonstrated about 20 years ago that altruism within families enormously alters how they respond to shocks and public policies that redistribute resources among members. It was shown that exogenous redistributions of resources from an altruist to her beneficiaries (or vice versa) may not affect the welfare of anyone because the altruist would try to reduce her gifts by the amount redistributed (Becker 1974). Barro (1974) derived this result in an intergenerational context, which cast doubt on the common assumption that government deficits and related fiscal policies have real effects on the economy.

The "Rotten Kid Theorem"—the name is very popular even when critics disagree with the analysis—carries the discussion of altruism further, for it shows how the behavior of selfish individuals is affected by altruism. Under some conditions, even selfish persons (of course, most parents believe that the best example of selfish beneficiaries and altruistic benefactors is selfish children with altruistic parents) are induced to act *as though* they are altruistic toward their benefactors because that raises their own selfish welfare. They act this way because otherwise gifts from their benefactors would be reduced enough to make them worse off (see Becker [1974] and the elaboration and qualifications to the analysis in Lindbeck and Weibull [1988], Bergstrom [1989], and Becker [1991, pp. 9–13]).

The Bible, Plato's *Republic,* and other early writings discussed the treatment of young children by their parents and of elderly parents by adult children. Both the elderly and children need care: in one case because of declining health and energy, and in the other because of biological growth and dependency. A powerful implication of the economic analysis of relations within families is that these two issues are closely related.

Parents who leave sizable bequests do not need old-age support because instead they help out their children. I mentioned earlier one well-known implication of this: under certain conditions, budget deficits and social security payments to the elderly have no real effects

because parents simply offset the bigger taxes in the future on their children through larger bequests.

It is much less appreciated that altruistic parents who leave bequests also tend to invest more in their children's skills, habits, and values. For they gain from financing all investments in the education and skills of children that yield a higher rate of return than the return on savings. They can indirectly save for old age by investing in children, and then reducing bequests when elderly. Both parents and children would be better off when parents make all investments in children that yield a higher return than that on savings, and then adjust bequests to the efficient level of investment (see sec. *A* of the Appendix for a formal demonstration).

However, even in rich countries, many parents do not plan on leaving bequests. These parents want old-age support, and they "underinvest" in their children's education and other care. They underinvest because they cannot compensate themselves for greater spending on children by reducing bequests since they do not plan on leaving any.

Both the children and parents would be better off if the parents agreed to invest more in the children in return for a commitment by the children to care for them when they need help. But how can such a commitment be enforced? Economists and lawyers usually recommend a written contract to ensure commitment, but can you imagine a society that will enforce contracts between adults and 10-year-olds or teenagers?

Part of my current research considers an indirect way to generate commitments when promises and written agreements are not binding. I shall describe briefly some of this new work because it carries the economic approach to the family onto uncharted ground related to the rational formation of preferences within families.

Parental attitudes and behavior have an enormous influence on their children. Parents who are alcoholic or are addicted to crack create a bizarre atmosphere for impressionable youngsters, whereas parents with stable values who transmit knowledge and inspire their children favorably influence both what their children are capable of and what they want to do. The economic approach can contribute insights into the formation of preferences through childhood experiences without necessarily adopting the Freudian emphasis on the primacy of what happened during the first few months of life.

Again, I am trying to model a commonsense idea, namely, that the attitudes and values of adults are enormously influenced by their childhood experiences. An Indian doctor living in the United States may love curry because he acquired a strong taste for it while growing

up in India, or a woman may forever fear men because she was sexually abused as a child.

Through its assumption of forward-looking behavior, the economic point of view implies that parents try to anticipate the effect of what happens to children on their attitudes and behavior when adults. These effects help determine the kind of care parents provide. For example, parents worried about old-age support may try to instill in their children feelings of guilt, obligation, duty, and filial love that indirectly, but still very effectively, can "commit" children to helping them out.

Economists have too narrow a perspective on commitments. "Manipulating" the experiences of others to influence their preferences may appear to be inefficient and fraught with uncertainty, but it can be the most effective way available to obtain commitment. Economic theory, especially game theory, needs to incorporate guilt, affection, and related attitudes into preferences in order to have a deeper understanding of when commitments are "credible" (see sec. *B* of the Appendix for a formal discussion).

Parents who do not leave bequests may be willing to make their children feel guiltier precisely because they gain more utility from greater old-age consumption than they lose from an equal reduction in children's consumption. This type of behavior may be considerably more common than suggested by the number of families that actually do leave bequests, for parents with young children often do not know whether they will be financially secure when they are old. They may try to protect themselves against ill health, unemployment, and other hazards of old age by instilling in their children a willingness to help out if that becomes necessary.

This analysis of the link between childhood experiences and adult preferences is closely related to work on rational habit formation (see Becker and Murphy [1988]; also see the discussion by Kandel and Lazear [1992] of the creation of guilt among employees). The formation of preferences is rational in the sense that parental spending on children partly depends on the anticipated effects of childhood experiences on adult attitudes and behavior. I do not have time to consider the behavior of children—such as crying and acting "cute"—that tries in turn to influence the attitudes of parents.

Many economists, including me, have excessively relied on altruism to tie together the interests of family members. Recognition of the connection between childhood experiences and future behavior reduces the need to rely on altruism in families. But it does not return the analysis to a narrow focus on self-interest, for it partially replaces altruism by feelings of obligation, anger, and other attitudes usually neglected by models of rational behavior.

If children are expected to help out in old age—perhaps because of guilt or related motivations—even parents who are not very loving would invest more in the children's human capital and save less to provide for their old age. (For a proof, see sec. *C* of the Appendix.) But equation (A12) of the Appendix shows that altruistic parents always prefer small increases in their own consumption when old to equal increases in their children's *if* they have made their children feel guilty. This means that such parents always underinvest in the children's human capital. This shows directly why creating guilt has costs and is not fully efficient.

Altruistic family heads who do not plan to leave bequests try to create a "warm" atmosphere in their families, so that members are willing to come to the assistance of those experiencing financial and other difficulties. This conclusion is relevant to discussions of so-called family values, a subject that received attention during the recent presidential campaign in the United States. Parents help determine the values of children—including their feelings of obligation, duty, and love—but what parents try to do can be greatly affected by public policies and changes in economic and social conditions.

Consider, for example, a program that transfers resources to the elderly, perhaps especially to poorer families who do not leave bequests, that reduces the elderly's dependence on children. According to the earlier analysis I gave, parents who do not need support when they become old do not try as hard to make children more loyal or guiltier or otherwise feel as well disposed toward their parents. This means that programs such as social security that significantly help the elderly would encourage family members to drift apart emotionally, not by accident but as maximizing responses to those policies.

Other changes in the modern world that have altered family values include increased geographical mobility, the greater wealth that comes with economic growth, better capital and insurance markets, higher divorce rates, smaller families, and publicly funded health care. These developments have generally made people better off, but they have also weakened the personal relations within families between husbands and wives, parents and children, and among more distant relatives, partly by reducing the incentives to invest in *creating* closer relations.

VI. Concluding Comments

An important step in extending the traditional analysis of individual rational choice is to incorporate into the theory a much richer class of attitudes, preferences, and calculations. This step is prominent in all the examples I consider. The analysis of discrimination includes in

preferences a dislike of—prejudice against—members of particular groups, such as blacks or women. In deciding whether to engage in illegal activities, potential criminals are assumed to act as though they consider both the gains and the risks, including the likelihood they will be caught and severity of punishments. In human capital theory, people rationally evaluate the benefits and costs of activities, such as education, training, expenditures on health, migration, and formation of habits that radically alter the way they are. The economic approach to the family assumes that even intimate decisions such as marriage, divorce, and family size are reached through weighing the advantages and disadvantages of alternative actions. The weights are determined by preferences that critically depend on the altruism and feelings of duty and obligation toward family members.

Since the economic, or rational choice, approach to behavior builds on a theory of individual decisions, criticisms of this theory usually concentrate on particular assumptions about how these decisions are made. Among other things, critics deny that individuals act consistently over time, and question whether behavior is forward-looking, particularly in situations that differ significantly from those usually considered by economists—such as those involving criminal, addictive, family, or political behavior. This is not the place to go into a detailed response to the criticisms, so I simply assert that no approach of comparable generality has yet been developed that offers serious competition to rational choice theory.

I have intentionally chosen certain topics for my research—such as addiction—to probe the boundaries of rational choice theory. William Blake said that you never know what is enough until you see what is more than enough (Jon Elster brought this proverb to my attention). My work may have sometimes assumed too much rationality, but I believe it has been an antidote to the extensive research that does not credit people with enough rationality.

While the economic approach to behavior builds on a theory of individual choice, it is not mainly concerned with individuals. It uses theory at the micro level as a powerful tool to derive implications at the group or macro level. Rational individual choice is combined with assumptions about technologies and other determinants of opportunities, equilibrium in market and nonmarket situations, and laws, norms, and traditions to obtain results concerning the behavior of groups. It is mainly because the theory derives implications at the macro level that it is of interest to policymakers and those studying differences among countries and cultures.

None of the theories considered in this lecture aims for the greatest generality; instead, each tries to derive concrete implications about behavior that can be tested with survey and other data. Disputes over

whether punishments deter crime, whether the lower earnings of women compared to those of men are mainly due to discrimination or lesser human capital, or whether no-fault divorce laws increase divorce rates—all raise questions about the empirical relevance of predictions derived from a theory based on individual rationality.

A close relation between theory and empirical testing helps prevent both the theoretical analysis and the empirical research from becoming sterile. Empirically oriented theories encourage the development of new sources and types of data, the way human capital theory stimulated the use of survey data, especially panels. At the same time, puzzling empirical results force changes in theory, as models of altruism and family preferences have been enriched to cope with the finding that parents in Western countries tend to bequeath equal amounts to different children.

I have been impressed by how many economists want to work on social issues rather than those forming the traditional core of economics. At the same time, specialists from fields that do consider social questions are often attracted to the economic way of modeling behavior because of the analytical power provided by the assumption of individual rationality. Thriving schools of rational choice theorists and empirical researchers are active in sociology, law, political science, and history and, to a lesser extent, in anthropology and psychology. The rational choice model provides the most promising basis presently available for a unified approach to the analysis of the social world by scholars from different social sciences.

Appendix

A

To develop a formal analysis, suppose that each person lives for three periods—young (y), middle age (m), and old age (o)—and has one child at the beginning of period m. A child's youth overlaps his parent's middle age, and a child's middle age overlaps his parent's old age. The utility parents get from altruism is assumed to be separable from the utilities produced by their own consumption.

A simple utility function of parents (V_p) incorporating these assumptions is

$$V_p = u_{mp} + \beta u_{op} + \beta a V_c, \tag{A1}$$

where β is the discount rate, and the degree of altruism rises with a. For selfish parents, $a = 0$. I do not permit parents to be sadistic toward children ($a < 0$), although the analysis is easily generalized to include sadists.

Each person works and earns income only during middle age. It is possible to save then to provide consumption for old age (Z_{op}) by accumulating assets with a yield of R_k. Parents influence children's earnings by investing in their

human capital. The marginal yield on these investments (R_h) is defined as

$$R_h = \frac{dE_c}{dh}, \tag{A2}$$

where E_c is the earnings of children at middle age, and h is the amount invested. This yield is assumed to decline as more is invested in children: $dR_h/dh \leq 0$.

Parents must also decide whether to leave bequests, denoted by k_c. If parents can consume at different ages, leave bequests, or invest in the child's human capital, their budget constraint is

$$Z_{mp} + h + \frac{Z_{op}}{R_k} + \frac{k_c}{R_k} = A_p, \tag{A3}$$

where A is the present value of resources.

One first-order condition to maximize parental utility determines their optimal consumption at middle and old age:

$$u'_{mp} = \beta R_k u'_{op} = \lambda_p, \tag{A4}$$

where λ_p is the parents' marginal utility of wealth. Another condition determines whether they give bequests:

$$\beta a V'_c \leq \frac{\lambda_p}{R_k} = \beta u'_{op}, \tag{A5}$$

and the last determines investments in the human capital of children:

$$R_h \beta a V'_c = \lambda_p. \tag{A6}$$

Equation (A6) assumes that the first-order condition for investment in human capital is a strict equality, that some human capital is always invested in children. This can be justified with an Inada-type condition that small investments in human capital yield very high rates of return. In rich economies such as Sweden or the United States, investments in basic knowledge and nutrition of children presumably do yield a very good return. As long as parents are not completely selfish—as long as $a > 0$—then such a condition does always imply positive investment in human capital. For completely selfish parents, equation (A6) would become an inequality.

Equation (A4) determines the accumulation of assets to finance old-age consumption. Whether parents leave bequests or want old-age support from their children is determined by the inequality in (A5). If this is a strict inequality, parents want support and would not leave bequests.

That inequality can be written in a more revealing way. If children also maximize their utility, then the envelope theorem implies that

$$a u'_{mc} < u'_{op} \quad \text{whenever} \quad a V'_c < u'_{op} \quad \text{since} \quad V'_c = u'_{mc}. \tag{A7}$$

Equation (A7) has the intuitive interpretation that parents do not give bequests when the utility the parents get from their children consuming a dollar more at middle age is less than the utility they get from a dollar more of their own consumption at old age. Obviously, such an inequality holds for completely selfish parents since the left-hand sides of equations (A5) and (A7) are zero when a is zero. The weaker the altruism (the smaller a), the more parents want from children.

Combining equations (A5) and (A6) gives

$$\frac{\lambda_p}{R_h} \leq \frac{\lambda_p}{R_k}, \quad \text{or } R_h \geq R_k. \tag{A8}$$

Equation (A8) implies that the marginal rate of return on human capital equals the return on assets when parents give bequests, and it is greater than the asset return when parents do not give bequests. Parents can help children either by investing in their human capital or by leaving them assets. Since they want to maximize the advantage to children, given the cost to themselves—parents are not sadistic—they help in the most efficient form.

Consequently, if strict inequality holds in equation (A8), they would not give bequests, for the best way to help children when the marginal return on human capital exceeds that on assets is to invest only in human capital. They leave bequests only when they get the same marginal return on both (some of these results have been derived in Becker and Tomes [1986]).

B

To analyze in a simple way the influence of parents over the formation of children's preferences, suppose parents can take actions x and y when children are young that affect their preferences when adults. I use the assumption of separability to write the utility function of middle-aged children as

$$V_c = u_{mc} + H(y) - G(x, g) + \beta u_{oc} + \dots. \tag{A9}$$

I assume that $H' > 0$ and $G_x > 0$, which means that an increase in y raises the utility of children, but an increase in x lowers their utility. Interpret H for concreteness as "happiness" and G as the "guilt" children feel toward their parents, so that greater x makes children feel guiltier. The question is, Why would nonsadistic parents want to make their children feel guilty?

The variable g is the key to understanding why. This measures the contribution of children to the old-age support of parents; let us assume that children feel less guilty when they contribute more ($G_g < 0$). If $G_{gx} > 0$, then greater x both raises children's guilt and stimulates more giving by them.

The budget constraint of parents becomes

$$Z_{mp} + h + x + y + \frac{Z_{op}}{R_k} + \frac{k_c}{R_k} = A_p + \frac{g}{R_k}. \tag{A10}$$

The first-order condition for the optimal y is

$$\beta a H' \leq \lambda_p. \tag{A11}$$

Since $H' > 0$, it is easy to understand why an altruistic parent may try to affect children's preferences through y since an increase in y makes children happier.

The first-order condition for x is more interesting, for even altruistic parents may want to make their children feel guilty if that sufficiently raises old-age support. This first-order condition can be written as

$$\frac{dV_p}{dx} = \frac{dg}{dx}\beta(u'_{op} - a u'_{mc}) - \beta a \frac{dG}{dx} \leq \lambda_p, \tag{A12}$$

where dG/dx incorporates the induced change in g. The second term in the middle expression is negative to altruistic parents because greater x does raise

children's guilt, which lowers the utility of these parents ($a > 0$). However, guilt also induces children to increase old-age support, as given by dg/dx. The magnitude of this response determines whether it is worthwhile for parents to make children feel guiltier.

Increased old-age support from children has two partially offsetting effects on the welfare of altruistic parents. On the one hand, it raises their old-age consumption and utility, as given by u'_{op}. On the other hand, it lowers children's consumption and, hence, the utility of altruistic parents, as given by $-au'_{mc}$. This means that altruistic parents who leave bequests never try to make children feel guiltier, for $u'_{op} = au'_{mc}$ for these parents. Since $dG/dx > 0$, they must be worse off when their children feel guiltier.

Equations (A5) and (A12) imply that

$$\frac{dg}{dx} - \frac{aG_x}{u'_{op}} = R_x \leq R_k. \tag{A13}$$

The marginal rate of return to altruistic parents from making children feel guiltier (given by R_x) nets out the parents' evaluation of the loss in children's utility from their guilt. Selfish parents ($a = 0$) ignore this loss and simply compare the effects of x and k on their consumption at old age.

C

Combine the first-order conditions in equations (A5) and (A6) to get

$$\frac{u'_{op}}{au'_{mc}} = \frac{R_h}{R_k}. \tag{A14}$$

Both sides of this equation exceed unity when parents do not give bequests. Since greater old-age support from children lowers the left-hand side by lowering the numerator and raising the denominator, the right-hand side must also fall to be in a utility-maximizing equilibrium. But since R_k is given by market conditions, the right-hand side can fall only if R_h falls, which implies greater investment in children when parents expect greater old-age support from children. Even completely selfish parents ($a = 0$) might invest in children if that would sufficiently increase the expected old-age support from guilty children.

References

Arrow, Kenneth J. "Models of Job Discrimination." In *Racial Discrimination in Economic Life,* edited by Anthony H. Pascal. Lexington, Mass.: Lexington Books, 1972.

———. "The Theory of Discrimination." In *Discrimination in Labor Markets,* edited by Orley Ashenfelter and Albert Rees. Princeton, N.J.: Princeton Univ. Press, 1973.

Arrow, Kenneth J., and Kurz, Mordecai. *Public Investment, the Rate of Return, and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Univ. Press (for Resources for the Future), 1970.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82 (November/December 1974): 1095–1117.

Barro, Robert J., and Sala-i-Martin, Xavier. "Convergence." *J.P.E.* 100 (April 1992): 223–51.

Beccaria, Cesare, marchese di. *On Crimes and Punishment.* Indianapolis: Hackett, 1986. Translation of *Dei delitti e delle pene* (1797).
Becker, Gary S. *The Economics of Discrimination.* Chicago: Univ. Chicago Press, 1957. 2d ed. 1971.
———. "Investment in Human Capital: A Theoretical Analysis." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 9–49.
———. *Human Capital.* New York: Columbia Univ. Press (for NBER), 1964. 2d ed. 1975.
———. "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.
———. "Crime and Punishment: An Economic Approach." *J.P.E.* 76 (March/April 1968): 169–217.
———. "A Theory of Social Interactions." *J.P.E.* 82 (November/December 1974): 1063–93.
———. *A Treatise on the Family.* Cambridge, Mass.: Harvard Univ. Press, 1981. Enl. ed. 1991.
Becker, Gary S., and Chiswick, Barry R. "Education and the Distribution of Earnings." *A.E.R. Papers and Proc.* 56 (May 1966): 358–69.
Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "An Economic Analysis of Marital Instability." *J.P.E.* 85 (December 1977): 1141–87.
Becker, Gary S., and Murphy, Kevin M. "A Theory of Rational Addiction." *J.P.E.* 96 (August 1988): 675–700.
Becker, Gary S.; Murphy, Kevin M.; and Tamura, Robert. "Human Capital, Fertility, and Economic Growth." *J.P.E.* 98, no. 5, pt. 2 (October 1990): S12–S37.
Becker, Gary S., and Stigler, George J. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *J. Legal Studies* 3 (January 1974): 1–18. Reprinted in *Chicago Studies in Political Economy,* by George J. Stigler. Chicago: Univ. Chicago Press, 1988.
Becker, Gary S., and Tomes, Nigel. "Human Capital and the Rise and Fall of Families." *J. Labor Econ.* 4, no. 3, pt. 2 (July 1986): S1–S39.
Bentham, Jeremy. *Theory of Legislation.* New York: Harcourt, Brace, 1931.
Bergstrom, Theodore C. "A Fresh Look at the Rotten Kid Theorem—and Other Household Mysteries." *J.P.E.* 97 (October 1989): 1138–59.
Boserup, Ester. "Inequality between the Sexes." In *The New Palgrave: A Dictionary of Economics,* edited by John Eatwell, Murray Milgate, and Peter Newman. New York: Stockton, 1987.
Cain, Bruce E.; Ferejohn, John; and Fiorina, Morris. *The Personal Vote: Constituency Service and Electoral Independence.* Cambridge, Mass.: Harvard Univ. Press, 1987.
Cain, Glen G. "The Economic Analysis of Labor Market Discrimination: A Survey." In *Handbook of Labor Economics,* vol. 1, edited by Orley Ashenfelter and Richard Layard. Handbooks in Economics Series, no. 5. New York: Elsevier Sci., 1986.
Cass, David. "Optimum Growth in an Aggregative Model of Capital Accumulation." *Rev. Econ. Studies* 32 (July 1965): 233–40.
Denison, Edward F. *Sources of Economic Growth in the United States.* Washington: Comm. Econ. Development, 1962.
Edgeworth, Francis Y. "Equal Pay to Men and Women for Equal Work." *Econ. J.* 32 (December 1922): 431–57.
Ehrlich, Isaac. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *J.P.E.* 81 (May/June 1973): 521–65.

———. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *A.E.R.* 65 (June 1975): 397–417.
Farrell, C., and Mandel, M. "Uncommon Sense." *Bus. Week* (October 26, 1992), pp. 36–37.
Fawcett, Millicent G. "Equal Pay for Equal Work." *Econ. J.* 28 (March 1918): 1–6.
Goldin, Claudia. *Understanding the Gender Gap: An Economic History of American Women.* Series on Long-Term Factors in Economic Development. New York: Oxford Univ. Press (for NBER), 1990.
Grossman, Sanford J., and Hart, Oliver D. "An Analysis of the Principal-Agent Problem." *Econometrica* 51 (January 1983): 7–45.
Hernandez, Donald. *When Households Continue, Discontinue, and Form.* Washington: U.S. Bur. Census, 1992.
Hutt, William H. *The Economics of the Colour Bar: A Study of the Economic Origins and Consequences of Racial Segregation in South Africa.* London: Deutsch (for Inst. Econ. Affairs), 1964.
Kandel, Eugene, and Lazear, Edward P. "Peer Pressure and Partnerships." *J.P.E.* 100 (August 1992): 801–17.
Landes, William M., and Posner, Richard A. "The Private Enforcement of Law." *J. Legal Studies* 4 (January 1975): 1–46.
Lindbeck, Assar, and Weibull, Jörgen W. "Altruism and Time Consistency: The Economics of Fait Accompli." *J.P.E.* 96 (December 1988): 1165–82.
Linder, Staffan Burenstam. *The Harried Leisure Class.* New York: Columbia Univ. Press, 1970.
Loury, Glenn C. "Incentive Effects of Affirmative Action." *Ann. American Acad. Polit. and Soc. Sci.* 523 (September 1992): 19–29.
Lucas, Robert E., Jr. "On the Mechanics of Economic Development." *J. Monetary Econ.* 22 (July 1988): 3–42.
Lundahl, Mats. *Apartheid in Theory and Practice: An Economic Analysis.* Boulder, Colo.: Westview, 1992.
McElroy, Marjorie B., and Horney, Mary Jean. "Nash-bargained Household Decisions: Toward a Generalization of the Theory of Demand." *Internat. Econ. Rev.* 22 (June 1981): 333–49.
Meltzer, David. "Mortality Decline, the Demographic Transition and Economic Growth." Ph.D. dissertation, Univ. Chicago, 1992.
Menninger, Karl. *The Crime of Punishment.* New York: Viking, 1966.
Mincer, Jacob. "Labor Force Participation of Married Women." In *Aspects of Labor Economics.* Universities–National Bureau Committee for Economic Research, no. 14. Princeton, N.J.: Princeton Univ. Press (for NBER), 1962.
———. *Schooling, Experience, and Earnings.* New York: Columbia Univ. Press (for NBER), 1974.
Murphy, Kevin M., and Welch, Finis. "The Structure of Wages." *Q.J.E.* 107 (February 1992): 285–326.
Myrdal, Gunnar. *An American Dilemma: The Negro Problem and Modern Democracy.* 2 vols. New York: Random House, 1944.
National Research Council. Panel of Research on Deterrent and Incapacitative Effects. *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates,* edited by Alfred Blumstein, Jacqueline Cohen, and Daniel Nagin. Washington: Nat. Acad. Sci., 1978.
Oi, Walter Y. "Labor as a Quasi-fixed Factor." *J.P.E.* 70 (December 1962): 538–55.
O'Neill, June. "The Trend in the Male-Female Wage Gap in the United States." *J. Labor Econ.* 3, no. 1, pt. 2 (January 1985): S91–S116.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *A.E.R.* 62 (September 1972): 659–61.
Polinsky, A. Mitchell, and Shavell, Steven. "The Optimal Use of Fines and Imprisonment." *J. Public Econ.* 24 (June 1984): 89–99.
Posner, Richard A. *Economic Analysis of Law.* 3d ed. Boston: Little, Brown, 1986.
Psacharopoulos, George. "Returns to Education: A Further International Update and Implications." *J. Human Resources* 20 (Fall 1985): 583–604.
Romer, Paul M. "Increasing Returns and Long-Run Growth." *J.P.E.* 94 (October 1986): 1002–37.
Schultz, Theodore W. *The Economic Value of Education.* New York: Columbia Univ. Press, 1963.
Soffer, O. "Before Beringia: Late Pleistocene Bio-social Transformations and the Colonization of Northern Eurasia." In *Chronostratigraphy of the Paleolithic in North Central, East Asia and America.* Novosibirsk: Acad. Sci. USSR, 1990.
Stigler, George J. "The Optimum Enforcement of Laws." *J.P.E.* 78 (May/June 1970): 526–36.
U.S. Sentencing Commission. *Federal Sentencing Guidelines Manual.* Washington: Government Printing Office, 1992.
Williamson, Oliver E. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting.* New York: Free Press, 1985.

AN INQUIRY INTO THE NATURE AND CAUSES OF THE WEALTH OF NATIONS

by

Adam Smith

INTRODUCTION AND PLAN OF THE WORK

THE ANNUAL LABOUR of every nation is the fund which originally supplies it with all the necessaries and conveniencies of life which it annually consumes, and which consist always either in the immediate produce of that labour, or in what is purchased with that produce from other nations.

According, therefore, as this produce, or what is purchased with it, bears a greater or smaller proportion to the number of those who are to consume it, the nation will be better or worse supplied with all the necessaries and conveniencies for which it has occasion.

But this proportion must in every nation be regulated by two different circumstances: first, by the skill, dexterity, and judgment with which its labour is generally applied; and, secondly, by the proportion between the number of those who are employed in useful labour, and that of those who are not so employed. Whatever be the soil, climate, or extent of territory of any particular nation, the abundance or scantiness of its annual supply must, in that particular situation, depend upon those two circumstances.

The abundance or scantiness of this supply, too, seems to depend more upon the former of those two circumstances than upon the latter. Among the savage nations of hunters and fishers, every individual who is able to work is more or less employed in useful labour, and endeavours to provide, as well as he can, the necessaries and conveniencies of life, for himself, and such of his family or tribe as are either too old, or too young, or too infirm, to go a-hunting and fishing. Such nations, however, are so miserably poor, that, from mere want, they are frequently reduced, or at least think themselves reduced, to the necessity sometimes of directly destroying, and sometimes of abandoning their infants, their old people,

and those afflicted with lingering diseases, to perish with hunger, or to be devoured by wild beasts. Among civilized and thriving nations, on the contrary, though a great number of people do not labour at all, many of whom consume the produce of ten times, frequently of a hundred times, more labour than the greater part of those who work; yet the produce of the whole labour of the society is so great, that all are often abundantly supplied; and a workman, even of the lowest and poorest order, if he is frugal and industrious, may enjoy a greater share of the necessaries and conveniencies of life than it is possible for any savage to acquire.

The causes of this improvement in the productive powers of labour, and the order according to which its produce is naturally distributed among the different ranks and conditions of men in the society, make the subject of the first book of this Inquiry.

Whatever be the actual state of the skill, dexterity, and judgment, with which labour is applied in any nation, the abundance or scantiness of its annual supply must depend, during the continuance of that state, upon the proportion between the number of those who are annually employed in useful labour, and that of those who are not so employed. The number of useful and productive labourers, it will hereafter appear, is everywhere in proportion to the quantity of capital stock which is employed in setting them to work, and to the particular way in which it is so employed. The second book, therefore, treats of the nature of capital stock, of the manner in which it is gradually accumulated, and of the different quantities of labour which it puts into motion, according to the different ways in which it is employed.

Nations tolerably well advanced as to skill, dexterity, and judgment, in the application of labour, have followed very different plans in the general conduct or direction of it; and those plans have not all been equally favourable to the greatness of its produce. The policy of some nations has given extraordinary encouragement to the industry of the country; that of others to the industry of towns. Scarce any nation has dealt equally and impartially with every sort of industry. Since the down-fall of the Roman empire, the policy of Europe has been more favourable to arts, manufactures, and commerce, the industry of towns, than to agriculture, the Industry of the country. The circumstances which seem to have introduced and established this policy are explained in the third book.

Though those different plans were, perhaps, first introduced by the private interests and prejudices of particular orders of men, without any regard to, or foresight of, their consequences upon the general welfare of the society; yet they have given occasion to very different theories of political economy; of which some magnify the importance of that industry which is carried on in towns, others of that which is carried on in the country. Those theories have had a

considerable influence, not only upon the opinions of men of learning, but upon the public conduct of princes and sovereign states. I have endeavoured, in the fourth book, to explain as fully and distinctly as I can those different theories, and the principal effects which they have produced in different ages and nations.

To explain in what has consisted the revenue of the great body of the people, or what has been the nature of those funds, which, in different ages and nations, have supplied their annual consumption, is the object of these four first books. The fifth and last book treats of the revenue of the sovereign, or commonwealth. In this book I have endeavoured to shew, first, what are the necessary expenses of the sovereign, or commonwealth; which of those expenses ought to be defrayed by the general contribution of the whole society, and which of them, by that of some particular part only, or of some particular members of it: secondly, what are the different methods in which the whole society may be made to contribute towards defraying the expenses incumbent on the whole society, and what are the principal advantages and inconveniencies of each of those methods; and, thirdly and lastly, what are the reasons and causes which have induced almost all modern governments to mortgage some part of this revenue, or to contract debts; and what have been the effects of those debts upon the real wealth, the annual produce of the land and labour of the society.

BOOK I

OF THE CAUSES OF IMPROVEMENT IN THE PRODUCTIVE POWERS OF LABOUR, AND OF THE ORDER ACCORDING TO WHICH ITS PRODUCE IS NATURALLY DISTRIBUTED AMONG THE DIFFERENT RANKS OF THE PEOPLE.

CHAPTER I

OF THE DIVISION OF LABOUR

THE GREATEST IMPROVEMENTS in the productive powers of labour, and the greater part of the skill, dexterity, and judgment, with which it is anywhere directed, or applied, seem to have been the effects of the division of labour. The effects of the division of labour, in the general business of society, will be more easily understood, by considering in what manner it operates in some particular manufactures. It is commonly supposed to be carried furthest in some very trifling ones; not perhaps that it really is carried further in them than in others of more

importance: but in those trifling manufactures which are destined to supply the small wants of but a small number of people, the whole number of workmen must necessarily be small; and those employed in every different branch of the work can often be collected into the same workhouse, and placed at once under the view of the spectator.

In those great manufactures, on the contrary, which are destined to supply the great wants of the great body of the people, every different branch of the work employs so great a number of workmen, that it is impossible to collect them all into the same workhouse. We can seldom see more, at one time, than those employed in one single branch. Though in such manufactures, therefore, the work may really be divided into a much greater number of parts, than in those of a more trifling nature, the division is not near so obvious, and has accordingly been much less observed.

To take an example, therefore, from a very trifling manufacture, but one in which the division of labour has been very often taken notice of, the trade of a pin-maker: a workman not educated to this business (which the division of labour has rendered a distinct trade, nor acquainted with the use of the machinery employed in it (to the invention of which the same division of labour has probably given occasion), could scarce, perhaps, with his utmost industry, make one pin in a day, and certainly could not make twenty. But in the way in which this business is now carried on, not only the whole work is a peculiar trade, but it is divided into a number of branches, of which the greater part are likewise peculiar trades. One man draws out the wire; another straights it; a third cuts it; a fourth points it; a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on is a peculiar business; to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufactories, are all performed by distinct hands, though in others the same man will sometimes perform two or three of them. I have seen a small manufactory of this kind, where ten men only were employed, and where some of them consequently performed two or three distinct operations. But though they were very poor, and therefore but indifferently accommodated with the necessary machinery, they could, when they exerted themselves, make among them about twelve pounds of pins in a day. There are in a pound upwards of four thousand pins of a middling size. Those ten persons, therefore, could make among them upwards of forty-eight thousand pins in a day. Each person, therefore, making a tenth part of forty-eight thousand pins, might be considered as making four thousand eight hundred pins in a day. But if they had all

wrought separately and independently, and without any of them having been educated to this peculiar business, they certainly could not each of them have made twenty, perhaps not one pin in a day; that is, certainly, not the two hundred and fortieth, perhaps not the four thousand eight hundredth, part of what they are at present capable of performing, in consequence of a proper division and combination of their different operations.

In every other art and manufacture, the effects of the division of labour are similar to what they are in this very trifling one, though, in many of them, the labour can neither be so much subdivided, nor reduced to so great a simplicity of operation. The division of labour, however, so far as it can be introduced, occasions, in every art, a proportionable increase of the productive powers of labour. The separation of different trades and employments from one another, seems to have taken place in consequence of this advantage. This separation, too, is generally carried furthest in those countries which enjoy the highest degree of industry and improvement; what is the work of one man, in a rude state of society, being generally that of several in an improved one. In every improved society, the farmer is generally nothing but a farmer; the manufacturer, nothing but a manufacturer. The labour, too, which is necessary to produce any one complete manufacture, is almost always divided among a great number of hands. How many different trades are employed in each branch of the linen and woollen manufactures, from the growers of the flax and the wool, to the bleachers and smoothers of the linen, or to the dyers and dressers of the cloth! The nature of agriculture, indeed, does not admit of so many subdivisions of labour, nor of so complete a separation of one business from another, as manufactures. It is impossible to separate so entirely the business of the grazier from that of the corn-farmer, as the trade of the carpenter is commonly separated from that of the smith. The spinner is almost always a distinct person from the, weaver; but the ploughman, the harrower, the sower of the seed, and the reaper of the corn, are often the same. The occasions for those different sorts of labour returning with the different seasons of the year, it is impossible that one man should be constantly employed in any one of them. This impossibility of making so complete and entire a separation of all the different branches of labour employed in agriculture, is perhaps the reason why the improvement of the productive powers of labour, in this art, does not always keep pace with their improvement in manufactures. The most opulent nations, indeed, generally excel all their neighbours in agriculture as well as in manufactures; but they are commonly more distinguished by their superiority in the latter than in the former. Their lands are in general better cultivated, and having more labour and expense bestowed

upon them, produce more in proportion to the extent and natural fertility of the ground. But this superiority of produce is seldom much more than in proportion to the superiority of labour and expense. In agriculture, the labour of the rich country is not always much more productive than that of the poor; or, at least, it is never so much more productive, as it commonly is in manufactures. The corn of the rich country, therefore, will not always, in the same degree of goodness, come cheaper to market than that of the poor. The corn of Poland, in the same degree of goodness, is as cheap as that of France, notwithstanding the superior opulence and improvement of the latter country. The corn of France is, in the corn-provinces, fully as good, and in most years nearly about the same price with the corn of England, though, in opulence and improvement, France is perhaps inferior to England. The corn-lands of England, however, are better cultivated than those of France, and the corn-lands of France are said to be much better cultivated than those of Poland. But though the poor country, notwithstanding the inferiority of its cultivation, can, in some measure, rival the rich in the cheapness and goodness of its corn, it can pretend to no such competition in its manufactures, at least if those manufactures suit the soil, climate, and situation, of the rich country. The silks of France are better and cheaper than those of England, because the silk manufacture, at least under the present high duties upon the importation of raw silk, does not so well suit the climate of England as that of France. But the hardware and the coarse woollens of England are beyond all comparison superior to those of France, and much cheaper, too, in the same degree of goodness. In Poland there are said to be scarce any manufactures of any kind, a few of those coarser household manufactures excepted, without which no country can well subsist.

This great increase in the quantity of work, which, in consequence of the division of labour, the same number of people are capable of performing, is owing to three different circumstances; first, to the increase of dexterity in every particular workman; secondly, to the saving of the time which is commonly lost in passing from one species of work to another; and, lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many.

First, the improvement of the dexterity of the workmen, necessarily increases the quantity of the work he can perform; and the division of labour, by reducing every man's business to some one simple operation, and by making this operation the sole employment of his life, necessarily increases very much the dexterity of the workman. A common smith, who, though accustomed to handle the hammer, has never been used to make nails, if, upon some particular occasion, he is obliged to attempt it, will scarce, I

am assured, be able to make above two or three hundred nails in a day, and those, too, very bad ones. A smith who has been accustomed to make nails, but whose sole or principal business has not been that of a nailer, can seldom, with his utmost diligence, make more than eight hundred or a thousand nails in a day. I have seen several boys, under twenty years of age, who had never exercised any other trade but that of making nails, and who, when they exerted themselves, could make, each of them, upwards of two thousand three hundred nails in a day. The making of a nail, however, is by no means one of the simplest operations. The same person blows the bellows, stirs or mends the fire as there is occasion, heats the iron, and forges every part of the nail: in forging the head, too, he is obliged to change his tools. The different operations into which the making of a pin, or of a metal button, is subdivided, are all of them much more simple, and the dexterity of the person, of whose life it has been the sole business to perform them, is usually much greater. The rapidity with which some of the operations of those manufactures are performed, exceeds what the human hand could, by those who had never seen them, he supposed capable of acquiring.

Secondly, The advantage which is gained by saving the time commonly lost in passing from one sort of work to another, is much greater than we should at first view be apt to imagine it. It is impossible to pass very quickly from one kind of work to another, that is carried on in a different place, and with quite different tools. A country weaver, who cultivates a small farm, must loose a good deal of time in passing from his loom to the field, and from the field to his loom. When the two trades can be carried on in the same workhouse, the loss of time is, no doubt, much less. It is, even in this case, however, very considerable. A man commonly saunters a little in turning his hand from one sort of employment to another. When he first begins the new work, he is seldom very keen and hearty; his mind, as they say, does not go to it, and for some time he rather trifles than applies to good purpose. The habit of sauntering, and of indolent careless application, which is naturally, or rather necessarily, acquired by every country workman who is obliged to change his work and his tools every half hour, and to apply his hand in twenty different ways almost every day of his life, renders him almost always slothful and lazy, and incapable of any vigorous application, even on the most pressing occasions. Independent, therefore, of his deficiency in point of dexterity, this cause alone must always reduce considerably the quantity of work which he is capable of performing.

Thirdly, and lastly, everybody must be sensible how much labour is facilitated and abridged by the application of proper machinery. It is unnecessary to give any example. I shall only observe, there-

fore, that the invention of all those machines by which labour is so much facilitated and abridged, seems to have been originally owing to the division of labour. Men are much more likely to discover easier and readier methods of attaining any object, when the whole attention of their minds is directed towards that single object, than when it is dissipated among a great variety of things. But, in consequence of the division of labour, the whole of every man's attention comes naturally to be directed towards some one very simple object. It is naturally to be expected, therefore, that some one or other of those who are employed in each particular branch of labour should soon find out easier and readier methods of performing their own particular work, whenever the nature of it admits of such improvement. A great part of the machines made use of in those manufactures in which labour is most subdivided, were originally the invention of common workmen, who, being each of them employed in some very simple operation, naturally turned their thoughts towards finding out easier and readier methods of performing it. Whoever has been much accustomed to visit such manufactures, must frequently have been shewn very pretty machines, which were the inventions of such workmen, in order to facilitate and quicken their own particular part of the work. In the first fire engines {this was the current designation for steam engines}, a boy was constantly employed to open and shut alternately the communication between the boiler and the cylinder, according as the piston either ascended or descended. One of those boys, who loved to play with his companions, observed that, by tying a string from the handle of the valve which opened this communication to another part of the machine, the valve would open and shut without his assistance, and leave him at liberty to divert himself with his play-fellows. One of the greatest improvements that has been made upon this machine, since it was first invented, was in this manner the discovery of a boy who wanted to save his own labour.

All the improvements in machinery, however, have by no means been the inventions of those who had occasion to use the machines. Many improvements have been made by the ingenuity of the makers of the machines, when to make them became the business of a peculiar trade; and some by that of those who are called philosophers, or men of speculation, whose trade it is not to do any thing, but to observe every thing, and who, upon that account, are often capable of combining together the powers of the most distant and dissimilar objects in the progress of society, philosophy or speculation becomes, like every other employment, the principal or sole trade and occupation of a particular class of citizens. Like every other employment, too, it is subdivided into a great number of different branches, each of which affords occupa-

tion to a peculiar tribe or class of philosophers; and this subdivision of employment in philosophy, as well as in every other business, improve dexterity, and saves time. Each individual becomes more expert in his own peculiar branch, more work is done upon the whole, and the quantity of science is considerably increased by it.

It is the great multiplication of the productions of all the different arts, in consequence of the division of labour, which occasions, in a well-governed society, that universal opulence which extends itself to the lowest ranks of the people. Every workman has a great quantity of his own work to dispose of beyond what he himself has occasion for; and every other workman being exactly in the same situation, he is enabled to exchange a great quantity of his own goods for a great quantity or, what comes to the same thing, for the price of a great quantity of theirs. He supplies them abundantly with what they have occasion for, and they accommodate him as amply with what he has occasion for, and a general plenty diffuses itself through all the different ranks of the society.

Observe the accommodation of the most common artificer or daylabourer in a civilized and thriving country, and you will perceive that the number of people, of whose industry a part, though but a small part, has been employed in procuring him this accommodation, exceeds all computation. The woollen coat, for example, which covers the day-labourer, as coarse and rough as it may appear, is the produce of the joint labour of a great multitude of workmen. The shepherd, the sorter of the wool, the wool-comber or carder, the dyer, the scribbler, the spinner, the weaver, the fuller, the dresser, with many others, must all join their different arts in order to complete even this homely production. How many merchants and carriers, besides, must have been employed in transporting the materials from some of those workmen to others who often live in a very distant part of the country? How much commerce and navigation in particular, how many ship-builders, sailors, sail-makers, rope-makers, must have been employed in order to bring together the different drugs made use of by the dyer, which often come from the remotest corners of the world? What a variety of labour, too, is necessary in order to produce the tools of the meanest of those workmen! To say nothing of such complicated machines as the ship of the sailor, the mill of the fuller, or even the loom of the weaver, let us consider only what a variety of labour is requisite in order to form that very simple machine, the shears with which the shepherd clips the wool. The miner, the builder of the furnace for smelting the ore the feller of the timber, the burner of the charcoal to be made use of in the smelting-house, the brickmaker, the bricklayer, the workmen who attend the furnace, the millwright, the forger, the smith, must all of them join their different arts in order to produce them. Were we to

examine, in the same manner, all the different parts of his dress and household furniture, the coarse linen shirt which he wears next his skin, the shoes which cover his feet, the bed which he lies on, and all the different parts which compose it, the kitchen-grate at which he prepares his victuals, the coals which he makes use of for that purpose, dug from the bowels of the earth, and brought to him, perhaps, by a long sea and a long land-carriage, all the other utensils of his kitchen, all the furniture of his table, the knives and forks, the earthen or pewter plates upon which he serves up and divides his victuals, the different hands employed in preparing his bread and his beer, the glass window which lets in the heat and the light, and keeps out the wind and the rain, with all the knowledge and art requisite for preparing that beautiful and happy invention, without which these northern parts of the world could scarce have afforded a very comfortable habitation, together with the tools of all the different workmen employed in producing those different conveniencies; if we examine, I say, all these things, and consider what a variety of labour is employed about each of them, we shall be sensible that, without the assistance and co-operation of many thousands, the very meanest person in a civilized country could not be provided, even according to, what we very falsely imagine, the easy and simple manner in which he is commonly accommodated. Compared, indeed, with the more extravagant luxury of the great, his accommodation must no doubt appear extremely simple and easy; and yet it may be true, perhaps, that the accommodation of an European prince does not always so much exceed that of an industrious and frugal peasant, as the accommodation of the latter exceeds that of many an African king, the absolute masters of the lives and liberties of ten thousand naked savages.

The Journal of

LAW & ECONOMICS

VOLUME III OCTOBER 1960

THE PROBLEM OF SOCIAL COST

R. H. COASE
University of Virginia

I. The Problem To Be Examined[1]

This paper is concerned with those actions of business firms which have harmful effects on others. The standard example is that of a factory the smoke from which has harmful effects on those occupying neighbouring properties. The economic analysis of such a situation has usually proceeded in terms of a divergence between the private and social product of the factory, in which economists have largely followed the treatment of Pigou in *The Economics of Welfare*. The conclusions to which this kind of analysis seems to have led most economists is that it would be desirable to make the owner of the factory liable for the damage caused to those injured by the smoke, or alternatively, to place a tax on the factory owner varying with the amount of smoke produced and equivalent in money terms to the damage it would cause, or finally, to exclude the factory from residential districts (and presumably from other

[1] This article, although concerned with a technical problem of economic analysis, arose out of the study of the Political Economy of Broadcasting which I am now conducting. The argument of the present article was implicit in a previous article dealing with the problem of allocating radio and television frequencies (The Federal Communications Commission, 2 J. Law & Econ. [1959]) but comments which I have received seemed to suggest that it would be desirable to deal with the question in a more explicit way and without reference to the original problem for the solution of which the analysis was developed.

areas in which the emission of smoke would have harmful effects on others). It is my contention that the suggested courses of action are inappropriate, in that they lead to results which are not necessarily, or even usually, desirable.

II. The Reciprocal Nature of the Problem

The traditional approach has tended to obscure the nature of the choice that has to be made. The question is commonly thought of as one in which A inflicts harm on B and what has to be decided is: how should we restrain A? But this is wrong. We are dealing with a problem of a reciprocal nature. To avoid the harm to B would inflict harm on A. The real question that has to be decided is: should A be allowed to harm B or should B be allowed to harm A? The problem is to avoid the more serious harm. I instanced in my previous article[2] the case of a confectioner the noise and vibrations from whose machinery disturbed a doctor in his work. To avoid harming the doctor would inflict harm on the confectioner. The problem posed by this case was essentially whether it was worth while, as a result of restricting the methods of production which could be used by the confectioner, to secure more doctoring at the cost of a reduced supply of confectionery products. Another example is afforded by the problem of straying cattle which destroy crops on neighbouring land. If it is inevitable that some cattle will stray, an increase in the supply of meat can only be obtained at the expense of a decrease in the supply of crops. The nature of the choice is clear: meat or crops. What answer should be given is, of course, not clear unless we know the value of what is obtained as well as the value of what is sacrificed to obtain it. To give another example, Professor George J. Stigler instances the contamination of a stream.[3] If we assume that the harmful effect of the pollution is that it kills the fish, the question to be decided is: is the value of the fish lost greater or less than the value of the product which the contamination of the stream makes possible. It goes almost without saying that this problem has to be looked at in total *and* at the margin.

III. The Pricing System with Liability for Damage

I propose to start my analysis by examining a case in which most economists would presumably agree that the problem would be solved in a completely satisfactory manner: when the damaging business has to pay for all damage caused *and* the pricing system works smoothly (strictly this means that the operation of a pricing system is without cost).

A good example of the problem under discussion is afforded by the case of straying cattle which destroy crops growing on neighbouring land. Let us suppose that a farmer and a cattle-raiser are operating on neighbouring proper-

[2] Coase, The Federal Communications Commission, 2 J. Law & Econ. 26–27 (1959).

[3] G. J. Stigler, The Theory of Price 105 (1952).

ties. Let us further suppose that, without any fencing between the properties, an increase in the size of the cattle-raiser's herd increases the total damage to the farmer's crops. What happens to the marginal damage as the size of the herd increases is another matter. This depends on whether the cattle tend to follow one another or to roam side by side, on whether they tend to be more or less restless as the size of the herd increases and on other similar factors. For my immediate purpose, it is immaterial what assumption is made about marginal damage as the size of the herd increases.

To simplify the argument, I propose to use an arithmetical example. I shall assume that the annual cost of fencing the farmer's property is $9 and that the price of the crop is $1 per ton. Also, I assume that the relation between the number of cattle in the herd and the annual crop loss is as follows:

Number in Herd (Steers)	Annual Crop Loss (Tons)	Crop Loss per Additional Steer (Tons)
1	1	1
2	3	2
3	6	3
4	10	4

Given that the cattle-raiser is liable for the damage caused, the additional annual cost imposed on the cattle-raiser if he increased his herd from, say, 2 to 3 steers is $3 and in deciding on the size of the herd, he will take this into account along with his other costs. That is, he will not increase the size of the herd unless the value of the additional meat produced (assuming that the cattle-raiser slaughters the cattle), is greater than the additional costs that this will entail, including the value of the additional crops destroyed. Of course, if, by the employment of dogs, herdsmen, aeroplanes, mobile radio and other means, the amount of damage can be reduced, these means will be adopted when their cost is less than the value of the crop which they prevent being lost. Given that the annual cost of fencing is $9, the cattle-raiser who wished to have a herd with 4 steers or more would pay for fencing to be erected and maintained, assuming that other means of attaining the same end would not do so more cheaply. When the fence is erected, the marginal cost due to the liability for damage becomes zero, except to the extent that an increase in the size of the herd necessitates a stronger and therefore more expensive fence because more steers are liable to lean against it at the same time. But, of course, it may be cheaper for the cattle-raiser not to fence and to pay for the damaged crops, as in my arithmetical example, with 3 or fewer steers.

It might be thought that the fact that the cattle-raiser would pay for all crops damaged would lead the farmer to increase his planting if a cattle-raiser came to occupy the neighbouring property. But this is not so. If the crop was previously sold in conditions of perfect competition, marginal cost was equal

to price for the amount of planting undertaken and any expansion would have reduced the profits of the farmer. In the new situation, the existence of crop damage would mean that the farmer would sell less on the open market but his receipts for a given production would remain the same, since the cattle-raiser would pay the market price for any crop damaged. Of course, if cattle-raising commonly involved the destruction of crops, the coming into existence of a cattle-raising industry might raise the price of the crops involved and farmers would then extend their planting. But I wish to confine my attention to the individual farmer.

I have said that the occupation of a neighbouring property by a cattle-raiser would not cause the amount of production, or perhaps more exactly the amount of planting, by the farmer to increase. In fact, if the cattle-raising has any effect, it will be to decrease the amount of planting. The reason for this is that, for any given tract of land, if the value of the crop damaged is so great that the receipts from the sale of the undamaged crop are less than the total costs of cultivating that tract of land, it will be profitable for the farmer and the cattle-raiser to make a bargain whereby that tract of land is left uncultivated. This can be made clear by means of an arithmetical example. Assume initially that the value of the crop obtained from cultivating a given tract of land is $12 and that the cost incurred in cultivating this tract of land is $10, the net gain from cultivating the land being $2. I assume for purposes of simplicity that the farmer owns the land. Now assume that the cattle-raiser starts operations on the neighbouring property and that the value of the crops damaged is $1. In this case $11 is obtained by the farmer from sale on the market and $1 is obtained from the cattle-raiser for damage suffered and the net gain remains $2. Now suppose that the cattle-raiser finds it profitable to increase the size of his herd, even though the amount of damage rises to $3; which means that the value of the additional meat production is greater than the additional costs, including the additional $2 payment for damage. But the total payment for damage is now $3. The net gain to the farmer from cultivating the land is still $2. The cattle-raiser would be better off if the farmer would agree not to cultivate his land for any payment less than $3. The farmer would be agreeable to not cultivating the land for any payment greater than $2. There is clearly room for a mutually satisfactory bargain which would lead to the abandonment of cultivation.[4] But the same argument applies not only to the whole tract cultivated by the farmer but also to any

[4] The argument in the text has proceeded on the assumption that the alternative to cultivation of the crop is abandonment of cultivation altogether. But this need not be so. There may be crops which are less liable to damage by cattle but which would not be as profitable as the crop grown in the absence of damage. Thus, if the cultivation of a new crop would yield a return to the farmer of $1 instead of $2, and the size of the herd which would cause $3 damage with the old crop would cause $1 damage with the new crop, it would be profitable to the cattle-raiser to pay any sum less than $2 to induce the farmer

subdivision of it. Suppose, for example, that the cattle have a well-defined route, say, to a brook or to a shady area. In these circumstances, the amount of damage to the crop along the route may well be great and if so, it could be that the farmer and the cattle-raiser would find it profitable to make a bargain whereby the farmer would agree not to cultivate this strip of land.

But this raises a further possibility. Suppose that there is such a well-defined route. Suppose further that the value of the crop that would be obtained by cultivating this strip of land is $10 but that the cost of cultivation is $11. In the absence of the cattle-raiser, the land would not be cultivated. However, given the presence of the cattle-raiser, it could well be that if the strip was cultivated, the whole crop would be destroyed by the cattle. In which case, the cattle-raiser would be forced to pay $10 to the farmer. It is true that the farmer would lose $1. But the cattle-raiser would lose $10. Clearly this is a situation which is not likely to last indefinitely since neither party would want this to happen. The aim of the farmer would be to induce the cattle-raiser to make a payment in return for an agreement to leave this land uncultivated. The farmer would not be able to obtain a payment greater than the cost of fencing off this piece of land nor so high as to lead the cattle-raiser to abandon the use of the neighbouring property. What payment would in fact be made would depend on the shrewdness of the farmer and the cattle-raiser as bargainers. But as the payment would not be so high as to cause the cattle-raiser to abandon this location and as it would not vary with the size of the herd, such an agreement would not affect the allocation of resources but would merely alter the distribution of income and wealth as between the cattle-raiser and the farmer.

I think it is clear that if the cattle-raiser is liable for damage caused and the pricing system works smoothly, the reduction in the value of production elsewhere will be taken into account in computing the additional cost involved in increasing the size of the herd. This cost will be weighed against the value of the additional meat production and, given perfect competition in the cattle industry, the allocation of resources in cattle-raising will be optimal. What needs to be emphasized is that the fall in the value of production elsewhere which would be taken into account in the costs of the cattle-raiser may well be less than the damage which the cattle would cause to the crops in the ordinary course of events. This is because it is possible, as a result of market transactions, to discontinue cultivation of the land. This is desirable in all

---

to change his crop (since this would reduce damage liability from $3 to $1) and it would be profitable for the farmer to do so if the amount received was more than $1 (the reduction in his return caused by switching crops). In fact, there would be room for a mutually satisfactory bargain in all cases in which a change of crop would reduce the amount of damage by more than it reduces the value of the crop (excluding damage)—in all cases, that is, in which a change in the crop cultivated would lead to an increase in the value of production.

cases in which the damage that the cattle would cause, and for which the cattle-raiser would be willing to pay, exceeds the amount which the farmer would pay for use of the land. In conditions of perfect competition, the amount which the farmer would pay for the use of the land is equal to the difference between the value of the total production when the factors are employed on this land and the value of the additional product yielded in their next best use (which would be what the farmer would have to pay for the factors). If damage exceeds the amount the farmer would pay for the use of the land, the value of the additional product of the factors employed elsewhere would exceed the value of the total product in this use after damage is taken into account. It follows that it would be desirable to abandon cultivation of the land and to release the factors employed for production elsewhere. A procedure which merely provided for payment for damage to the crop caused by the cattle but which did not allow for the possibility of cultivation being discontinued would result in too small an employment of factors of production in cattle-raising and too large an employment of factors in cultivation of the crop. But given the possibility of market transactions, a situation in which damage to crops exceeded the rent of the land would not endure. Whether the cattle-raiser pays the farmer to leave the land uncultivated or himself rents the land by paying the land-owner an amount slightly greater than the farmer would pay (if the farmer was himself renting the land), the final result would be the same and would maximise the value of production. Even when the farmer is induced to plant crops which it would not be profitable to cultivate for sale on the market, this will be a purely short-term phenomenon and may be expected to lead to an agreement under which the planting will cease. The cattle-raiser will remain in that location and the marginal cost of meat production will be the same as before, thus having no long-run effect on the allocation of resources.

IV. The Pricing System with No Liability for Damage

I now turn to the case in which, although the pricing system is assumed to work smoothly (that is, costlessly), the damaging business is not liable for any of the damage which it causes. This business does not have to make a payment to those damaged by its actions. I propose to show that the allocation of resources will be the same in this case as it was when the damaging business was liable for damage caused. As I showed in the previous case that the allocation of resources was optimal, it will not be necessary to repeat this part of the argument.

I return to the case of the farmer and the cattle-raiser. The farmer would suffer increased damage to his crop as the size of the herd increased. Suppose that the size of the cattle-raiser's herd is 3 steers (and that this is the size of the herd that would be maintained if crop damage was not taken into account). Then the farmer would be willing to pay up to $3 if the cattle-

raiser would reduce his herd to 2 steers, up to $5 if the herd were reduced to 1 steer and would pay up to $6 if cattle-raising was abandoned. The cattle-raiser would therefore receive $3 from the farmer if he kept 2 steers instead of 3. This $3 foregone is therefore part of the cost incurred in keeping the third steer. Whether the $3 is a payment which the cattle-raiser has to make if he adds the third steer to his herd (which it would be if the cattle-raiser was liable to the farmer for damage caused to the crop) or whether it is a sum of money which he would have received if he did not keep a third steer (which it would be if the cattle-raiser was not liable to the farmer for damage caused to the crop) does not affect the final result. In both cases $3 is part of the cost of adding a third steer, to be included along with the other costs. If the increase in the value of production in cattle-raising through increasing the size of the herd from 2 to 3 is greater than the additional costs that have to be incurred (including the $3 damage to crops), the size of the herd will be increased. Otherwise, it will not. The size of the herd will be the same whether the cattle-raiser is liable for damage caused to the crop or not.

It may be argued that the assumed starting point—a herd of 3 steers—was arbitrary. And this is true. But the farmer would not wish to pay to avoid crop damage which the cattle-raiser would not be able to cause. For example, the maximum annual payment which the farmer could be induced to pay could not exceed $9, the annual cost of fencing. And the farmer would only be willing to pay this sum if it did not reduce his earnings to a level that would cause him to abandon cultivation of this particular tract of land. Furthermore, the farmer would only be willing to pay this amount if he believed that, in the absence of any payment by him, the size of the herd maintained by the cattle raiser would be 4 or more steers. Let us assume that this is the case. Then the farmer would be willing to pay up to $3 if the cattle raiser would reduce his herd to 3 steers, up to $6 if the herd were reduced to 2 steers, up to $8 if one steer only were kept and up to $9 if cattle-raising were abandoned. It will be noticed that the change in the starting point has not altered the amount which would accrue to the cattle-raiser if he reduced the size of his herd by any given amount. It is still true that the cattle-raiser could receive an additional $3 from the farmer if he agreed to reduce his herd from 3 steers to 2 and that the $3 represents the value of the crop that would be destroyed by adding the third steer to the herd. Although a different belief on the part of the farmer (whether justified or not) about the size of the herd that the cattle-raiser would maintain in the absence of payments from him may affect the total payment he can be induced to pay, it is not true that this different belief would have any effect on the size of the herd that the cattle-raiser will actually keep. This will be the same as it would be if the cattle-raiser had to pay for damage caused by his cattle, since a receipt foregone of a given amount is the equivalent of a payment of the same amount.

It might be thought that it would pay the cattle-raiser to increase his herd

above the size that he would wish to maintain once a bargain had been made, in order to induce the farmer to make a larger total payment. And this may be true. It is similar in nature to the action of the farmer (when the cattle-raiser was liable for damage) in cultivating land on which, as a result of an agreement with the cattle-raiser, planting would subsequently be abandoned (including land which would not be cultivated at all in the absence of cattle-raising). But such manoeuvres are preliminaries to an agreement and do not affect the long-run equilibrium position, which is the same whether or not the cattle-raiser is held responsible for the crop damage brought about by his cattle.

It is necessary to know whether the damaging business is liable or not for damage caused since without the establishment of this initial delimitation of rights there can be no market transactions to transfer and recombine them. But the ultimate result (which maximises the value of production) is independent of the legal position if the pricing system is assumed to work without cost.

V. The Problem Illustrated Anew

The harmful effects of the activities of a business can assume a wide variety of forms. An early English case concerned a building which, by obstructing currents of air, hindered the operation of a windmill.[5] A recent case in Florida concerned a building which cast a shadow on the cabana, swimming pool and sunbathing areas of a neighbouring hotel.[6] The problem of straying cattle and the damaging of crops which was the subject of detailed examination in the two preceding sections, although it may have appeared to be rather a special case, is in fact but one example of a problem which arises in many different guises. To clarify the nature of my argument and to demonstrate its general applicability, I propose to illustrate it anew by reference to four actual cases.

Let us first reconsider the case of *Sturges v. Bridgman*[7] which I used as an illustration of the general problem in my article on "The Federal Communications Commission." In this case, a confectioner (in Wigmore Street) used two mortars and pestles in connection with his business (one had been in operation in the same position for more than 60 years and the other for more than 26 years). A doctor then came to occupy neighbouring premises (in Wimpole Street). The confectioner's machinery caused the doctor no harm until, eight years after he had first occupied the premises, he built a consulting room at the end of his garden right against the confectioner's kitchen. It was then found that the noise and vibration caused by the confectioner's machin-

[5] See Gale on Easements 237–39 (13th ed. M. Bowles 1959).

[6] See Fontainebleu Hotel Corp. v. Forty-Five Twenty-Five, Inc., 114 So. 2d 357 (1959).

[7] 11 Ch. D. 852 (1879).

ery made it difficult for the doctor to use his new consulting room. "In particular . . . the noise prevented him from examining his patients by auscultation[8] for diseases of the chest. He also found it impossible to engage with effect in any occupation which required thought and attention." The doctor therefore brought a legal action to force the confectioner to stop using his machinery. The courts had little difficulty in granting the doctor the injunction he sought. "Individual cases of hardship may occur in the strict carrying out of the principle upon which we found our judgment, but the negation of the principle would lead even more to individual hardship, and would at the same time produce a prejudicial effect upon the development of land for residential purposes."

The court's decision established that the doctor had the right to prevent the confectioner from using his machinery. But, of course, it would have been possible to modify the arrangements envisaged in the legal ruling by means of a bargain between the parties. The doctor would have been willing to waive his right and allow the machinery to continue in operation if the confectioner would have paid him a sum of money which was greater than the loss of income which he would suffer from having to move to a more costly or less convenient location or from having to curtail his activities at this location or, as was suggested as a possibility, from having to build a separate wall which would deaden the noise and vibration. The confectioner would have been willing to do this if the amount he would have to pay the doctor was less than the fall in income he would suffer if he had to change his mode of operation at this location, abandon his operation or move his confectionery business to some other location. The solution of the problem depends essentially on whether the continued use of the machinery adds more to the confectioner's income than it subtracts from the doctor's.[9] But now consider the situation if the confectioner had won the case. The confectioner would then have had the right to continue operating his noise and vibration-generating machinery without having to pay anything to the doctor. The boot would have been on the other foot: the doctor would have had to pay the confectioner to induce him to stop using the machinery. If the doctor's income would have fallen more through continuance of the use of this machinery than it added to the income of the confectioner, there would clearly be room for a bargain whereby the doctor paid the confectioner to stop using the machinery. That is to say, the circumstances in which it would not pay the confectioner to continue to use the machinery and to compensate the doctor for the losses that this would bring (if the doctor had the right to prevent the confectioner's using his

[8] Auscultation is the act of listening by ear or stethoscope in order to judge by sound the condition of the body.

[9] Note that what is taken into account is the change in income after allowing for alterations in methods of production, location, character of product, etc.

machinery) would be those in which it would be in the interest of the doctor to make a payment to the confectioner which would induce him to discontinue the use of the machinery (if the confectioner had the right to operate the machinery). The basic conditions are exactly the same in this case as they were in the example of the cattle which destroyed crops. With costless market transactions, the decision of the courts concerning liability for damage would be without effect on the allocation of resources. It was of course the view of the judges that they were affecting the working of the economic system—and in a desirable direction. Any other decision would have had "a prejudicial effect upon the development of land for residential purposes," an argument which was elaborated by examining the example of a forge operating on a barren moor, which was later developed for residual purposes. The judges' view that they were settling how the land was to be used would be true only in the case in which the costs of carrying out the necessary market transactions exceeded the gain which might be achieved by any rearrangement of rights. And it would be desirable to preserve the areas (Wimpole Street or the moor) for residential or professional use (by giving non-industrial users the right to stop the noise, vibration, smoke, etc., by injunction) only if the value of the additional residential facilities obtained was greater than the value of cakes or iron lost. But of this the judges seem to have been unaware.

Another example of the same problem is furnished by the case of *Cooke v. Forbes*.[10] One process in the weaving of cocoa-nut fibre matting was to immerse it in bleaching liquids after which it was hung out to dry. Fumes from a manufacturer of sulphate of ammonia had the effect of turning the matting from a bright to a dull and blackish colour. The reason for this was that the bleaching liquid contained chloride of tin, which, when affected by sulphuretted hydrogen, is turned to a darker colour. An injunction was sought to stop the manufacturer from emitting the fumes. The lawyers for the defendant argued that if the plaintiff "were not to use . . . a particular bleaching liquid, their fibre would not be affected; that their process is unusual, not according to the custom of the trade, and even damaging to their own fabrics." The judge commented: ". . . it appears to me quite plain that a person has a right to carry on upon his own property a manufacturing process in which he uses chloride of tin, or any sort of metallic dye, and that his neighbour is not at liberty to pour in gas which will interfere with his manufacture. If it can be traced to the neighbour, then, I apprehend, clearly he will have a right to come here and ask for relief." But in view of the fact that the damage was accidental and occasional, that careful precautions were taken and that there was no exceptional risk, an injunction was refused, leaving the plaintiff to bring an action for damages if he wished. What the subsequent developments

[10] L. R. 5 Eq. 166 (1867–1868).

were I do not know. But it is clear that the situation is essentially the same as that found in *Sturges v. Bridgman,* except that the cocoa-nut fibre matting manufacturer could not secure an injunction but would have to seek damages from the sulphate of ammonia manufacturer. The economic analysis of the situation is exactly the same as with the cattle which destroyed crops. To avoid the damage, the sulphate of ammonia manufacturer could increase his precautions or move to another location. Either course would presumably increase his costs. Alternatively he could pay for the damage. This he would do if the payments for damage were less than the additional costs that would have to be incurred to avoid the damage. The payments for damage would then become part of the cost of production of sulphate of ammonia. Of course, if, as was suggested in the legal proceedings, the amount of damage could be eliminated by changing the bleaching agent (which would presumably increase the costs of the matting manufacturer) and if the additional cost was less than the damage that would otherwise occur, it should be possible for the two manufacturers to make a mutually satisfactory bargain whereby the new bleaching agent was used. Had the court decided against the matting manufacturer, as a consequence of which he would have had to suffer the damage without compensation, the allocation of resources would not have been affected. It would pay the matting manufacturer to change his bleaching agent if the additional cost involved was less than the reduction in damage. And since the matting manufacturer would be willing to pay the sulphate of ammonia manufacturer an amount up to his loss of income (the increase in costs or the damage suffered) if he would cease his activities, this loss of income would remain a cost of production for the manufacturer of sulphate of ammonia. This case is indeed analytically exactly the same as the cattle example.

Bryant v. Lefever[11] raised the problem of the smoke nuisance in a novel form. The plaintiff and the defendants were occupiers of adjoining houses, which were of about the same height.

> Before 1876 the plaintiff was able to light a fire in any room of his house without the chimneys smoking; the two houses had remained in the same condition some thirty or forty years. In 1876 the defendants took down their house, and began to rebuild it. They carried up a wall by the side of the plaintiff's chimneys much beyond its original height, and stacked timber on the roof of their house, and thereby caused the plaintiff's chimneys to smoke whenever he lighted fires.

The reason, of course, why the chimneys smoked was that the erection of the wall and the stacking of the timber prevented the free circulation of air. In a trial before a jury, the plaintiff was awarded damages of £40. The case then went to the Court of Appeals where the judgment was reversed. Bramwell, L.J., argued:

[11] 4 C.P.D. 172 (1878–1879).

. . . it is said, and the jury have found, that the defendants have done that which caused a nuisance to the plaintiff's house. We think there is no evidence of this. No doubt there is a nuisance, but it is not of the defendant's causing. They have done nothing in causing the nuisance. Their house and their timber are harmless enough. It is the plaintiff who causes the nuisance by lighting a coal fire in a place the chimney of which is placed so near the defendants' wall, that the smoke does not escape, but comes into the house. Let the plaintiff cease to light his fire, let him move his chimney, let him carry it higher, and there would be no nuisance. Who then, causes it? It would be very clear that the plaintiff did, if he had built his house or chimney after the defendants had put up the timber on theirs, and it is really the same though he did so before the timber was there. But (what is in truth the same answer), if the defendants cause the nuisance, they have a right to do so. If the plaintiff has not the right to the passage of air, except subject to the defendants' right to build or put timber on their house, then his right is subject to their right, and though a nuisance follows from the exercise of their right, they are not liable.

And Cotton, L.J., said:

Here it is found that the erection of the defendants' wall has sensibly and materially interfered with the comfort of human existence in the plaintiff's house, and it is said this is a nuisance for which the defendants are liable. Ordinarily this is so, but the defendants have done so, not by sending on to the plaintiff's property any smoke or noxious vapour, but by interrupting the egress of smoke from the plaintiff's house in a way to which . . . the plaintiff has no legal right. The plaintiff creates the smoke, which interferes with his comfort. Unless he has . . . a right to get rid of this in a particular way which has been interfered with by the defendants, he cannot sue the defendants, because the smoke made by himself, for which he has not provided any effectual means of escape, causes him annoyance. It is as if a man tried to get rid of liquid filth arising on his own land by a drain into his neighbour's land. Until a right had been acquired by user, the neighbour might stop the drain without incurring liability by so doing. No doubt great inconvenience would be caused to the owner of the property on which the liquid filth arises. But the act of his neighbour would be a lawful act, and he would not be liable for the consequences attributable to the fact that the man had accumulated filth without providing any effectual means of getting rid of it.

I do not propose to show that any subsequent modification of the situation, as a result of bargains between the parties (conditioned by the cost of stacking the timber elsewhere, the cost of extending the chimney higher, etc.), would have exactly the same result whatever decision the courts had come to since this point has already been adequately dealt with in the discussion of the cattle example and the two previous cases. What I shall discuss is the argument of the judges in the Court of Appeals that the smoke nuisance was not caused by the man who erected the wall but by the man who lit the fires. The novelty of the situation is that the smoke nuisance was suffered by the man who lit the fires and not by some third person. The question is not a trivial

one since it lies at the heart of the problem under discussion. Who caused the smoke nuisance? The answer seems fairly clear. The smoke nuisance was caused both by the man who built the wall *and* by the man who lit the fires. Given the fires, there would have been no smoke nuisance without the wall; given the wall, there would have been no smoke nuisance without the fires. Eliminate the wall *or* the fires and the smoke nuisance would disappear. On the marginal principle it is clear that *both* were responsible and *both* should be forced to include the loss of amenity due to the smoke as a cost in deciding whether to continue the activity which gives rise to the smoke. And given the possibility of market transactions, this is what would in fact happen. Although the wall-builder was not liable legally for the nuisance, as the man with the smoking chimneys would presumably be willing to pay a sum equal to the monetary worth to him of eliminating the smoke, this sum would therefore become for the wall-builder, a cost of continuing to have the high wall with the timber stacked on the roof.

The judges' contention that it was the man who lit the fires who alone caused the smoke nuisance is true only if we assume that the wall is the given factor. This is what the judges did by deciding that the man who erected the higher wall had a legal right to do so. The case would have been even more interesting if the smoke from the chimneys had injured the timber. Then it would have been the wall-builder who suffered the damage. The case would then have closely paralleled *Sturges v. Bridgman* and there can be little doubt that the man who lit the fires would have been liable for the ensuing damage to the timber, in spite of the fact that no damage had occurred until the high wall was built by the man who owned the timber.

Judges have to decide on legal liability but this should not confuse economists about the nature of the economic problem involved. In the case of the cattle and the crops, it is true that there would be no crop damage without the cattle. It is equally true that there would be no crop damage without the crops. The doctor's work would not have been disturbed if the confectioner had not worked his machinery; but the machinery would have disturbed no one if the doctor had not set up his consulting room in that particular place. The matting was blackened by the fumes from the sulphate of ammonia manufacturer; but no damage would have occurred if the matting manufacturer had not chosen to hang out his matting in a particular place and to use a particular bleaching agent. If we are to discuss the problem in terms of causation, both parties cause the damage. If we are to attain an optimum allocation of resources, it is therefore desirable that both parties should take the harmful effect (the nuisance) into account in deciding on their course of action. It is one of the beauties of a smoothly operating pricing system that, as has already been explained, the fall in the value of production due to the harmful effect would be a cost for both parties.

Bass v. Gregory[12] will serve as an excellent final illustration of the problem. The plaintiffs were the owners and tenant of a public house called the Jolly Anglers. The defendant was the owner of some cottages and a yard adjoining the Jolly Anglers. Under the public house was a cellar excavated in the rock. From the cellar, a hole or shaft had been cut into an old well situated in the defendant's yard. The well therefore became the ventilating shaft for the cellar. The cellar "had been used for a particular purpose in the process of brewing, which, without ventilation, could not be carried on." The cause of the action was that the defendant removed a grating from the mouth of the well, "so as to stop or prevent the free passage of air from [the] cellar upwards through the well. . . ." What caused the defendant to take this step is not clear from the report of the case. Perhaps "the air . . . impregnated by the brewing operations" which "passed up the well and out into the open air" was offensive to him. At any rate, he preferred to have the well in his yard stopped up. The court had first to determine whether the owners of the public house could have a legal right to a current of air. If they were to have such a right, this case would have to be distinguished from *Bryant v. Lefever* (already considered). This, however, presented no difficulty. In this case, the current of air was confined to "a strictly defined channel." In the case of *Bryant v. Lefever,* what was involved was "the general current of air common to all mankind." The judge therefore held that the owners of the public house could have the right to a current of air whereas the owner of the private house in *Bryant v. Lefever* could not. An economist might be tempted to add "but the air moved all the same." However, all that had been decided at this stage of the argument was that there could be a legal right, not that the owners of the public house possessed it. But evidence showed that the shaft from the cellar to the well had existed for over forty years and that the use of the well as a ventilating shaft must have been known to the owners of the yard since the air, when it emerged, smelt of the brewing operations. The judge therefore held that the public house had such a right by the "doctrine of lost grant." This doctrine states "that if a legal right is proved to have existed and been exercised for a number of years the law ought to presume that it had a legal origin."[13] So the owner of the cottages and yard had to unstop the well and endure the smell.

[12] 25 Q.B.D. 481 (1890).

[13] It may be asked why a lost grant could not also be presumed in the case of the confectioner who had operated one mortar for more than 60 years. The answer is that until the doctor built the consulting room at the end of his garden there was no nuisance. So the nuisance had not continued for many years. It is true that the confectioner in his affidavit referred to "an invalid lady who occupied the house upon one occasion, about thirty years before" who "requested him if possible to discontinue the use of the mortars before eight o'clock in the morning" and that there was some evidence that the garden wall had been subjected to vibration. But the court had little difficulty in disposing of this line of argument: ". . . this vibration, even if it existed at all, was so slight, and the com-

The reasoning employed by the courts in determining legal rights will often seem strange to an economist because many of the factors on which the decision turns are, to an economist, irrelevant. Because of this, situations which are, from an economic point of view, identical will be treated quite differently by the courts. The economic problem in all cases of harmful effects is how to maximise the value of production. In the case of *Bass v. Gregory* fresh air was drawn in through the well which facilitated the production of beer but foul air was expelled through the well which made life in the adjoining houses less pleasant. The economic problem was to decide which to choose: a lower cost of beer and worsened amenities in adjoining houses or a higher cost of beer and improved amenities. In deciding this question, the "doctrine of lost grant" is about as relevant as the colour of the judge's eyes. But it has to be remembered that the immediate question faced by the courts is *not* what shall be done by whom *but* who has the legal right to do what. It is always possible to modify by transactions on the market the initial legal delimitation of rights. And, of course, if such market transactions are costless, such a rearrangement of rights will always take place if it would lead to an increase in the value of production.

VI. The Cost of Market Transactions Taken into Account

The argument has proceeded up to this point on the assumption (explicit in Sections III and IV and tacit in Section V) that there were no costs involved in carrying out market transactions. This is, of course, a very unrealistic assumption. In order to carry out a market transaction it is necessary to discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading up to a bargain, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed, and so on. These operations are often extremely costly, sufficiently costly at any rate to prevent many transactions that would be carried out in a world in which the pricing system worked without cost.

In earlier sections, when dealing with the problem of the rearrangement of legal rights through the market, it was argued that such a rearrangement would be made through the market whenever this would lead to an increase in the value of production. But this assumed costless market transactions. Once the costs of carrying out market transactions are taken into account it is clear that such a rearrangement of rights will only be undertaken when the increase in the value of production consequent upon the rearrangement

---

plaint, if it can be called a complaint, of the invalid lady . . . was of so trifling a character, that . . . the Defendant's acts would not have given rise to any proceeding either at law or in equity" (11 Ch.D. 863). That is, the confectioner had not committed a nuisance until the doctor built his consulting room.

is greater than the costs which would be involved in bringing it about. When it is less, the granting of an injunction (or the knowledge that it would be granted) or the liability to pay damages may result in an activity being discontinued (or may prevent its being started) which would be undertaken if market transactions were costless. In these conditions the initial delimitation of legal rights does have an effect on the efficiency with which the economic system operates. One arrangement of rights may bring about a greater value of production than any other. But unless this is the arrangement of rights established by the legal system, the costs of reaching the same result by altering and combining rights through the market may be so great that this optimal arrangement of rights, and the greater value of production which it would bring, may never be achieved. The part played by economic considerations in the process of delimiting legal rights will be discussed in the next section. In this section, I will take the initial delimitation of rights and the costs of carrying out market transactions as given.

It is clear that an alternative form of economic organisation which could achieve the same result at less cost than would be incurred by using the market would enable the value of production to be raised. As I explained many years ago, the firm represents such an alternative to organising production through market transactions.[14] Within the firm individual bargains between the various cooperating factors of production are eliminated and for a market transaction is substituted an administrative decision. The rearrangement of production then takes place without the need for bargains between the owners of the factors of production. A landowner who has control of a large tract of land may devote his land to various uses taking into account the effect that the interrelations of the various activities will have on the net return of the land, thus rendering unnecessary bargains between those undertaking the various activities. Owners of a large building or of several adjoining properties in a given area may act in much the same way. In effect, using our earlier terminology, the firm would acquire the legal rights of all the parties and the rearrangement of activities would not follow on a rearrangement of rights by contract, but as a result of an administrative decision as to how the rights should be used.

It does not, of course, follow that the administrative costs of organising a transaction through a firm are inevitably less than the costs of the market transactions which are superseded. But where contracts are peculiarly difficult to draw up and an attempt to describe what the parties have agreed to do or not to do (e.g. the amount and kind of a smell or noise that they may make or will not make) would necessitate a lengthy and highly involved document, and, where, as is probable, a long-term contract would be desir-

[14] See Coase, The Nature of the Firm, 4 Economica, New Series, 386 (1937). Reprinted in Readings in Price Theory, 331 (1952).

able;[15] it would be hardly surprising if the emergence of a firm or the extension of the activities of an existing firm was not the solution adopted on many occasions to deal with the problem of harmful effects. This solution would be adopted whenever the administrative costs of the firm were less than the costs of the market transactions that it supersedes and the gains which would result from the rearrangement of activities greater than the firm's costs of organising them. I do not need to examine in great detail the character of this solution since I have explained what is involved in my earlier article.

But the firm is not the only possible answer to this problem. The administrative costs of organising transactions within the firm may also be high, and particularly so when many diverse activities are brought within the control of a single organisation. In the standard case of a smoke nuisance, which may affect a vast number of people engaged in a wide variety of activities, the administrative costs might well be so high as to make any attempt to deal with the problem within the confines of a single firm impossible. An alternative solution is direct Government regulation. Instead of instituting a legal system of rights which can be modified by transactions on the market, the government may impose regulations which state what people must or must not do and which have to be obeyed. Thus, the government (by statute or perhaps more likely through an administrative agency) may, to deal with the problem of smoke nuisance, decree that certain methods of production should or should not be used (e.g. that smoke preventing devices should be installed or that coal or oil should not be burned) or may confine certain types of business to certain districts (zoning regulations).

The government is, in a sense, a super-firm (but of a very special kind) since it is able to influence the use of factors of production by administrative decision. But the ordinary firm is subject to checks in its operations because of the competition of other firms, which might administer the same activities at lower cost and also because there is always the alternative of market transactions as against organisation within the firm if the administrative costs become too great. The government is able, if it wishes, to avoid the market altogether, which a firm can never do. The firm has to make market agreements with the owners of the factors of production that it uses. Just as the government can conscript or seize property, so it can decree that factors of production should only be used in such-and-such a way. Such authoritarian methods save a lot of trouble (for those doing the organising). Furthermore, the government has at its disposal the police and the other law enforcement agencies to make sure that its regulations are carried out.

It is clear that the government has powers which might enable it to get some things done at a lower cost than could a private organisation (or at any

[15] For reasons explained in my earlier article, see Readings in Price Theory, n. 14 at 337.

rate one without special governmental powers). But the governmental administrative machine is not itself costless. It can, in fact, on occasion be extremely costly. Furthermore, there is no reason to suppose that the restrictive and zoning regulations, made by a fallible administration subject to political pressures and operating without any competitive check, will necessarily always be those which increase the efficiency with which the economic system operates. Furthermore, such general regulations which must apply to a wide variety of cases will be enforced in some cases in which they are clearly inappropriate. From these considerations it follows that direct governmental regulation will not necessarily give better results than leaving the problem to be solved by the market or the firm. But equally there is no reason why, on occasion, such governmental administrative regulation should not lead to an improvement in economic efficiency. This would seem particularly likely when, as is normally the case with the smoke nuisance, a large number of people are involved and in which therefore the costs of handling the problem through the market or the firm may be high.

There is, of course, a further alternative, which is to do nothing about the problem at all. And given that the costs involved in solving the problem by regulations issued by the governmental administrative machine will often be heavy (particularly if the costs are interpreted to include all the consequences which follow from the Government engaging in this kind of activity), it will no doubt be commonly the case that the gain which would come from regulating the actions which give rise to the harmful effects will be less than the costs involved in Government regulation.

The discussion of the problem of harmful effects in this section (when the costs of market transactions are taken into account) is extremely inadequate. But at least it has made clear that the problem is one of choosing the appropriate social arrangement for dealing with the harmful effects. All solutions have costs and there is no reason to suppose that government regulation is called for simply because the problem is not well handled by the market or the firm. Satisfactory views on policy can only come from a patient study of how, in practice, the market, firms and governments handle the problem of harmful effects. Economists need to study the work of the broker in bringing parties together, the effectiveness of restrictive covenants, the problems of the large-scale real-estate development company, the operation of Government zoning and other regulating activities. It is my belief that economists, and policy-makers generally, have tended to over-estimate the advantages which come from governmental regulation. But this belief, even if justified, does not do more than suggest that government regulation should be curtailed. It does not tell us where the boundary line should be drawn. This, it seems to me, has to come from a detailed investigation of the actual results

of handling the problem in different ways. But it would be unfortunate if this investigation were undertaken with the aid of a faulty economic analysis. The aim of this article is to indicate what the economic approach to the problem should be.

VII. The Legal Delimitation of Rights and the Economic Problem

The discussion in Section V not only served to illustrate the argument but also afforded a glimpse at the legal approach to the problem of harmful effects. The cases considered were all English but a similar selection of American cases could easily be made and the character of the reasoning would have been the same. Of course, if market transactions were costless, all that matters (questions of equity apart) is that the rights of the various parties should be well-defined and the results of legal actions easy to forecast. But as we have seen, the situation is quite different when market transactions are so costly as to make it difficult to change the arrangement of rights established by the law. In such cases, the courts directly influence economic activity. It would therefore seem desirable that the courts should understand the economic consequences of their decisions and should, insofar as this is possible without creating too much uncertainty about the legal position itself, take these consequences into account when making their decisions. Even when it is possible to change the legal delimitation of rights through market transactions, it is obviously desirable to reduce the need for such transactions and thus reduce the employment of resources in carrying them out.

A thorough examination of the presuppositions of the courts in trying such cases would be of great interest but I have not been able to attempt it. Nevertheless it is clear from a cursory study that the courts have often recognized the economic implications of their decisions and are aware (as many economists are not) of the reciprocal nature of the problem. Furthermore, from time to time, they take these economic implications into account, along with other factors, in arriving at their decisions. The American writers on this subject refer to the question in a more explicit fashion than do the British. Thus, to quote Prosser on Torts, a person may

> make use of his own property or . . . conduct his own affairs at the expense of some harm to his neighbors. He may operate a factory whose noise and smoke cause some discomfort to others, so long as he keeps within reasonable bounds. It is only when his conduct is unreasonable, *in the light of its utility and the harm which results* [italics added], that it becomes a nuisance. . . . . As it was said in an ancient case in regard to candle-making in a town, "Le utility del chose excusera le noisomeness del stink."
>
> The world must have factories, smelters, oil refineries, noisy machinery and blasting, even at the expense of some inconvenience to those in the vicinity and the

plaintiff may be required to accept some not unreasonable discomfort for the general good.[16]

The standard British writers do not state as explicitly as this that a comparison between the utility and harm produced is an element in deciding whether a harmful effect should be considered a nuisance. But similar views, if less strongly expressed, are to be found.[17] The doctrine that the harmful effect must be substantial before the court will act is, no doubt, in part a reflection of the fact that there will almost always be some gain to offset the harm. And in the reports of individual cases, it is clear that the judges have had in mind what would be lost as well as what would be gained in deciding whether to grant an injunction or award damages. Thus, in refusing to prevent the destruction of a prospect by a new building, the judge stated:

> I know no general rule of common law, which . . . says, that building so as to stop another's prospect is a nuisance. Was that the case, there could be no great towns; and I must grant injunctions to all the new buildings in this town. . . .[18]

In *Webb v. Bird*[19] it was decided that it was not a nuisance to build a schoolhouse so near a windmill as to obstruct currents of air and hinder the working of the mill. An early case seems to have been decided in an opposite direction. Gale commented:

> In old maps of London a row of windmills appears on the heights to the north of London. Probably in the time of King James it was thought an alarming circumstance, as affecting the supply of food to the city, that anyone should build so near them as to take the wind out from their sails.[20]

In one of the cases discussed in section V, *Sturges v. Bridgman,* it seems clear that the judges were thinking of the economic consequences of alternative decisions. To the argument that if the principle that they seemed to be following

[16] See W. L. Prosser, The Law of Torts 398–99, 412 (2d ed. 1955). The quotation about the ancient case concerning candle-making is taken from Sir James Fitzjames Stephen, A General View of the Criminal Law of England 106 (1890). Sir James Stephen gives no reference. He perhaps had in mind *Rex. v. Ronkett,* included in Seavey, Keeton and Thurston, Cases on Torts 604 (1950). A similar view to that expressed by Prosser is to be found in F. V. Harper and F. James, The Law of Torts 67–74 (1956); Restatement, Torts §§826, 827 and 828.

[17] See Winfield on Torts 541–48 (6th ed. T. E. Lewis 1954); Salmond on the Law of Torts 181–90 (12th ed. R.F.V. Heuston 1957); H. Street, The Law of Torts 221–29 (1959).

[18] Attorney General v. Doughty, 2 Ves. Sen. 453, 28 Eng. Rep. 290 (Ch. 1752). Compare in this connection the statement of an American judge, quoted in Prosser, op. cit. supra n. 16 at 413 n. 54: "Without smoke, Pittsburgh would have remained a very pretty village," Musmanno, J., in Versailles Borough v. McKeesport Coal & Coke Co., 1935, 83 Pitts. Leg. J. 379, 385.

[19] 10 C.B. (N.S.) 268, 142 Eng. Rep. 445 (1861); 13 C.B. (N.S.) 841, 143 Eng. Rep. 332 (1863).

[20] See Gale on Easements 238, n. 6 (13th ed. M. Bowles 1959).

were carried out to its logical consequences, it would result in the most serious practical inconveniences, for a man might go—say into the midst of the tanneries of *Bermondsey,* or into any other locality devoted to any particular trade or manufacture of a noisy or unsavoury character, and by building a private residence upon a vacant piece of land put a stop to such trade or manufacture altogether,

the judges answered that

whether anything is a nuisance or not is a question to be determined, not merely by an abstract consideration of the thing itself, but in reference to its circumstances; What would be a nuisance in *Belgrave Square* would not necessarily be so in *Bermondsey;* and where a locality is devoted to a particular trade or manufacture carried on by the traders or manufacturers in a particular and established manner not constituting a public nuisance, Judges and juries would be justified in finding, and may be trusted to find, that the trade or manufacture so carried on in that locality is not a private or actionable wrong.[21]

That the character of the neighborhood is relevant in deciding whether something is, or is not, a nuisance, is definitely established.

He who dislikes the noise of traffic must not set up his abode in the heart of a great city. He who loves peace and quiet must not live in a locality devoted to the business of making boilers or steamships.[22]

What has emerged has been described as "planning and zoning by the judiciary."[23] Of course there are sometimes considerable difficulties in applying the criteria.[24]

An interesting example of the problem is found in *Adams v. Ursell*[25] in which a fried fish shop in a predominantly working-class district was set up near houses of "a much better character." England without fish-and-chips is a contradiction in terms and the case was clearly one of high importance. The judge commented:

It was urged that an injunction would cause great hardship to the defendant and to the poor people who get food at his shop. The answer to that is that it does not follow that the defendant cannot carry on his business in another more suitable place somewhere in the neighbourhood. It by no means follows that because a fried fish shop is a nuisance in one place it is a nuisance in another.

In fact, the injunction which restrained Mr. Ursell from running his shop did not even extend to the whole street. So he was presumably able to move to other premises near houses of "a much worse character," the inhabitants

[21] 11 Ch.D. 865 (1879).

[22] Salmond on the Law of Torts 182 (12th ed. R.F.V. Heuston 1957).

[23] C. M. Haar, Land-Use Planning, A Casebook on the Use, Misuse, and Re-use of Urban Land 95 (1959).

[24] See, for example, Rushmer v. Polsue and Alfieri, Ltd. [1906] 1 Ch. 234, which deals with the case of a house in a quiet situation in a noisy district.

[25] [1913] 1 Ch. 269.

of which would no doubt consider the availability of fish-and-chips to outweigh the pervading odour and "fog or mist" so graphically described by the plaintiff. Had there been no other "more suitable place in the neighbourhood," the case would have been more difficult and the decision might have been different. What would "the poor people" have had for food? No English judge would have said: "Let them eat cake."

The courts do not always refer very clearly to the economic problem posed by the cases brought before them but it seems probable that in the interpretation of words and phrases like "reasonable" or "common or ordinary use" there is some recognition, perhaps largely unconscious and certainly not very explicit, of the economic aspects of the questions at issue. A good example of this would seem to be the judgment in the Court of Appeals in *Andreae v. Selfridge and Company Ltd.*[26] In this case, a hotel (in Wigmore Street) was situated on part of an island site. The remainder of the site was acquired by Selfridges which demolished the existing buildings in order to erect another in their place. The hotel suffered a loss of custom in consequence of the noise and dust caused by the demolition. The owner of the hotel brought an action against Selfridges for damages. In the lower court, the hotel was awarded £4,500 damages. The case was then taken on appeal.

The judge who had found for the hotel proprietor in the lower court said:

> I cannot regard what the defendants did on the site of the first operation as having been commonly done in the ordinary use and occupation of land or houses. It is neither usual nor common, in this country, for people to excavate a site to a depth of 60 feet and then to erect upon that site a steel framework and fasten the steel frames together with rivets. . . . Nor is it, I think, a common or ordinary use of land, in this country, to act as the defendants did when they were dealing with the site of their second operation—namely, to demolish all the houses that they had to demolish, five or six of them I think, if not more, and to use for the purpose of demolishing them pneumatic hammers.

Sir Wilfred Greene, M.R., speaking for the Court of Appeals, first noted

> that when one is dealing with temporary operations, such as demolition and re-building, everybody has to put up with a certain amount of discomfort, because operations of that kind cannot be carried on at all without a certain amount of noise and a certain amount of dust. Therefore, the rule with regard to interference must be read subject to this qualification. . . .

He then referred to the previous judgment:

> With great respect to the learned judge, I take the view that he has not approached this matter from the correct angle. It seems to me that it is not possible to say . . . that the type of demolition, excavation and construction in which the defendant company was engaged in the course of these operations was of such an abnormal and unusual nature as to prevent the qualification to which I have referred coming

[26] [1938] 1 Ch. 1.

into operation. It seems to me that, when the rule speaks of the common or ordinary use of land, it does not mean that the methods of using land and building on it are in some way to be stabilised for ever. As time goes on new inventions or new methods enable land to be more profitably used, either by digging down into the earth or by mounting up into the skies. Whether, from other points of view, that is a matter which is desirable for humanity is neither here nor there; but it is part of the normal use of land, to make use upon your land, in the matter of construction, of what particular type and what particular depth of foundations and particular height of building may be reasonable, in the circumstances, and in view of the developments of the day. . . . Guests at hotels are very easily upset. People coming to this hotel, who were accustomed to a quiet outlook at the back, coming back and finding demolition and building going on, may very well have taken the view that the particular merit of this hotel no longer existed. That would be a misfortune for the plaintiff; but assuming that there was nothing wrong in the defendant company's works, assuming the defendant company was carrying on the demolition and its building, productive of noise though it might be, with all reasonable skill, and taking all reasonable precautions not to cause annoyance to its neighbors, then the planitiff might lose all her clients in the hotel because they have lost the amenities of an open and quiet place behind, but she would have no cause of complaint. . . . [But those] who say that their interference with the comfort of their neighbors is justified because their operations are normal and usual and conducted with proper care and skill are under a specific duty . . . to use that reasonable and proper care and skill. It is not a correct attitude to take to say: 'We will go on and do what we like until somebody complains!' . . . Their duty is to take proper precautions and to see that the nuisance is reduced to a minimum. It is no answer for them to say: 'But this would mean that we should have to do the work more slowly than we would like to do it, or it would involve putting us to some extra expense.' All these questions are matters of common sense and degree, and quite clearly it would be unreasonable to expect people to conduct their work so slowly or so expensively, for the purpose of preventing a transient inconvenience, that the cost and trouble would be prohibitive. . . . In this case, the defendant company's attitude seems to have been to go on until somebody complained, and, further, that its desire to hurry its work and conduct it according to its own ideas and its own convenience was to prevail if there was a real conflict between it and the comfort of its neighbors. That . . . is not carrying out the obligation of using reasonable care and skill. . . . The effect comes to this . . . the plaintiff suffered an actionable nuisance; . . . she is entitled, not to a nominal sum, but to a substantial sum, based upon those principles . . . but in arriving at the sum . . . I have discounted any loss of custom . . . which might be due to the general loss of amenities owing to what was going on at the back. . . .

The upshot was that the damages awarded were reduced from £4,500 to £1,000.

The discussion in this section has, up to this point, been concerned with court decisions arising out of the common law relating to nuisance. Delimitation of rights in this area also comes about because of statutory enactments. Most economists would appear to assume that the aim of governmental

action in this field is to extend the scope of the law of nuisance by designating as nuisances activities which would not be recognized as such by the common law. And there can be no doubt that some statutes, for example, the Public Health Acts, have had this effect. But not all Government enactments are of this kind. The effect of much of the legislation in this area is to protect businesses from the claims of those they have harmed by their actions. There is a long list of legalized nuisances.

The position has been summarized in *Halsbury's Laws of England* as follows:

> Where the legislature directs that a thing shall in all events be done or authorises certain works at a particular place for a specific purposes or grants powers with the intention that they shall be exercised, although leaving some discretion as to the mode of exercise, no action will lie at common law for nuisance or damage which is the inevitable result of carrying out the statutory powers so conferred. This is so whether the act causing the damage is authorised for public purposes or private profit. Acts done under powers granted by persons to whom Parliament has delegated authority to grant such powers, for example, under provisional orders of the Board of Trade, are regarded as having been done under statutory authority. In the absence of negligence it seems that a body exercising statutory powers will not be liable to an action merely because it might, by acting in a different way, have minimised an injury.

Instances are next given of freedom from liability for acts authorized:

> An action has been held not to be against a body exercising its statutory powers without negligence in respect of the flooding of land by water escaping from watercourses, from water pipes, from drains, or from a canal; the escape of fumes from sewers; the escape of sewage: the subsidence of a road over a sewer; vibration or noise caused by a railway; fires caused by authorised acts; the pollution of a stream where statutory requirements to use the best known method of purifying before discharging the effluent have been satisfied; interference with a telephone or telegraph system by an elctric tramway; the insertion of poles for tramways in the subsoil; annoyance caused by things reasonably necessary for the excavation of authorised works; accidental damage caused by the placing of a grating in a roadway; the escape of tar acid; or interference with the access of a frontager by a street shelter or safety railings on the edge of a pavement.[27]

The legal position in the United States would seem to be essentially the same as in England, except that the power of the legislatures to authorize what would otherwise be nuisances under the common law, at least without giving compensation to the person harmed, is somewhat more limited, as it is subject to constitutional restrictions.[28] Nonetheless, the power is there and cases more or less identical with the English cases can be found. The

[27] See 30 Halsbury, Law of England 690–91 (3d ed. 1960), Article on Public Authorities and Public Officers.

[28] See Prosser, op. cit. supra n. 16 at 421; Harper and James, op. cit. supra n. 16 at 86–87.

question has arisen in an acute form in connection with airports and the operation of aeroplanes. The case of *Delta Air Corporation v. Kersey, Kersey v. City of Atlanta*[29] is a good example. Mr. Kersey bought land and built a house on it. Some years later the City of Atlanta constructed an airport on land immediately adjoining that of Mr. Kersey. It was explained that his property was "a quiet, peaceful and proper location for a home before the airport was built, but dust, noises and low flying of airplanes caused by the operation of the airport have rendered his property unsuitable as a home," a state of affairs which was described in the report of the case with a wealth of distressing detail. The judge first referred to an earlier case, *Thrasher v. City of Atlanta*[30] in which it was noted that the City of Atlanta had been expressly authorized to operate an airport.

> By this franchise aviation was recognised as a lawful business and also as an enterprise affected with a public interest . . . all persons using [the airport] in the manner contemplated by law are within the protection and immunity of the franchise granted by the municipality. An airport is not a nuisance per se, although it might become such from the manner of its construction or operation.

Since aviation was a lawful business affected with a public interest and the construction of the airport was autorized by statute, the judge next referred to *Georgia Railroad and Banking Co. v. Maddox*[31] in which it was said:

> Where a railroad terminal yard is located and its construction authorized, under statutory powers, if it be constructed and operated in a proper manner, it cannot be adjudged a nuisance. Accordingly, injuries and inconveniences to persons residing near such a yard, from noises of locomotives, rumbling of cars, vibrations produced thereby, and smoke, cinders, soot and the like, which result from the ordinary and necessary, therefore proper, use and operation of such a yard, are not nuisances, but are the necessary concomitants of the franchise granted.

In view of this, the judge decided that the noise and dust complained of by Mr. Kersey "may be deemed to be incidental to the proper operation of an airport, and as such they cannot be said to constitute a nuisance." But the complaint against low flying was different:

> . . . can it be said that flights . . . at such a low height [25 to 50 feet above Mr. Kersey's house] as to be imminently dangerous to . . . life and health . . . are a necessary concomitant of an airport? We do not think this question can be answered in the affirmative. No reason appears why the city could not obtain lands of an area [sufficiently large] . . . as not to require such low flights. . . . For the sake of public convenience adjoining-property owners must suffer such inconvenience from noise and dust as result from the usual and proper operation of an airport, but their private rights are entitled to preference in the eyes of the law where the inconvenience is not one demanded by a properly constructed and operated airport.

[29] Supreme Court of Georgia. 193 Ga. 862, 20 S.E. 2d 245 (1942).

[30] 178 Ga. 514, 173 S.E. 817 (1934).

[31] 116 Ga. 64, 42 S.E. 315 (1902).

Of course this assumed that the City of Atlanta could prevent the low flying and continue to operate the airport. The judge therefore added:

> From all that appears, the conditions causing the low flying may be remedied; but if on the trial it should appear that it is indispensable to the public interest that the airport should continue to be operated in its present condition, it may be said that the petitioner should be denied injunctive relief.

In the course of another aviation case, *Smith v. New England Aircraft Co.*,[32] the court surveyed the law in the United States regarding the legalizing of nuisances and it is apparent that, in the broad, it is very similar to that found in England:

> It is the proper function of the legislative department of government in the exercise of the police power to consider the problems and risks that arise from the use of new inventions and endeavor to adjust private rights and harmonize conflicting interests by comprehensive statutes for the public welfare. . . . There are . . . analogies where the invasion of the airspace over underlying land by noise, smoke, vibration, dust and disagreeable odors, having been authorized by the legislative department of government and not being in effect a condemnation of the property although in some measure depreciating its market value, must be borne by the landowner without compensation or remedy. Legislative sanction makes that lawful which otherwise might be a nuisance. Examples of this are damages to adjacent land arising from smoke, vibration and noise in the operation of a railroad . . . ; the noise of ringing factory bells . . . ; the abatement of nuisances . . . ; the erection of steam engines and furnaces . . . ; unpleasant odors connected with sewers, oil refining and storage of naphtha. . . .

Most economists seem to be unaware of all this. When they are prevented from sleeping at night by the roar of jet planes overhead (publicly authorized and perhaps publicly operated), are unable to think (or rest) in the day because of the noise and vibration from passing trains (publicly authorized and perhaps publicly operated), find it difficult to breathe because of the odour from a local sewage farm (publicly authorized and perhaps publicly operated) and are unable to escape because their driveways are blocked by a road obstruction (without any doubt, publicly devised), their nerves frayed and mental balance disturbed, they proceed to declaim about the disadvantages of private enterprise and the need for Government regulation.

While most economists seem to be under a misapprehension concerning the character of the situation with which they are dealing, it is also the case that the activities which they would like to see stopped or curtailed may well be socially justified. It is all a question of weighing up the gains that would accrue from eliminating these harmful effects against the gains that accrue from allowing them to continue. Of course, it is likely that an extension of Government economic activity will often lead to this protection against

[32] 270 Mass. 511, 523, 170 N.E. 385, 390 (1930).

action for nuisance being pushed further than is desirable. For one thing, the Government is likely to look with a benevolent eye on enterprises which it is itself promoting. For another, it is possible to describe the committing of a nuisance by public enterprise in a much more pleasant way than when the same thing is done by private enterprise. In the words of Lord Justice Sir Alfred Denning:

> . . . the significance of the social revolution of today is that, whereas in the past the balance was much too heavily in favor of the rights of property and freedom of contract, Parliament has repeatedly intervened so as to give the public good its proper place.[33]

There can be little doubt that the Welfare State is likely to bring an extension of that immunity from liability for damage, which economists have been in the habit of condemning (although they have tended to assume that this immunity was a sign of too little Government intervention in the economic system). For example, in Britain, the powers of local authorities are regarded as being either absolute or conditional. In the first category, the local authority has no discretion in exercising the power conferred on it. "The absolute power may be said to cover all the necessary consequences of its direct operation even if such consequences amount to nuisance." On the other hand, a conditional power may only be exercised in such a way that the consequences do not constitute a nuisance.

> It is the intention of the legislature which determines whether a power is absolute or conditional. . . . [As] there is the possibility that the social policy of the legislature may change from time to time, a power which in one era would be construed as being conditional, might in another era be interpreted as being absolute in order to further the policy of the Welfare State. This point is one which should be borne in mind when considering some of the older cases upon this aspect of the law of nuisance.[34]

It would seem desirable to summarize the burden of this long section. The problem which we face in dealing with actions which have harmful effects is not simply one of restraining those responsible for them. What has to be decided is whether the gain from preventing the harm is greater than the loss which would be suffered elsewhere as a result of stopping the action which produces the harm. In a world in which there are costs of rearranging the rights established by the legal system, the courts, in cases relating to nuisance, are, in effect, making a decision on the economic problem and determining how resources are to be employed. It was argued that the courts are conscious of this and that they often make, although not always in a very explicit fashion, a comparison between what would be gained and what lost by preventing

[33] See Sir Alfred Denning, Freedom Under the Law 71 (1949).

[34] M. B. Cairns, The Law of Tort in Local Government 28–32 (1954).

actions which have harmful effects. But the delimitation of rights is also the result of statutory enactments. Here we also find evidence of an appreciation of the reciprocal nature of the problem. While statutory enactments add to the list of nuisances, action is also taken to legalize what would otherwise be nuisances under the common law. The kind of situation which economists are prone to consider as requiring corrective Government action is, in fact, often the result of Government action. Such action is not necessarily unwise. But there is a real danger that extensive Government intervention in the economic system may lead to the protection of those responsible for harmful effects being carried too far.

VIII. Pigou's Treatment in "The Economics of Welfare"

The fountainhead for the modern economic analysis of the problem discussed in this article is Pigou's *Economics of Welfare* and, in particular, that section of Part II which deals with divergences between social and private net products which come about because

> one person A, in the course of rendering some service, for which payment is made, to a second person B, incidentally also renders services or disservices to other persons (not producers of like services), of such a sort that payment cannot be exacted from the benefited parties or compensation enforced on behalf of the injured parties.[35]

Pigou tells us that his aim in Part II of *The Economics of Welfare* is

> to ascertain how far the free play of self-interest, acting under the existing legal system, tends to distribute the country's resources in the way most favorable to the production of a large national dividend, and how far it is feasible for State action to improve upon 'natural' tendencies.[36]

To judge from the first part of this statement, Pigou's purpose is to discover whether any improvements could be made in the existing arrangements which determine the use of resources. Since Pigou's conclusion is that improvements could be made, one might have expected him to continue by saying that he proposed to set out the changes required to bring them about. Instead, Pigou adds a phrase which contrasts "natural" tendencies with State action, which seems in some sense to equate the present arrangements with "natural" tendencies and to imply that what is required to bring about these improvements is State action (if feasible). That this is more or less Pigou's position is evident from Chapter I of Part II.[37] Pigou starts by referring to "optimistic

[35] A. C. Pigou, The Economics of Welfare 183 (4th ed. 1932). My references will all be to the fourth edition but the argument and examples examined in this article remained substantially unchanged from the first edition in 1920 to the fourth in 1932. A large part (but not all) of this analysis had appeared previously in Wealth and Welfare (1912).

[36] *Id.* at xii.

[37] *Id.* at 127–30.

followers of the classical economists"[38] who have argued that the value of production would be maximised if the Government refrained from any interference in the economic system and the economic arrangements were those which came about "naturally." Pigou goes on to say that if self-interest does promote economic welfare, it is because human institutions have been devised to make it so. (This part of Pigou's argument, which he develops with the aid of a quotation from Cannan, seems to me to be essentially correct.) Pigou concludes:

> But even in the most advanced States there are failures and imperfections. . . . there are many obstacles that prevent a community's resources from being distributed . . . in the most efficient way. The study of these constitutes our present problem. . . . its purposes is essentially practical. It seeks to bring into clearer light some of the ways in which it now is, or eventually may become, feasible for governments to control the play of economic forces in such wise as to promote the economic welfare, and through that, the total welfare, of their citizens as a whole.[39]

Pigou's underlying thought would appear to be: Some have argued that no State action is needed. But the system has performed as well as it has because of State action. Nonetheless, there are still imperfections. What additional State action is required?

If this is a correct summary of Pigou's position, its inadequacy can be demonstrated by examining the first example he gives of a divergence between private and social products.

> It might happen . . . that costs are thrown upon people not directly concerned, through, say, uncompensated damage done to surrounding woods by sparks from railway engines. All such effects must be included—some of them will be positive, others negative elements—in reckoning up the social net product of the marginal increment of any volume of resources turned into any use or place.[40]

The example used by Pigou refers to a real situation. In Britain, a railway does not normally have to compensate those who suffer damage by fire caused by sparks from an engine. Taken in conjunction with what he says in Chapter 9 of Part II, I take Pigou's policy recommendations to be, first, that there should be State action to correct this "natural" situation and, second, that the railways should be forced to compensate those whose woods are burnt. If this is a correct interpretation of Pigou's position, I would argue that the first recommendation is based on a misapprehension of the facts and that the second is not necessarily desirable.

[38] In Wealth and Welfare, Pigou attributes the "optimism" to Adam Smith himself and not to his followers. He there refers to the "highly optimistic theory of Adam Smith that the national dividend, in given circumstances of demand and supply, tends 'naturally' to a maximum" (p. 104).

[39] Pigou, op. cit. supra n. 35 at 129–30.

[40] *Id.* at 134.

Let us consider the legal position. Under the heading "Sparks from engines," we find the following in Halsbury's Laws of England:

> If railway undertakers use steam engines on their railway without express statutory authority to do so, they are liable, irrespective of any negligence on their part, for fires caused by sparks from engines. Railway undertakers are, however, generally given statutory authority to use steam engines on their railway; accordingly, if an engine is constructed with the precautions which science suggests against fire and is used without negligence, they are not responsible at common law for any damage which may be done by sparks. . . . In the construction of an engine the undertaker is bound to use all the discoveries which science has put within its reach in order to avoid doing harm, provided they are such as it is reasonable to require the company to adopt, having proper regard to the likelihood of the damage and to the cost and convenience of the remedy; but it is not negligence on the part of an undertaker if it refuses to use an apparatus the efficiency of which is open to bona fide doubt.

To this general rule, there is a statutory exception arising from the Railway (Fires) Act, 1905, as amended in 1923. This concerns agricultural land or agricultural crops.

> In such a case the fact that the engine was used under statutory powers does not affect the liability of the company in an action for the damage. . . . These provisions, however, only apply where the claim for damage . . . does not exceed £ 200, [£ 100 in the 1905 Act] and where written notice of the occurrence of the fire and the intention to claim has been sent to the company within seven days of the occurrence of the damage and particulars of the damage in writing showing the amount of the claim in money not exceeding £200 have been sent to the company within twenty-one days.

Agricultural land does not include moorland or buildings and agricultural crops do not include those led away or stacked.[41] I have not made a close study of the parliamentary history of this statutory exception, but to judge from debates in the House of Commons in 1922 and 1923, this exception was probably designed to help the smallholder.[42]

Let us return to Pigou's example of uncompensated damage to surrounding woods caused by sparks from railway engines. This is presumably intended to show how it is possible "for State action to improve on 'natural' tendencies." If we treat Pigou's example as referring to the position before 1905, or as being an arbitrary example (in that he might just as well have written "surrounding buildings" instead of "surrounding woods"), then it is clear that the reason why compensation was not paid must have been that the railway had statutory authority to run steam engines (which relieved it of liability for fires caused by sparks). That this was the legal position was

[41] See 31 Halsbury, Laws of England 474–75 (3d ed. 1960), Article on Railways and Canals, from which this summary of the legal position, and all quotations, are taken.

[42] See 152 H.C. Deb. 2622–63 (1922); 161 H.C. Deb. 2935–55 (1923).

established in 1860, in a case, oddly enough, which concerned the burning of surrounding woods by a railway,[43] and the law on this point has not been changed (apart from the one exception) by a century of railway legislation, including nationalisation. If we treat Pigou's example of "uncompensated damage done to surrounding woods by sparks from railway engines" literally, and assume that it refers to the period after 1905, then it is clear that the reason why compensation was not paid must have been that the damage was more than £100 (in the first edition of *The Economics of Welfare*) or more than £200 (in later editions) or that the owner of the wood failed to notify the railway in writing within seven days of the fire or did not send particulars of the damage, in writing, within twenty-one days. In the real world, Pigou's example could only exist as a result of a deliberate choice of the legislature. It is not, of course, easy to imagine the construction of a railway in a state of nature. The nearest one can get to this is presumably a railway which uses steam engines "without express statutory authority." However, in this case the railway would be obliged to compensate those whose woods it burnt down. That is to say, compensation would be paid in the absence of Government action. The only circumstances in which compensation would not be paid would be those in which there had been Government action. It is strange that Pigou, who clearly thought it desirable that compensation should be paid, should have chosen this particular example to demonstrate how it is possible "for State action to improve on 'natural' tendencies."

Pigou seems to have had a faulty view of the facts of the situation. But it also seems likely that he was mistaken in his economic analysis. It is not necessarily desirable that the railway should be required to compensate those who suffer damage by fires caused by railway engines. I need not show here that, if the railway could make a bargain with everyone having property adjoining the railway line and there were no costs involved in making such bargains, it would not matter whether the railway was liable for damage caused by fires or not. This question has been treated at length in earlier sections. The problem is whether it would be desirable to make the railway liable in conditions in which it is too expensive for such bargains to be made. Pigou clearly thought it was desirable to force the railway to pay compensation and it is easy to see the kind of argument that would have led him to this conclusion. Suppose a railway is considering whether to run an additional train or to increase the speed of an existing train or to install spark-preventing devices on its engines. If the railway were not liable for fire damage, then, when making these decisions, it would not take into account as a cost the increase in damage resulting from the additional train or the faster train or the failure to install spark-preventing devices. This is the source of the di-

[43] Vaughan v. Taff Vale Railway Co., 3 H. and N. 743 (Ex. 1858) and 5 H. and N. 679 (Ex. 1860).

vergence between private and social net products. It results in the railway performing acts which will lower the value of total production—and which it would not do if it were liable for the damage. This can be shown by means of an arithmetical example.

Consider a railway, which is *not* liable for damage by fires caused by sparks from its engines, which runs two trains per day on a certain line. Suppose that running one train per day would enable the railway to perform services worth $150 per annum and running two trains a day would enable the railway to perform services worth $250 per annum. Suppose further that the cost of running one train is $50 per annum and two trains $100 per annum. Assuming perfect competition, the cost equals the fall in the value of production elsewhere due to the employment of additional factors of production by the railway. Clearly the railway would find it profitable to run two trains per day. But suppose that running one train per day would destroy by fire crops worth (on an average over the year) $60 and two trains a day would result in the destruction of crops worth $120. In these circumstances running one train per day would raise the value of total production but the running of a second train would reduce the value of total production. The second train would enable additional railway services worth $100 per annum to be performed. But the fall in the value of production elsewhere would be $110 per annum; $50 as a result of the employment of additional factors of production and $60 as a result of the destruction of crops. Since it would be better if the second train were not run and since it would not run if the railway were liable for damage caused to crops, the conclusion that the railway should be made liable for the damage seems irresistable. Undoubtedly it is this kind of reasoning which underlies the Pigovian position.

The conclusion that it would be better if the second train did not run is correct. The conclusion that it is desirable that the railway should be made liable for the damage it causes is wrong. Let us change our assumption concerning the rule of liability. Suppose that the railway is liable for damage from fires caused by sparks from the engine. A farmer on lands adjoining the railway is then in the position that, if his crop is destroyed by fires caused by the railway, he will receive the market price from the railway; but if his crop is not damaged, he will receive the market price by sale. It therefore becomes a matter of indifference to him whether his crop is damaged by fire or not. The position is very different when the railway is *not* liable. Any crop destruction through railway-caused fires would then reduce the receipts of the farmer. He would therefore take out of cultivation any land for which the damage is likely to be greater than the net return of the land (for reasons explained at length in Section III). A change from a regime in which the railway is *not* liable for damage to one in which it *is* liable is likely therefore to lead to an increase in the amount of cultivation on lands adjoining the

railway. It will also, of course, lead to an increase in the amount of crop destruction due to railway-caused fires.

Let us return to our arithmetical example. Assume that, with the changed rule of liability, there is a doubling in the amount of crop destruction due to railway-caused fires. With one train per day, crops worth $120 would be destroyed each year and two trains per day would lead to the destruction of crops worth $240. We saw previously that it would not be profitable to run the second train if the railway had to pay $60 per annum as compensation for damage. With damage at $120 per annum the loss from running the second train would be $60 greater. But now let us consider the first train. The value of the transport services furnished by the first train is $150. The cost of running the train is $50. The amount that the railway would have to pay out as compensation for damage is $120. It follows that it would not be profitable to run any trains. With the figures in our example we reach the following result: if the railway is not liable for fire-damage, two trains per day would be run; if the railway is liable for fire-damage, it would cease operations altogether. Does this mean that it is better that there should be no railway? This question can be resolved by considering what would happen to the value of total production if it were decided to exempt the railway from liability for fire-damage, thus bringing it into operation (with two trains per day).

The operation of the railway would enable transport services worth $250 to be performed. It would also mean the employment of factors of production which would reduce the value of production elsewhere by $100. Furthermore it would mean the destruction of crops worth $120. The coming of the railway will also have led to the abandonment of cultivation of some land. Since we know that, had this land been cultivated, the value of the crops destroyed by fire would have been $120, and since it is unlikely that the total crop on this land would have been destroyed, it seems reasonable to suppose that the value of the crop yield on this land would have been higher than this. Assume it would have been $160. But the abandonment of cultivation would have released factors of production for employment elsewhere. All we know is that the amount by which the value of production elsewhere will increase will be less than $160. Suppose that it is $150. Then the gain from operating the railway would be $250 (the value of the transport services) minus $100 (the cost of the factors of production) minus $120 (the value of crops destroyed by fire) minus $160 (the fall in the value of crop production due to the abandonment of cultivation) plus $150 (the value of production elsewhere of the released factors of production). Overall, operating the railway will increase the value of total production by $20. With these figures it is clear that it is better that the railway should not be liable for the damage it causes, thus enabling it to operate profitably. Of course, by altering the

figures, it could be shown that there are other cases in which it would be desirable that the railway should be liable for the damage it causes. It is enough for my purpose to show that, from an economic point of view, a situation in which there is "uncompensated damage done to surrounding woods by sparks from railway engines" is not necessarily undesirable. Whether it is desirable or not depends on the particular circumstances.

How is it that the Pigovian analysis seems to give the wrong answer? The reason is that Pigou does not seem to have noticed that his analysis is dealing with an entirely different question. The analysis as such is correct. But it is quite illegitimate for Pigou to draw the particular conclusion he does. The question at issue is not whether it is desirable to run an additional train or a faster train or to install smoke-preventing devices; the question at issue is whether it is desirable to have a system in which the railway has to compensate those who suffer damage from the fires which it causes or one in which the railway does not have to compensate them. When an economist is comparing alternative social arrangements, the proper procedure is to compare the total social product yielded by these different arrangements. The comparison of private and social products is neither here nor there. A simple example will demonstrate this. Imagine a town in which there are traffic lights. A motorist approaches an intersection and stops because the light is red. There are no cars approaching the intersection on the other street. If the motorist ignored the red signal, no accident would occur and the total product would increase because the motorist would arrive earlier at his destination. Why does he not do this? The reason is that if he ignored the light he would be fined. The private product from crossing the street is less than the social product. Should we conclude from this that the total product would be greater if there were no fines for failing to obey traffic signals? The Pigovian analysis shows us that it is possible to conceive of better worlds than the one in which we live. But the problem is to devise practical arrangements which will correct defects in one part of the system without causing more serious harm in other parts.

I have examined in considerable detail one example of a divergence between private and social products and I do not propose to make any further examination of Pigou's analytical system. But the main discussion of the problem considered in this article is to be found in that part of Chapter 9 in Part II which deals with Pigou's second class of divergence and it is of interest to see how Pigou develops his argument. Pigou's own description of this second class of divergence was quoted at the beginning of this section. Pigou distinguishes between the case in which a person renders services for which he receives no payment and the case in which a person renders disservices and compensation is not given to the injured parties. Our main attention has, of course, centred on this second case. It is therefore rather

astonishing to find, as was pointed out to me by Professor Francesco Forte, that the problem of the smoking chimney—the "stock instance"[44] or "classroom example"[45] of the second case—is used by Pigou as an example of the first case (services rendered without payment) and is never mentioned, at any rate explicitly, in connection with the second case.[46] Pigou points out that factory owners who devote resources to preventing their chimneys from smoking render services for which they receive no payment. The implication, in the light of Pigou's discussion later in the chapter, is that a factory owner with a smokey chimney should be given a bounty to induce him to install smoke-preventing devices. Most modern economists would suggest that the owner of the factory with the smokey chimney should be taxed. It seems a pity that economists (apart from Professor Forte) do not seem to have noticed this feature of Pigou's treatment since a realisation that the problem could be tackled in either of these two ways would probably have led to an explicit recognition of its reciprocal nature.

In discussing the second case (disservices without compensation to those damaged), Pigou says that they are rendered "when the owner of a site in a residential quarter of a city builds a factory there and so destroys a great part of the amenities of neighbouring sites; or, in a less degree, when he uses his site in such a way as to spoil the lighting of the house opposite; or when he invests resources in erecting buildings in a crowded centre, which by contracting the air-space and the playing room of the neighbourhood, tend to injure the health and efficiency of the families living there."[47] Pigou is, of course, quite right to describe such actions as "uncharged disservices." But he is wrong when he describes these actions as "anti-social."[48] They may or may not be. It is necessary to weigh the harm against the good that will result. Nothing could be more "anti-social" than to oppose any action which causes any harm to anyone.

The example with which Pigou opens his discussion of "uncharged disservices" is not, as I have indicated, the case of the smokey chimney but the case of the overrunning rabbits: ". . . incidental uncharged disservices are rendered to third parties when the game-preserving activities of one occupier involve the overrunning of a neighbouring occupier's land by rabbits. . . ." This example is of extraordinary interest, not so much because the economic

[44] Sir Dennis Robertson, I Lectures on Economic Principles 162 (1957).

[45] E. J. Mishan, The Meaning of Efficiency in Economics, 189 The Bankers' Magazine 482 (June 1960).

[46] Pigou, op. cit. supra n. 35 at 184.

[47] *Id.* at 185–86.

[48] *Id.* at 186 n.1. For similar unqualified statements see Pigou's lecture "Some Aspects of the Housing Problem" in B. S. Rowntree and A. C. Pigou, Lectures on Housing, in 18 Manchester Univ. Lectures (1914).

analysis of the case is essentially any different from that of the other examples, but because of the peculiarities of the legal position and the light it throws on the part which economics can play in what is apparently the purely legal question of the delimitation of rights.

The problem of legal liability for the actions of rabbits is part of the general subject of liability for animals.[49] I will, although with reluctance, confine my discussion to rabbits. The early cases relating to rabbits concerned the relations between the lord of the manor and commoners, since, from the thirteenth century on, it became usual for the lord of the manor to stock the commons with conies (rabbits), both for the sake of the meat and the fur. But in 1597, in *Boulston*'s case, an action was brought by one landowner against a neighbouring landowner, alleging that the defendant had made coney-burrows and that the conies had increased and had destroyed the plaintiff's corn. The action failed for the reason that

> . . . so soon as the coneys come on his neighbor's land he may kill them, for they are ferae naturae, and he who makes the coney-boroughs has no property in them, and he shall not be punished for the damage which the coneys do in which he has no property, and which the other may lawfully kill.[50]

As *Boulston*'s case has been treated as binding—Bray, J., in 1919, said that he was not aware that *Boulston*'s case has ever been overruled or questioned[51]—Pigou's rabbit example undoubtedly represented the legal position at the time *The Economics of Welfare* was written.[52] And in this case, it is not far from the truth to say that the state of affairs which Pigou describes came about because of an absence of Government action (at any rate in the form of statutory enactments) and was the result of "natural" tendencies.

Nonetheless, *Boulston*'s case is something of a legal curiousity and Professor Williams makes no secret of his distaste for this decision:

[49] See G. L. Williams, Liability for Animals—An Account of the Development and Present Law of Tortious Liability for Animals, Distress Damage Feasant and the Duty to Fence, in Great Britain, Northern Ireland and the Common Law Dominions (1939). Part Four, "The Action of Nuisance, in Relation to Liability for Animals," 236–62, is especially relevant to our discussion. The problem of liability for rabbits is discussed in this part, 238–47. I do not know how far the common law in the United State regarding liability for animals has diverged from that in Britain. In some Western States of the United States, the English common law regarding the duty to fence has not been followed, in part because "the considerable amount of open, uncleared land made it a matter of public policy to allow cattle to run at large" (Williams, *op. cit. supra* 227). This affords a good example of how a different set of circumstances may make it economically desirable to change the legal rule regarding the delimitation of rights.

[50] 5 Coke (Vol. 3) 104 b. 77 Eng. Rep., 216, 217.

[51] See Stearn v. Prentice Bros. Ltd., (1919) 1 K.B., 395, 397.

[52] I have not looked into recent cases. The legal position has also been modified by statutory enactments.

The conception of liability in nuisance as being based upon ownership is the result, apparently, of a confusion with the action of cattle-trespass, and runs counter both to principle and to the medieval authorities on the escape of water, smoke and filth. . . . The prerequisite of any satisfactory treatment of the subject is the final abandonment of the pernicious doctrine in *Boulston*'s case. . . . Once *Boulston*'s case disappears, the way will be clear for a rational restatement of the whole subject, on lines that will harmonize with the principles prevailing in the rest of the law of nuisance.[53]

The judges in *Boulston*'s case were, of course, aware that their view of the matter depended on distinguishing this case from one involving nuisance:

This cause is not like to the cases put, on the other side, of erecting a lime-kiln, dye-house, or the like; for there the annoyance is by the act of the parties who make them; but it is not so here, for the conies of themselves went into the plaintiff's land, and he might take them when they came upon his land, and make profit of them.[54]

Professor Williams comments:

Once more the atavistic idea is emerging that the animals are guilty and not the landowner. It is not, of course, a satisfactory principle to introduce into a modern law of nuisance. If A. erects a house or plants a tree so that the rain runs or drips from it on to B.'s land, this is A.'s act for which he is liable; but if A. introduces rabbits into his land so that they escape from it into B.'s, this is the act of the rabbits for which A. is not liable—such is the specious distinction resulting from *Boulston*'s case.[55]

It has to be admitted that the decision in *Boulston*'s case seems a little odd. A man may be liable for damage caused by smoke or unpleasant smells, without it being necessary to determine whether he owns the smoke or the smell. And the rule in *Boulston*'s case has not always been followed in cases dealing with other animals. For example, in *Bland v. Yates*,[56] it was decided that an injunction could be granted to prevent someone from keeping an *unusual and excessive* collection of manure in which flies bred and which infested a neighbour's house. The question of who owned the flies was not raised. An economist would not wish to object because legal reasoning sometimes appears a little odd. But there is a sound economic reason for supporting Professor Williams' view that the problem of liability for animals (and particularly rabbits) should be brought within the ordinary law of nuisance. The reason is not that the man who harbours rabbits is solely responsible for the damage; the man whose crops are eaten is equally responsible. And given that the costs of market transactions make a rearrange-

[53] Williams, op. cit. supra n. 49 at 242, 258.

[54] Boulston v. Hardy, Cro. Eliz., 547, 548, 77 Eng. Rep. 216.

[55] Williams, op. cit. supra n. 49 at 243.

[56] 58 Sol.J. 612 (1913–1914).

ment of rights impossible, unless we know the particular circumstances, we cannot say whether it is desirable or not to make the man who harbours rabbits responsible for the damage committed by the rabbits on neighbouring properties. The objection to the rule in *Boulston*'s case is that, under it, the harbourer of rabbits can *never* be liable. It fixes the rule of liability at one pole: and this is as undesirable, from an economic point of view, as fixing the rule at the other pole and making the harbourer of rabbits always liable. But, as we saw in Section VII, the law of nuisance, as it is in fact handled by the courts, is flexible and allows for a comparison of the utility of an act with the harm it produces. As Professor Williams says: "The whole law of nuisance is an attempt to reconcile and compromise between conflicting interests. . . ."[57] To bring the problem of rabbits within the ordinary law of nuisance would not mean *inevitably* making the harbourer of rabbits liable for damage committed by the rabbits. This is not to say that the sole task of the courts in such cases is to make a comparison between the harm and the utility of an act. Nor is it to be expected that the courts will always decide correctly after making such a comparison. But unless the courts act very foolishly, the ordinary law of nuisance would seem likely to give economically more satisfactory results than adopting a rigid rule. Pigou's case of the overrunning rabbits affords an excellent example of how problems of law and economics are interrelated, even though the correct policy to follow would seem to be different from that envisioned by Pigou.

Pigou allows one exception to his conclusion that there is a divergence between private and social products in the rabbit example. He adds: ". . . unless . . . the two occupiers stand in the relation of landlord and tenant, so that compensation is given in an adjustment of the rent."[58] This qualification is rather surprising since Pigou's first class of divergence is largely concerned with the difficulties of drawing up satisfactory contracts between landlords and tenants. In fact, all the recent cases on the problem of rabbits cited by Professor Williams involved disputes between landlords and tenants concerning sporting rights.[59] Pigou seems to make a distinction between the case in which no contract is possible (the second class) and that in which the contract is unsatisfactory (the first class). Thus he says that the second class of divergences between private and social net product

> cannot, like divergences due to tenancy laws, be mitigated by a modification of the contractual relation between any two contracting parties, because the divergence arises out of a service or disservice rendered to persons other than the contracting parties.[60]

[57] Williams, op. cit. supra n. 49 at 259.

[58] Pigou, op. cit. supra n. 35 at 185.

[59] Williams, op. cit. supra n. 49 at 244–47.

[60] Pigou, op. cit. supra n. 35 at 192.

But the reason why some activities are not the subject of contracts is exactly the same as the reason why some contracts are commonly unsatisfactory—it would cost too much to put the matter right. Indeed, the two cases are really the same since the contracts are unsatisfactory because they do not cover certain activities. The exact bearing of the discussion of the first class of divergence on Pigou's main argument is difficult to discover. He shows that in some circumstances contractual relations between landlord and tenant may result in a divergence between private and social products.[61] But he also goes on to show that Government-enforced compensation schemes and rent-controls will also produce divergences.[62] Furthermore, he shows that, when the Government is in a similar position to a private landlord, e.g. when granting a franchise to a public utility, exactly the same difficulties arise as when private individuals are involved.[63] The discussion is interesting but I have been unable to discover what general conclusions about economic policy, if any, Pigou expects us to draw from it.

Indeed, Pigou's treatment of the problems considered in this article is extremely elusive and the discussion of his views raises almost insuperable difficulties of interpretation. Consequently it is impossible to be sure that one has understood what Pigou really meant. Nevertheless, it is difficult to resist the conclusion, extraordinary though this may be in an economist of Pigou's stature, that the main source of this obscurity is that Pigou had not thought his position through.

IX. The Pigovian Tradition

It is strange that a doctrine as faulty as that developed by Pigou should have been so influential, although part of its success has probably been due to the lack of clarity in the exposition. Not being clear, it was never clearly wrong. Curiously enough, this obscurity in the source has not prevented the emergence of a fairly well-defined oral tradition. What economists think they learn from Pigou, and what they tell their students, which I term the Pigovian tradition, is reasonably clear. I propose to show the inadequacy of this Pigovian tradition by demonstrating that both the analysis and the policy conclusions which it supports are incorrect.

I do not propose to justify my view as to the prevailing opinion by copious references to the literature. I do this partly because the treatment in the literature is usually so fragmentary, often involving little more than a reference to Pigou plus some explanatory comment, that detailed examination would be inappropriate. But the main reason for this lack of reference is that the doctrine, although based on Pigou, must have been largely the product of an oral tradition. Certainly economists with whom I have discussed these problems have shown a unanimity of opinion which is quite

[61] *Id.* 174–75. [62] *Id.* 177–83. [63] *Id.* 175–77.

remarkable considering the meagre treatment accorded this subject in the literature. No doubt there are some economists who do not share the usual view but they must represent a small minority of the profession.

The approach to the problems under discussion is through an examination of the value of physical production. The private product is the value of the additional product resulting from a particular activity of a business. The social product equals the private product minus the fall in the value of production elsewhere for which no compensation is paid by the business. Thus, if 10 units of a factor (and no other factors) are used by a business to make a certain product with a value of $105; and the owner of this factor is not compensated for their use, which he is unable to prevent; and these 10 units of the factor would yield products in their best alternative use worth $100; then, the social product is $105 minus $100 or $5. If the business now pays for one unit of the factor and its price equals the value of its marginal product, then the social product rises to $15. If two units are paid for, the social product rises to $25 and so on until it reaches $105 when all units of the factor are paid for. It is not difficult to see why economists have so readily accepted this rather odd procedure. The analysis focusses on the individual business decision and since the use of certain resources is not allowed for in costs, receipts are reduced by the same amount. But, of course, this means that the value of the social product has no social significance whatsoever. It seems to me preferable to use the opportunity cost concept and to approach these problems by comparing the value of the product yielded by factors in alternative uses or by alternative arrangements. The main advantage of a pricing system is that it leads to the employment of factors in places where the value of the product yielded is greatest and does so at less cost than alternative systems (I leave aside that a pricing system also eases the problem of the redistribution of income). But if through some God-given natural harmony factors flowed to the places where the value of the product yielded was greatest without any use of the pricing system and consequently there was no compensation, I would find it a source of surprise rather than a cause for dismay.

The definition of the social product is queer but this does not mean that the conclusions for policy drawn from the analysis are necessarily wrong. However, there are bound to be dangers in an approach which diverts attention from the basic issues and there can be little doubt that it has been responsible for some of the errors in current doctrine. The belief that it is desirable that the business which causes harmful effects should be forced to compensate those who suffer damage (which was exhaustively discussed in section VIII in connection with Pigou's railway sparks example) is undoubtedly the result of not comparing the total product obtainable with alternative social arrangements.

The same fault is to be found in proposals for solving the problem of harmful effects by the use of taxes or bounties. Pigou lays considerable stress on this solution although he is, as usual, lacking in detail and qualified in his support.[64] Modern economists tend to think exclusively in terms of taxes and in a very precise way. The tax should be equal to the damage done and should therefore vary with the amount of the harmful effect. As it is not proposed that the proceeds of the tax should be paid to those suffering the damage, this solution is not the same as that which would force a business to pay compensation to those damaged by its actions, although economists generally do not seem to have noticed this and tend to treat the two solutions as being identical.

Assume that a factory which emits smoke is set up in a district previously free from smoke pollution, causing damage valued at $100 per annum. Assume that the taxation solution is adopted and that the factory owner is taxed $100 per annum as long as the factory emits the smoke. Assume further that a smoke-preventing device costing $90 per annum to run is available. In these circumstances, the smoke-preventing device would be installed. Damage of $100 would have been avoided at an expenditure of $90 and the factory-owner would be better off by $10 per annum. Yet the position achieved may not be optimal. Suppose that those who suffer the damage could avoid it by moving to other locations or by taking various precautions which would cost them, or be equivalent to a loss in income of, $40 per annum. Then there would be a gain in the value of production of $50 if the factory continued to emit its smoke and those now in the district moved elesewhere or made other adjustments to avoid the damage. If the factory owner is to be made to pay a tax equal to the damage caused, it would clearly be desirable to institute a double tax system and to make residents of the district pay an amount equal to the additional cost incurred by the factory owner (or the consumers of his products) in order to avoid the damage. In these conditions, people would not stay in the district or would take other measures to prevent the damage from occurring, when the costs of doing so were less than the costs that would be incurred by the producer to reduce the damage (the producer's object, of course, being not so much to reduce the damage as to reduce the tax payments). A tax system which was confined to a tax on the producer for damage caused would tend to lead to unduly high costs being incurred for the prevention of damage. Of course this could be avoided if it were possible to base the tax, not on the damage caused, but on the fall in the value of production (in its widest sense) resulting from the emission of smoke. But to do so would require a detailed knowledge of individual preferences and I am unable to imagine how the data needed for such a taxation system could be assembled. Indeed,

[64] *Id.* 192–4, 381 and Public Finance 94–100 (3d ed. 1947).

the proposal to solve the smoke-pollution and similar problems by the use of taxes bristles with difficulties: the problem of calculation, the difference between average and marginal damage, the interrelations between the damage suffered on different properties, etc. But it is unnecessary to examine these problems here. It is enough for my purpose to show that, even if the tax is exactly adjusted to equal the damage that would be done to neighboring properties as a result of the emission of each additional puff of smoke, the tax would not necessarily bring about optimal conditions. An increase in the number of people living or of business operating in the vicinity of the smoke-emitting factory will increase the amount of harm produced by a given emission of smoke. The tax that would be imposed would therefore increase with an increase in the number of those in the vicinity. This will tend to lead to a decrease in the value of production of the factors employed by the factory, either because a reduction in production due to the tax will result in factors being used elsewhere in ways which are less valuable, or because factors will be diverted to produce means for reducing the amount of smoke emitted. But people deciding to establish themselves in the vicinity of the factory will not take into account this fall in the value of production which results from their presence. This failure to take into account costs imposed on others is comparable to the action of a factory-owner in not taking into account the harm resulting from his emission of smoke. Without the tax, there may be too much smoke and too few people in the vicinity of the factory; but with the tax there may be too little smoke and too many people in the vicinity of the factory. There is no reason to suppose that one of these results is necessarily preferable.

I need not devote much space to discussing the similar error involved in the suggestion that smoke producing factories should, by means of zoning regulations, be removed from the districts in which the smoke causes harmful effects. When the change in the location of the factory results in a reduction in production, this obviously needs to be taken into account and weighed against the harm which would result from the factory remaining in that location. The aim of such regulation should not be to eliminate smoke pollution but rather to secure the optimum amount of smoke pollution, this being the amount which will maximise the value of production.

X. A Change of Approach

It is my belief that the failure of economists to reach correct conclusions about the treatment of harmful effects cannot be ascribed simply to a few slips in analysis. It stems from basic defects in the current approach to problems of welfare economics. What is needed is a change of approach.

Analysis in terms of divergencies between private and social products concentrates attention on particular deficiencies in the system and tends to

nourish the belief that any measure which will remove the deficiency is necessarily desirable. It diverts attention from those other changes in the system which are inevitably associated with the corrective measure, changes which may well produce more harm than the original deficiency. In the preceding sections of this article, we have seen many examples of this. But it is not necessary to approach the problem in this way. Economists who study problems of the firm habitually use an opportunity cost approach and compare the receipts obtained from a given combination of factors with alternative business arrangements. It would seem desirable to use a similar approach when dealing with questions of economic policy and to compare the total product yielded by alternative social arrangements. In this article, the analysis has been confined, as is usual in this part of economics, to comparisons of the value of production, as measured by the market. But it is, of course, desirable that the choice between different social arrangements for the solution of economic problems should be carried out in broader terms than this and that the total effect of these arrangements in all spheres of life should be taken into account. As Frank H. Knight has so often emphasized, problems of welfare economics must ultimately dissolve into a study of aesthetics and morals.

A second feature of the usual treatment of the problems discussed in this article is that the analysis proceeds in terms of a comparison between a state of laissez faire and some kind of ideal world. This approach inevitably leads to a looseness of thought since the nature of the alternatives being compared is never clear. In a state of laissez faire, is there a monetary, a legal or a political system and if so, what are they? In an ideal world, would there be a monetary, a legal or a political system and if so, what would they be? The answers to all these questions are shrouded in mystery and every man is free to draw whatever conclusions he likes. Actually very little analysis is required to show that an ideal world is better than a state of laissez faire, unless the definitions of a state of laissez faire and an ideal world happen to be the same. But the whole discussion is largely irrelevant for questions of economic policy since whatever we may have in mind as our ideal world, it is clear that we have not yet discovered how to get to it from where we are. A better approach would seem to be to start our analysis with a situation approximating that which actually exists, to examine the effects of a proposed policy change and to attempt to decide whether the new situation would be, in total, better or worse than the original one. In this way, conclusions for policy would have some relevance to the actual situation.

A final reason for the failure to develop a theory adequate to handle the problem of harmful effects stems from a faulty concept of a factor of production. This is usually thought of as a physical entity which the businessman acquires and uses (an acre of land, a ton of fertiliser) instead of as a

right to perform certain (physical) actions. We may speak of a person owning land and using it as a factor of production but what the land-owner in fact possesses is the right to carry out a circumscribed list of actions. The rights of a land-owner are not unlimited. It is not even always possible for him to remove the land to another place, for instance, by quarrying it. And although it may be possible for him to exclude some people from using "his" land, this may not be true of others. For example, some people may have the right to cross the land. Furthermore, it may or may not be possible to erect certain types of buildings or to grow certain crops or to use particular drainage systems on the land. This does not come about simply because of Government regulation. It would be equally true under the common law. In fact it would be true under any system of law. A system in which the rights of individuals were unlimited would be one in which there were no rights to acquire.

If factors of production are thought of as rights, it becomes easier to understand that the right to do something which has a harmful effect (such as the creation of smoke, noise, smells, etc.) is also a factor of production. Just as we may use a piece of land in such a way as to prevent someone else from crossing it, or parking his car, or building his house upon it, so we may use it in such a way as to deny him a view or quiet or unpolluted air. The cost of exercising a right (of using a factor of production) is always the loss which is suffered elsewhere in consequence of the exercise of that right—the inability to cross land, to park a car, to build a house, to enjoy a view, to have peace and quiet or to breathe clean air.

It would clearly be desirable if the only actions performed were those in which what was gained was worth more than what was lost. But in choosing between social arrangements within the context of which individual decisions are made, we have to bear in mind that a change in the existing system which will lead to an improvement in some decisions may well lead to a worsening of others. Furthermore we have to take into account the costs involved in operating the various social arrangements (whether it be the working of a market or of a government department), as well as the costs involved in moving to a new system. In devising and choosing between social arrangements we should have regard for the total effect. This, above all, is the change in approach which I am advocating.

Why Nations Fail

THE ORIGINS OF POWER, PROSPERITY, AND POVERTY

Daron Acemoglu and

James A. Robinson

Crown Publishers • New York

3.

THE MAKING OF PROSPERITY AND POVERTY

THE ECONOMICS OF THE 38TH PARALLEL

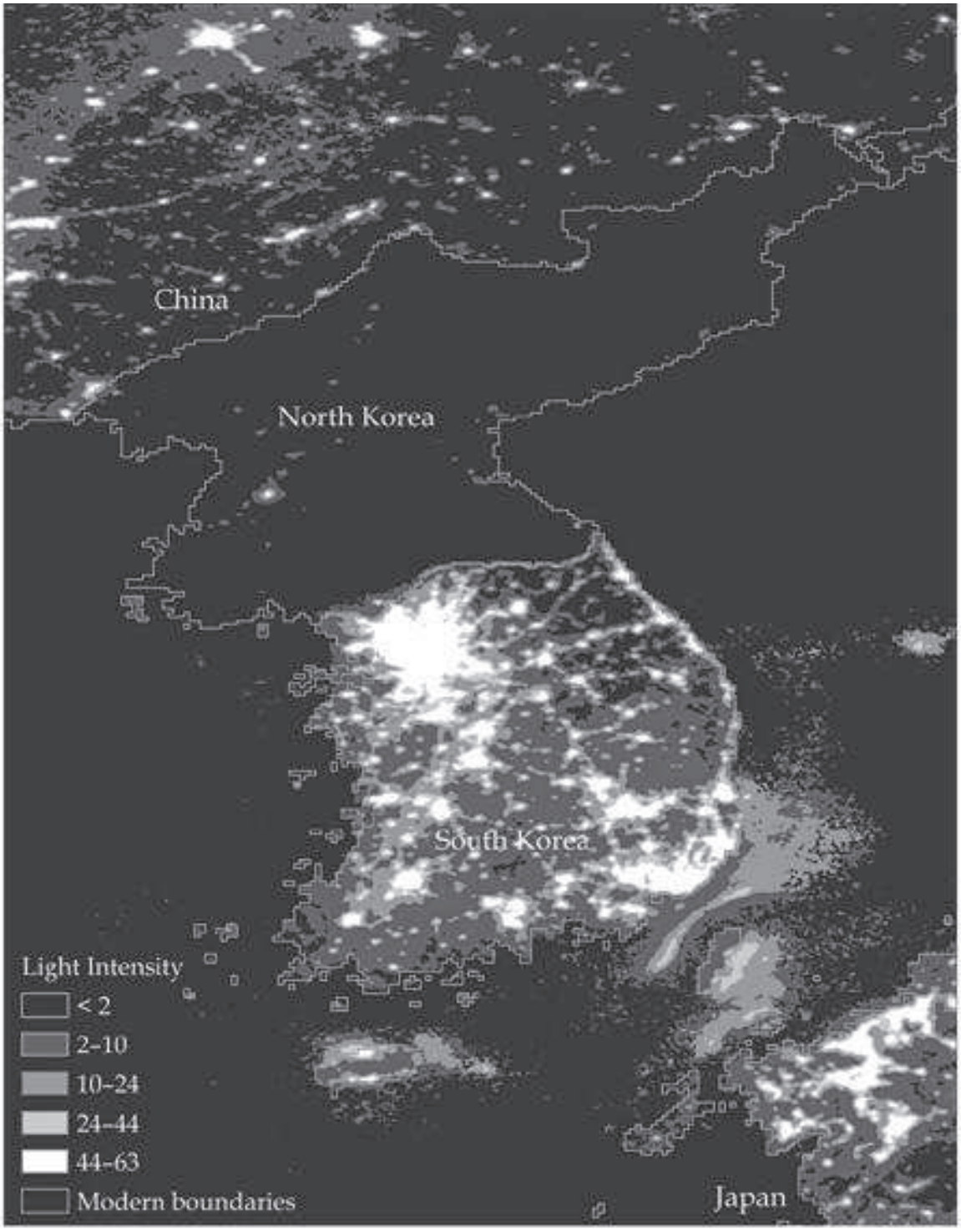
IN THE SUMMER OF 1945, as the Second World War was drawing to a close, the Japanese colony in Korea began to collapse. Within a month of Japan's August 15 unconditional surrender, Korea was divided at the 38th parallel into two spheres of influence. The South was administered by the United States. The North, by Russia. The uneasy peace of the cold war was shattered in June 1950 when the North Korean army invaded the South. Though initially the North Koreans made large inroads, capturing the capital city, Seoul, by the autumn, they were in full retreat. It was then that Hwang Pyŏng-Wŏn and his brother were separated. Hwang Pyŏng-Wŏn managed to hide and avoid being drafted into the North Korean army. He stayed in the South and worked as a pharmacist. His brother, a doctor working in Seoul treating wounded soldiers from the South Korean army, was taken north as the North Korean army retreated. Dragged apart in 1950, they met again in 2000 in Seoul for the first time in fifty years, after the two governments finally agreed to initiate a limited program of family reunification.

As a doctor, Hwang Pyŏng-Wŏn's brother had ended up working for the air force, a good job in a military dictatorship. But even those with privileges in North Korea don't do that well. When the brothers met, Hwang Pyŏng-Wŏn asked about how life was north of the 38th parallel. He had a car, but his brother didn't. "Do you have a telephone?" he asked his brother. "No," said his brother. "My daughter, who works at the Foreign Ministry, has a phone, but if you don't know the code you can't call." Hwang Pyŏng-Wŏn recalled how all the people from the North at the reunion were asking for money, so he offered some to his brother. But his brother said, "If I go back with money the

government will say, 'Give that money to us,' so keep it." Hwang Pyŏng-Wŏn noticed his brother's coat was threadbare: "Take off that coat and leave it, and when you go back wear this one," he suggested. "I can't do that," his brother replied. "This is just borrowed from the government to come here." Hwang Pyŏng-Wŏn recalled how when they parted, his brother was ill at ease and always nervous as though someone were listening. He was poorer than Hwang Pyŏng-Wŏn imagined. His brother said he lived well, but Hwang Pyŏng-Wŏn thought he looked awful and was thin as a rake.

The people of South Korea have living standards similar to those of Portugal and Spain. To the north, in the so-called Democratic People's Republic of Korea, or North Korea, living standards are akin to those of a sub-Saharan African country, about one-tenth of average living standards in South Korea. The health of North Koreans is in an even worse state; the average North Korean can expect to live ten years less than his cousins south of the 38th parallel. Map 7 illustrates in a dramatic way the economic gap between the Koreas. It plots data on the intensity of light at night from satellite images. North Korea is almost completely dark due to lack of electricity; South Korea is blazing with light.

These striking differences are not ancient. In fact, they did not exist prior to the end of the Second World War. But after 1945, the different governments in the North and the South adopted very different ways of organizing their economies. South Korea was led, and its early economic and political institutions were shaped, by the Harvard- and Princeton-educated, staunchly anticommunist Syngman Rhee, with significant support from the United States. Rhee was elected president in 1948. Forged in the midst of the Korean War and against the threat of communism spreading to the south of the 38th parallel, South Korea was no democracy. Both Rhee and his equally famous successor, General Park Chung-Hee, secured their places in history as authoritarian presidents. But both governed a market economy where private property was recognized, and after 1961, Park effectively threw the weight of the state behind rapid economic growth, channeling credit and subsidies to firms that were successful.



Map 7: Lights in South Korea and darkness in the North

The situation north of the 38th parallel was different. Kim Il-Sung, a leader of anti-Japanese communist partisans during the Second World War, established himself as dictator by 1947 and, with the help of the Soviet Union, introduced a rigid form of centrally planned economy as part of the so-called Juche system. Private property was outlawed, and markets were banned. Freedoms were curtailed not only in the marketplace, but in every sphere of

North Koreans' lives—except for those who happened to be part of the very small ruling elite around Kim Il-Sung and, later, his son and successor Kim Jong-Il.

It should not surprise us that the economic fortunes of South and North Korea diverged sharply. Kim Il-Sung's command economy and the Juche system soon proved to be a disaster. Detailed statistics are not available from North Korea, which is a secretive state, to say the least. Nonetheless, available evidence confirms what we know from the all-too-often recurring famines: not only did industrial production fail to take off, but North Korea in fact experienced a collapse in agricultural productivity. Lack of private property meant that few people had incentives to invest or to exert effort to increase or even maintain productivity. The stifling, repressive regime was inimical to innovation and the adoption of new technologies. But Kim Il-Sung, Kim Jong-Il, and their cronies had no intention of reforming the system, or introducing private property, markets, private contracts, or changing economic and political institutions. North Korea continues to stagnate economically.

Meanwhile, in the South, economic institutions encouraged investment and trade. South Korean politicians invested in education, achieving high rates of literacy and schooling. South Korean companies were quick to take advantage of the relatively educated population, the policies encouraging investment and industrialization, exports, and the transfer of technology. South Korea quickly became one of East Asia's "Miracle Economies," one of the most rapidly growing nations in the world.

By the late 1990s, in just about half a century, South Korean growth and North Korean stagnation led to a tenfold gap between the two halves of this once-united country—imagine what a difference a couple of centuries could make. The economic disaster of North Korea, which led to the starvation of millions, when placed against the South Korean economic success, is striking: neither culture nor geography nor ignorance can explain the divergent paths of North and South Korea. We have to look at institutions for an answer.

EXTRACTIVE AND INCLUSIVE ECONOMIC INSTITUTIONS

Countries differ in their economic success because of their different institutions, the rules influencing how the economy works, and the incentives that motivate people. Imagine teenagers in North and South Korea and what they expect from life. Those in the North grow up in poverty, without entrepreneurial initiative, creativity, or adequate education to prepare them for skilled work. Much of the education they receive at school is pure propaganda, meant to shore up the legitimacy of the regime; there are few books, let alone computers. After finishing school, everyone has to go into the army for ten years. These teenagers know that they will not be able to own property, start a business, or become more prosperous even if many people engage illegally in private economic activities to make a living. They also know that they will not have legal access to markets where they can use their skills or their earnings to purchase the goods they need and desire. They are even unsure about what kind of human rights they will have.

Those in the South obtain a good education, and face incentives that encourage them to exert effort and excel in their chosen vocation. South Korea is a market economy, built on private property. South Korean teenagers know that, if successful as entrepreneurs or workers, they can one day enjoy the fruits of their investments and efforts; they can improve their standard of living and buy cars, houses, and health care.

In the South the state supports economic activity. So it is possible for entrepreneurs to borrow money from banks and financial markets, for foreign companies to enter into partnerships with South Korean firms, for individuals to take up mortgages to buy houses. In the South, by and large, you are free to open any business you like. In the North, you are not. In the South, you can hire workers, sell your products or services, and spend your money in the marketplace in whichever way you want. In the North, there are only black markets. These different rules are the institutions under which North and South Koreans live.

Inclusive economic institutions, such as those in South Korea or in the United States, are those that allow and encourage participation by the great mass of people in

economic activities that make best use of their talents and skills and that enable individuals to make the choices they wish. To be inclusive, economic institutions must feature secure private property, an unbiased system of law, and a provision of public services that provides a level playing field in which people can exchange and contract; it also must permit the entry of new businesses and allow people to choose their careers.

THE CONTRAST OF South and North Korea, and of the United States and Latin America, illustrates a general principle. Inclusive economic institutions foster economic activity, productivity growth, and economic prosperity. Secure private property rights are central, since only those with such rights will be willing to invest and increase productivity. A businessman who expects his output to be stolen, expropriated, or entirely taxed away will have little incentive to work, let alone any incentive to undertake investments and innovations. But such rights must exist for the majority of people in society.

In 1680 the English government conducted a census of the population of its West Indian colony of Barbados. The census revealed that of the total population on the island of around 60,000, almost 39,000 were African slaves who were the property of the remaining one-third of the population. Indeed, they were mostly the property of the largest 175 sugar planters, who also owned most of the land. These large planters had secure and well-enforced property rights over their land and even over their slaves. If one planter wanted to sell slaves to another, he could do so and expect a court to enforce such a sale or any other contract he wrote. Why? Of the forty judges and justices of the peace on the island, twenty-nine of them were large planters. Also, the eight most senior military officials were all large planters. Despite well-defined, secure, and enforced property rights and contracts for the island's elite, Barbados did not have inclusive economic institutions, since two-thirds of the population were slaves with no access to education or economic opportunities, and no ability or incentive to use their talents or skills. Inclusive economic institutions require secure property rights and

economic opportunities not just for the elite but for a broad cross-section of society.

Secure property rights, the law, public services, and the freedom to contract and exchange all rely on the state, the institution with the coercive capacity to impose order, prevent theft and fraud, and enforce contracts between private parties. To function well, society also needs other public services: roads and a transport network so that goods can be transported; a public infrastructure so that economic activity can flourish; and some type of basic regulation to prevent fraud and malfeasance. Though many of these public services can be provided by markets and private citizens, the degree of coordination necessary to do so on a large scale often eludes all but a central authority. The state is thus inexorably intertwined with economic institutions, as the enforcer of law and order, private property, and contracts, and often as a key provider of public services. Inclusive economic institutions need and use the state.

The economic institutions of North Korea or of colonial Latin America—the *mita*, *encomienda*, or *repartimiento* described earlier—do not have these properties. Private property is nonexistent in North Korea. In colonial Latin America there was private property for Spaniards, but the property of the indigenous peoples was highly insecure. In neither type of society was the vast mass of people able to make the economic decisions they wanted to; they were subject to mass coercion. In neither type of society was the power of the state used to provide key public services that promoted prosperity. In North Korea, the state built an education system to inculcate propaganda, but was unable to prevent famine. In colonial Latin America, the state focused on coercing indigenous peoples. In neither type of society was there a level playing field or an unbiased legal system. In North Korea, the legal system is an arm of the ruling Communist Party, and in Latin America it was a tool of discrimination against the mass of people. We call such institutions, which have opposite properties to those we call inclusive, extractive economic institutions—extractive because such institutions are designed to extract incomes and wealth from one subset of society to benefit a different subset.

ENGINES OF PROSPERITY

Inclusive economic institutions create inclusive markets, which not only give people freedom to pursue the vocations in life that best suit their talents but also provide a level playing field that gives them the opportunity to do so. Those who have good ideas will be able to start businesses, workers will tend to go to activities where their productivity is greater, and less efficient firms can be replaced by more efficient ones. Contrast how people choose their occupations under inclusive markets to colonial Peru and Bolivia, where under the *mita*, many were forced to work in silver and mercury mines, regardless of their skills or whether they wanted to. Inclusive markets are not just free markets. Barbados in the seventeenth century also had markets. But in the same way that it lacked property rights for all but the narrow planter elite, its markets were far from inclusive; markets in slaves were in fact one part of the economic institutions systematically coercing the majority of the population and robbing them of the ability to choose their occupations and how they should utilize their talents.

Inclusive economic institutions also pave the way for two other engines of prosperity: technology and education. Sustained economic growth is almost always accompanied by technological improvements that enable people (labor), land, and existing capital (buildings, existing machines, and so on) to become more productive. Think of our great-great-grandparents, just over a century ago, who did not have access to planes or automobiles or most of the drugs and health care we now take for granted, not to mention indoor plumbing, air-conditioning, shopping malls, radio, or motion pictures; let alone information technology, robotics, or computer-controlled machinery. And going back a few more generations, the technological know-how and living standards were even more backward, so much so that we would find it hard to imagine how most people struggled through life. These improvements follow from science and from entrepreneurs such as Thomas Edison, who applied science to create profitable businesses. This process of innovation is made possible by economic institutions that encourage private property, uphold contracts, create a level

playing field, and encourage and allow the entry of new businesses that can bring new technologies to life. It should therefore be no surprise that it was U.S. society, not Mexico or Peru, that produced Thomas Edison, and that it was South Korea, not North Korea, that today produces technologically innovative companies such as Samsung and Hyundai.

Intimately linked to technology are the education, skills, competencies, and know-how of the workforce, acquired in schools, at home, and on the job. We are so much more productive than a century ago not just because of better technology embodied in machines but also because of the greater know-how that workers possess. All the technology in the world would be of little use without workers who knew how to operate it. But there is more to skills and competencies than just the ability to run machines. It is the education and skills of the workforce that generate the scientific knowledge upon which our progress is built and that enable the adaptation and adoption of these technologies in diverse lines of business. Though we saw in chapter 1 that many of the innovators of the Industrial Revolution and afterward, like Thomas Edison, were not highly educated, these innovations were much simpler than modern technology. Today technological change requires education both for the innovator and the worker. And here we see the importance of economic institutions that create a level playing field. The United States could produce, or attract from foreign lands, the likes of Bill Gates, Steve Jobs, Sergey Brin, Larry Page, and Jeff Bezos, and the hundreds of scientists who made fundamental discoveries in information technology, nuclear power, biotech, and other fields upon which these entrepreneurs built their businesses. The supply of talent was there to be harnessed because most teenagers in the United States have access to as much schooling as they wish or are capable of attaining. Now imagine a different society, for example the Congo or Haiti, where a large fraction of the population has no means of attending school, or where, if they manage to go to school, the quality of teaching is lamentable, where teachers do not show up for work, and even if they do, there may not be any books.

The low education level of poor countries is caused by

economic institutions that fail to create incentives for parents to educate their children and by political institutions that fail to induce the government to build, finance, and support schools and the wishes of parents and children. The price these nations pay for low education of their population and lack of inclusive markets is high. They fail to mobilize their nascent talent. They have many potential Bill Gateses and perhaps one or two Albert Einsteins who are now working as poor, uneducated farmers, being coerced to do what they don't want to do or being drafted into the army, because they never had the opportunity to realize their vocation in life.

The ability of economic institutions to harness the potential of inclusive markets, encourage technological innovation, invest in people, and mobilize the talents and skills of a large number of individuals is critical for economic growth. Explaining why so many economic institutions fail to meet these simple objectives is the central theme of this book.

Extractive and Inclusive Political Institutions

All economic institutions are created by society. Those of North Korea, for example, were forced on its citizens by the communists who took over the country in the 1940s, while those of colonial Latin America were imposed by Spanish conquistadors. South Korea ended up with very different economic institutions than the North because different people with different interests and objectives made the decisions about how to structure society. In other words, South Korea had different politics.

Politics is the process by which a society chooses the rules that will govern it. Politics surrounds institutions for the simple reason that while inclusive institutions may be good for the economic prosperity of a nation, some people or groups, such as the elite of the Communist Party of North Korea or the sugar planters of colonial Barbados, will be much better off by setting up institutions that are extractive. When there is conflict over institutions, what happens depends on which people or group wins out in the game of politics—who can get more support, obtain additional resources, and form more effective alliances. In short, who

wins depends on the distribution of political power in society.

The political institutions of a society are a key determinant of the outcome of this game. They are the rules that govern incentives in politics. They determine how the government is chosen and which part of the government has the right to do what. Political institutions determine who has power in society and to what ends that power can be used. If the distribution of power is narrow and unconstrained, then the political institutions are absolutist, as exemplified by the absolutist monarchies reigning throughout the world during much of history. Under absolutist political institutions such as those in North Korea and colonial Latin America, those who can wield this power will be able to set up economic institutions to enrich themselves and augment their power at the expense of society. In contrast, political institutions that distribute power broadly in society and subject it to constraints are pluralistic. Instead of being vested in a single individual or a narrow group, political power rests with a broad coalition or a plurality of groups.

There is obviously a close connection between pluralism and inclusive economic institutions. But the key to understanding why South Korea and the United States have inclusive economic institutions is not just their pluralistic political institutions but also their sufficiently centralized and powerful states. A telling contrast is with the East African nation of Somalia. As we will see later in the book, political power in Somalia has long been widely distributed—almost pluralistic. Indeed there is no real authority that can control or sanction what anyone does. Society is divided into deeply antagonistic clans that cannot dominate one another. The power of one clan is constrained only by the guns of another. This distribution of power leads not to inclusive institutions but to chaos, and at the root of it is the Somali state's lack of any kind of political centralization, or state centralization, and its inability to enforce even the minimal amount of law and order to support economic activity, trade, or even the basic security of its citizens.

Max Weber, who we met in the previous chapter, provided the most famous and widely accepted definition

of the state, identifying it with the “monopoly of legitimate violence” in society. Without such a monopoly and the degree of centralization that it entails, the state cannot play its role as enforcer of law and order, let alone provide public services and encourage and regulate economic activity. When the state fails to achieve almost any political centralization, society sooner or later descends into chaos, as did Somalia.

We will refer to political institutions that are sufficiently centralized and pluralistic as inclusive political institutions. When either of these conditions fails, we will refer to the institutions as extractive political institutions.

There is strong synergy between economic and political institutions. Extractive political institutions concentrate power in the hands of a narrow elite and place few constraints on the exercise of this power. Economic institutions are then often structured by this elite to extract resources from the rest of the society. Extractive economic institutions thus naturally accompany extractive political institutions. In fact, they must inherently depend on extractive political institutions for their survival. Inclusive political institutions, vesting power broadly, would tend to uproot economic institutions that expropriate the resources of the many, erect entry barriers, and suppress the functioning of markets so that only a few benefit.

In Barbados, for example, the plantation system based on the exploitation of slaves could not have survived without political institutions that suppressed and completely excluded the slaves from the political process. The economic system impoverishing millions for the benefit of a narrow communist elite in North Korea would also be unthinkable without the total political domination of the Communist Party.

This synergistic relationship between extractive economic and political institutions introduces a strong feedback loop: political institutions enable the elites controlling political power to choose economic institutions with few constraints or opposing forces. They also enable the elites to structure future political institutions and their evolution. Extractive economic institutions, in turn, enrich the same elites, and their economic wealth and power help consolidate their political dominance. In Barbados or in

Latin America, for example, the colonists were able to use their political power to impose a set of economic institutions that made them huge fortunes at the expense of the rest of the population. The resources these economic institutions generated enabled these elites to build armies and security forces to defend their absolutist monopoly of political power. The implication of course is that extractive political and economic institutions support each other and tend to persist.

There is in fact more to the synergy between extractive economic and political institutions. When existing elites are challenged under extractive political institutions and the newcomers break through, the newcomers are likewise subject to only a few constraints. They thus have incentives to maintain these political institutions and create a similar set of economic institutions, as Porfirio Díaz and the elite surrounding him did at the end of the nineteenth century in Mexico.

Inclusive economic institutions, in turn, are forged on foundations laid by inclusive political institutions, which make power broadly distributed in society and constrain its arbitrary exercise. Such political institutions also make it harder for others to usurp power and undermine the foundations of inclusive institutions. Those controlling political power cannot easily use it to set up extractive economic institutions for their own benefit. Inclusive economic institutions, in turn, create a more equitable distribution of resources, facilitating the persistence of inclusive political institutions.

It was not a coincidence that when, in 1618, the Virginia Company gave land, and freedom from their draconian contracts, to the colonists it had previously tried to coerce, the General Assembly in the following year allowed the colonists to begin governing themselves. Economic rights without political rights would not have been trusted by the colonists, who had seen the persistent efforts of the Virginia Company to coerce them. Neither would these economies have been stable and durable. In fact, combinations of extractive and inclusive institutions are generally unstable. Extractive economic institutions under inclusive political institutions are unlikely to survive for long, as our discussion of Barbados suggests.

Similarly, inclusive economic institutions will neither support nor be supported by extractive political ones. Either they will be transformed into extractive economic institutions to the benefit of the narrow interests that hold power, or the economic dynamism they create will destabilize the extractive political institutions, opening the way for the emergence of inclusive political institutions. Inclusive economic institutions also tend to reduce the benefits the elites can enjoy by ruling over extractive political institutions, since those institutions face competition in the marketplace and are constrained by the contracts and property rights of the rest of society.

WHY NOT ALWAYS CHOOSE PROSPERITY?

Political and economic institutions, which are ultimately the choice of society, can be inclusive and encourage economic growth. Or they can be extractive and become impediments to economic growth. Nations fail when they have extractive economic institutions, supported by extractive political institutions that impede and even block economic growth. But this means that the choice of institutions—that is, the politics of institutions—is central to our quest for understanding the reasons for the success and failure of nations. We have to understand why the politics of some societies lead to inclusive institutions that foster economic growth, while the politics of the vast majority of societies throughout history has led, and still leads today, to extractive institutions that hamper economic growth.

It might seem obvious that everyone should have an interest in creating the type of economic institutions that will bring prosperity. Wouldn't every citizen, every politician, and even a predatory dictator want to make his country as wealthy as possible?

Let's return to the Kingdom of Kongo we discussed earlier. Though this kingdom collapsed in the seventeenth century, it provided the name for the modern country that became independent from Belgian colonial rule in 1960. As an independent polity, Congo experienced almost unbroken economic decline and mounting poverty under the rule of Joseph Mobutu between 1965 and 1997. This

decline continued after Mobutu was overthrown by Laurent Kabila. Mobutu created a highly extractive set of economic institutions. The citizens were impoverished, but Mobutu and the elite surrounding him, known as Les Grosses Legumes (the Big Vegetables), became fabulously wealthy. Mobutu built himself a palace at his birthplace, Gbadolite, in the north of the country, with an airport large enough to land a supersonic Concord jet, a plane he frequently rented from Air France for travel to Europe. In Europe he bought castles and owned large tracts of the Belgian capital of Brussels.

Wouldn't it have been better for Mobutu to set up economic institutions that increased the wealth of the Congolese rather than deepening their poverty? If Mobutu had managed to increase the prosperity of his nation, would he not have been able to appropriate even more money, buy a Concord instead of renting one, have more castles and mansions, possibly a bigger and more powerful army? Unfortunately for the citizens of many countries in the world, the answer is no. Economic institutions that create incentives for economic progress may simultaneously redistribute income and power in such a way that a predatory dictator and others with political power may become worse off.

The fundamental problem is that there will necessarily be disputes and conflict over economic institutions. Different institutions have different consequences for the prosperity of a nation, how that prosperity is distributed, and who has power. The economic growth which can be induced by institutions creates both winners and losers. This was clear during the Industrial Revolution in England, which laid the foundations of the prosperity we see in the rich countries of the world today. It centered on a series of pathbreaking technological changes in steam power, transportation, and textile production. Even though mechanization led to enormous increases in total incomes and ultimately became the foundation of modern industrial society, it was bitterly opposed by many. Not because of ignorance or shortsightedness; quite the opposite. Rather, such opposition to economic growth has its own, unfortunately coherent, logic. Economic growth and technological change are accompanied by what the great economist

Joseph Schumpeter called creative destruction. They replace the old with the new. New sectors attract resources away from old ones. New firms take business away from established ones. New technologies make existing skills and machines obsolete. The process of economic growth and the inclusive institutions upon which it is based create losers as well as winners in the political arena and in the economic marketplace. Fear of creative destruction is often at the root of the opposition to inclusive economic and political institutions.

European history provides a vivid example of the consequences of creative destruction. On the eve of the Industrial Revolution in the eighteenth century, the governments of most European countries were controlled by aristocracies and traditional elites, whose major source of income was from landholdings or from trading privileges they enjoyed thanks to monopolies granted and entry barriers imposed by monarchs. Consistent with the idea of creative destruction, the spread of industries, factories, and towns took resources away from the land, reduced land rents, and increased the wages that landowners had to pay their workers. These elites also saw the emergence of new businessmen and merchants eroding their trading privileges. All in all, they were the clear economic losers from industrialization. Urbanization and the emergence of a socially conscious middle and working class also challenged the political monopoly of landed aristocracies. So with the spread of the Industrial Revolution the aristocracies weren't just the economic losers; they also risked becoming political losers, losing their hold on political power. With their economic and political power under threat, these elites often formed a formidable opposition against industrialization.

The aristocracy was not the only loser from industrialization. Artisans whose manual skills were being replaced by mechanization likewise opposed the spread of industry. Many organized against it, rioting and destroying the machines they saw as responsible for the decline of their livelihood. They were the Luddites, a word that has today become synonymous with resistance to technological change. John Kay, English inventor of the "flying shuttle" in 1733, one of the first significant improvements in the

mechanization of weaving, had his house burned down by Luddites in 1753. James Hargreaves, inventor of the "spinning jenny," a complementary revolutionary improvement in spinning, got similar treatment.

In reality, the artisans were much less effective than the landowners and elites in opposing industrialization. The Luddites did not possess the political power—the ability to affect political outcomes against the wishes of other groups—of the landed aristocracy. In England, industrialization marched on, despite the Luddites' opposition, because aristocratic opposition, though real, was muted. In the Austro-Hungarian and the Russian empires, where the absolutist monarchs and aristocrats had far more to lose, industrialization was blocked. In consequence, the economies of Austria-Hungary and Russia stalled. They fell behind other European nations, where economic growth took off during the nineteenth century.

The success and failure of specific groups notwithstanding, one lesson is clear: powerful groups often stand against economic progress and against the engines of prosperity. Economic growth is not just a process of more and better machines, and more and better educated people, but also a transformative and destabilizing process associated with widespread creative destruction. Growth thus moves forward only if not blocked by the economic losers who anticipate that their economic privileges will be lost and by the political losers who fear that their political power will be eroded.

Conflict over scarce resources, income and power, translates into conflict over the rules of the game, the economic institutions, which will determine the economic activities and who will benefit from them. When there is a conflict, the wishes of all parties cannot be simultaneously met. Some will be defeated and frustrated, while others will succeed in securing outcomes they like. Who the winners of this conflict are has fundamental implications for a nation's economic trajectory. If the groups standing against growth are the winners, they can successfully block economic growth, and the economy will stagnate.

The logic of why the powerful would not necessarily want to set up the economic institutions that promote economic success extends easily to the choice of political institutions.

In an absolutist regime, some elites can wield power to set up economic institutions they prefer. Would they be interested in changing political institutions to make them more pluralistic? In general not, since this would only dilute their political power, making it more difficult, maybe impossible, for them to structure economic institutions to further their own interests. Here again we see a ready source of conflict. The people who suffer from the extractive economic institutions cannot hope for absolutist rulers to voluntarily change political institutions and redistribute power in society. The only way to change these political institutions is to force the elite to create more pluralistic institutions.

In the same way that there is no reason why political institutions should automatically become pluralistic, there is no natural tendency toward political centralization. There would certainly be incentives to create more centralized state institutions in any society, particularly in those with no such centralization whatsoever. For example, in Somalia, if one clan created a centralized state capable of imposing order on the country, this could lead to economic benefits and make this clan richer. What stops this? The main barrier to political centralization is again a form of fear from change: any clan, group, or politician attempting to centralize power in the state will also be centralizing power in their own hands, and this is likely to meet the ire of other clans, groups, and individuals, who would be the political losers of this process. Lack of political centralization means not only lack of law and order in much of a territory but also there being many actors with sufficient powers to block or disrupt things, and the fear of their opposition and violent reaction will often deter many would-be centralizers. Political centralization is likely only when one group of people is sufficiently more powerful than others to build a state. In Somalia, power is evenly balanced, and no one clan can impose its will on any other. Therefore, the lack of political centralization persists.

THE LONG AGONY OF THE CONGO

There are few better, or more depressing, examples of the forces that explain the logic of why economic prosperity is

so persistently rare under extractive institutions or that illustrate the synergy between extractive economic and political institutions than the Congo. Portuguese and Dutch visitors to Kongo in the fifteenth and sixteenth centuries remarked on the "miserable poverty" there. Technology was rudimentary by European standards, with the Kongolese having neither writing, the wheel, nor the plow. The reason for this poverty, and the reluctance of Kongolese farmers to adopt better technologies when they learned of them, is clear from existing historical accounts. It was due to the extractive nature of the country's economic institutions.

As we have seen, the Kingdom of Kongo was governed by the king in Mbanza, subsequently São Salvador. Areas away from the capital were ruled by an elite who played the roles of governors of different parts of the kingdom. The wealth of this elite was based on slave plantations around São Salvador and the extraction of taxes from the rest of the country. Slavery was central to the economy, used by the elite to supply their own plantations and by Europeans on the coast. Taxes were arbitrary; one tax was even collected every time the king's beret fell off. To become more prosperous, the Kongolese people would have had to save and invest—for example, by buying plows. But it would not have been worthwhile, since any extra output that they produced using better technology would have been subject to expropriation by the king and his elite. Instead of investing to increase their productivity and selling their products in markets, the Kongolese moved their villages away from the market; they were trying to be as far away from the roads as possible, in order to reduce the incidence of plunder and to escape the reach of slave traders.

The poverty of the Kongo was therefore the result of extractive economic institutions that blocked all the engines of prosperity or even made them work in reverse. The Kongo's government provided very few public services to its citizens, not even basic ones, such as secure property rights or law and order. On the contrary, the government was itself the biggest threat to its subjects' property and human rights. The institution of slavery meant that the most fundamental market of all, an inclusive labor market where people can choose their occupation or jobs in ways that are

so crucial for a prosperous economy, did not exist. Moreover, long-distance trade and mercantile activities were controlled by the king and were open only to those associated with him. Though the elite quickly became literate after the Portuguese introduced writing, the king made no attempt to spread literacy to the great mass of the population.

Nevertheless, though "miserable poverty" was widespread, the Kongolese extractive institutions had their own impeccable logic: they made a few people, those with political power, very rich. In the sixteenth century, the king of Kongo and the aristocracy were able to import European luxury goods and were surrounded by servants and slaves.

The roots of the economic institutions of Kongolese society flowed from the distribution of political power in society and thus from the nature of political institutions. There was nothing to stop the king from taking people's possessions or bodies, other than the threat of revolt. Though this threat was real, it was not enough to make people or their wealth secure. The political institutions of Kongo were truly absolutist, making the king and the elite subject to essentially no constraints, and it gave no say to the citizens in the way their society was organized.

Of course, it is not difficult to see that the political institutions of Kongo contrast sharply with inclusive political institutions where power is constrained and broadly distributed. The absolutist institutions of Kongo were kept in place by the army. The king had a standing army of five thousand troops in the mid-seventeenth century, with a core of five hundred musketeers—a formidable force for its time. Why the king and the aristocracy so eagerly adopted European firearms is thus easy to understand.

There was no chance of sustained economic growth under this set of economic institutions and even incentives for generating temporary growth were highly limited. Reforming economic institutions to improve individual property rights would have made the Kongolese society at large more prosperous. But it is unlikely that the elite would have benefited from this wider prosperity. First, such reforms would have made the elite economic losers, by undermining the wealth that the slave trade and slave plantations brought them. Second, such reforms would

have been possible only if the political power of the king and the elite were curtailed. For instance, if the king continued to command his five hundred musketeers, who would have believed an announcement that slavery had been abolished? What would have stopped the king from changing his mind later on? The only real guarantee would have been a change in political institutions so that citizens gained some countervailing political power, giving them some say over taxation or what the musketeers did. But in this case it is dubious that sustaining the consumption and lifestyle of the king and the elite would have been high on their list of priorities. In this scenario, changes that would have created better economic institutions in society would have made the king and aristocracy political as well as economic losers.

The interaction of economic and political institutions five hundred years ago is still relevant for understanding why the modern state of Congo is still miserably poor today. The advent of European rule in this area, and deeper into the basin of the River Congo at the time of the "scramble for Africa" in the late nineteenth century, led to an insecurity of human and property rights even more egregious than that which characterized the precolonial Kongo. In addition, it reproduced the pattern of extractive institutions and political absolutism that empowered and enriched a few at the expense of the masses, though the few now were Belgian colonialists, most notably King Leopold II.

When Congo became independent in 1960, the same pattern of economic institutions, incentives, and performance reproduced itself. These Congolese extractive economic institutions were again supported by highly extractive political institutions. The situation was worsened because European colonialism created a polity, Congo, made up of many different precolonial states and societies that the national state, run from Kinshasa, had little control over. Though President Mobutu used the state to enrich himself and his cronies—for example, through the Zairianization program of 1973, which involved the mass expropriation of foreign economic interests—he presided over a noncentralized state with little authority over much of the country, and had to appeal to foreign assistance to stop the provinces of Katanga and Kasai from seceding in the

1960s. This lack of political centralization, almost to the point of total collapse of the state, is a feature that Congo shares with much of sub-Saharan Africa.

The modern Democratic Republic of Congo remains poor because its citizens still lack the economic institutions that create the basic incentives that make a society prosperous. It is not geography, culture, or the ignorance of its citizens or politicians that keep the Congo poor, but its extractive economic institutions. These are still in place after all these centuries because political power continues to be narrowly concentrated in the hands of an elite who have little incentive to enforce secure property rights for the people, to provide the basic public services that would improve the quality of life, or to encourage economic progress. Rather, their interests are to extract income and sustain their power. They have not used this power to build a centralized state, for to do so would create the same problems of opposition and political challenges that promoting economic growth would. Moreover, as in much of the rest of sub-Saharan Africa, infighting triggered by rival groups attempting to take control of extractive institutions destroyed any tendency for state centralization that might have existed.

The history of the Kingdom of Kongo, and the more recent history of the Congo, vividly illustrates how political institutions determine economic institutions and, through these, the economic incentives and the scope for economic growth. It also illustrates the symbiotic relationship between political absolutism and economic institutions that empower and enrich a few at the expense of many.

GROWTH UNDER EXTRACTIVE POLITICAL INSTITUTIONS

Congo today is an extreme example, with lawlessness and highly insecure property rights. However, in most cases such extremism would not serve the interest of the elite, since it would destroy all economic incentives and generate few resources to be extracted. The central thesis of this book is that economic growth and prosperity are associated with inclusive economic and political institutions, while extractive institutions typically lead to stagnation and poverty. But this implies neither that

extractive institutions can never generate growth nor that all extractive institutions are created equal.

There are two distinct but complementary ways in which growth under extractive political institutions can emerge. First, even if economic institutions are extractive, growth is possible when elites can directly allocate resources to high-productivity activities that they themselves control. A prominent example of this type of growth under extractive institutions was the Caribbean Islands between the sixteenth and eighteenth centuries. Most people were slaves, working under gruesome conditions in plantations, living barely above subsistence level. Many died from malnutrition and exhaustion. In Barbados, Cuba, Haiti, and Jamaica in the seventeenth and eighteenth centuries, a small minority, the planter elite, controlled all political power and owned all the assets, including all the slaves. While the majority had no rights, the planter elite's property and assets were well protected. Despite the extractive economic institutions that savagely exploited the majority of the population, these islands were among the richest places in the world, because they could produce sugar and sell it in world markets. The economy of the islands stagnated only when there was a need to shift to new economic activities, which threatened both the incomes and the political power of the planter elite.

Another example is the economic growth and industrialization of the Soviet Union from the first Five-Year Plan in 1928 until the 1970s. Political and economic institutions were highly extractive, and markets were heavily constrained. Nevertheless, the Soviet Union was able to achieve rapid economic growth because it could use the power of the state to move resources from agriculture, where they were very inefficiently used, into industry.

The second type of growth under extractive political institutions arises when the institutions permit the development of somewhat, even if not completely, inclusive economic institutions. Many societies with extractive political institutions will shy away from inclusive economic institutions because of fear of creative destruction. But the degree to which the elite manage to monopolize power varies across societies. In some, the position of the elite could be sufficiently secure that they may permit some

moves toward inclusive economic institutions when they are fairly certain that this will not threaten their political power. Alternatively, the historical situation could be such as to endow an extractive political regime with rather inclusive economic institutions, which they decide not to block. These provide the second way in which growth can take place under extractive political institutions.

The rapid industrialization of South Korea under General Park is an example. Park came to power via a military coup in 1961, but he did so in a society heavily supported by the United States and with an economy where economic institutions were essentially inclusive. Though Park's regime was authoritarian, it felt secure enough to promote economic growth, and in fact did so very actively—perhaps partly because the regime was not directly supported by extractive economic institutions. Differently from the Soviet Union and most other cases of growth under extractive institutions, South Korea transitioned from extractive political institutions toward inclusive political institutions in the 1980s. This successful transition was due to a confluence of factors.

By the 1970s, economic institutions in South Korea had become sufficiently inclusive that they reduced one of the strong rationales for extractive political institutions—the economic elite had little to gain from their own or the military's dominance of politics. The relative equality of income in South Korea also meant that the elite had less to fear from pluralism and democracy. The key influence of the United States, particularly given the threat from North Korea, also meant that the strong democracy movement that challenged the military dictatorship could not be repressed for long. Though General Park's assassination in 1979 was followed by another military coup, led by Chun Doo-hwan, Chun's chosen successor, Roh Tae-woo, initiated a process of political reforms that led to the consolidation of a pluralistic democracy after 1992. Of course, no transition of this sort took place in the Soviet Union. In consequence, Soviet growth ran out of steam, and the economy began to collapse in the 1980s and then totally fell apart in the 1990s.

Chinese economic growth today also has several commonalities with both the Soviet and South Korean

experiences. While the early stages of Chinese growth were spearheaded by radical market reforms in the agricultural sector, reforms in the industrial sector have been more muted. Even today, the state and the Communist Party play a central role in deciding which sectors and which companies will receive additional capital and will expand—in the process, making and breaking fortunes. As in the Soviet Union in its heyday, China is growing rapidly, but this is still growth under extractive institutions, under the control of the state, with little sign of a transition to inclusive political institutions. The fact that Chinese economic institutions are still far from fully inclusive also suggests that a South Korean–style transition is less likely, though of course not impossible.

It is worth noting that political centralization is key to both ways in which growth under extractive political institutions can occur. Without some degree of political centralization, the planter elite in Barbados, Cuba, Haiti, and Jamaica would not have been able to keep law and order and defend their own assets and property. Without significant political centralization and a firm grip on political power, neither the South Korean military elites nor the Chinese Communist Party would have felt secure enough to manufacture significant economic reforms and still manage to cling to power. And without such centralization, the state in the Soviet Union or China could not have been able to coordinate economic activity to channel resources toward high productivity areas. A major dividing line between extractive political institutions is therefore their degree of political centralization. Those without it, such as many in sub-Saharan Africa, will find it difficult to achieve even limited growth.

Even though extractive institutions can generate some growth, they will usually not generate sustained economic growth, and certainly not the type of growth that is accompanied by creative destruction. When both political and economic institutions are extractive, the incentives will not be there for creative destruction and technological change. For a while the state may be able to create rapid economic growth by allocating resources and people by fiat, but this process is intrinsically limited. When the limits are hit, growth stops, as it did in the Soviet Union in the

1970s. Even when the Soviets achieved rapid economic growth, there was little technological change in most of the economy, though by pouring massive resources into the military they were able to develop military technologies and even pull ahead of the United States in the space and nuclear race for a short while. But this growth without creative destruction and without broad-based technological innovation was not sustainable and came to an abrupt end.

In addition, the arrangements that support economic growth under extractive political institutions are, by their nature, fragile—they can collapse or can be easily destroyed by the infighting that the extractive institutions themselves generate. In fact, extractive political and economic institutions create a general tendency for infighting, because they lead to the concentration of wealth and power in the hands of a narrow elite. If another group can overwhelm and outmaneuver this elite and take control of the state, they will be the ones enjoying this wealth and power. Consequently, as our discussion of the collapse of the later Roman Empire and the Maya cities will illustrate (this page and this page), fighting to control the all-powerful state is always latent, and it will periodically intensify and bring the undoing of these regimes, as it turns into civil war and sometimes into total breakdown and collapse of the state. One implication of this is that even if a society under extractive institutions initially achieves some degree of state centralization, it will not last. In fact, the infighting to take control of extractive institutions often leads to civil wars and widespread lawlessness, enshrining a persistent absence of state centralization as in many nations in sub-Saharan Africa and some in Latin America and South Asia.

Finally, when growth comes under extractive political institutions but where economic institutions have inclusive aspects, as they did in South Korea, there is always the danger that economic institutions become more extractive and growth stops. Those controlling political power will eventually find it more beneficial to use their power to limit competition, to increase their share of the pie, or even to steal and loot from others rather than support economic progress. The distribution and ability to exercise power will ultimately undermine the very foundations of economic prosperity, unless political institutions are transformed from

extractive to inclusive.

What happened and why?

The anatomy of crisis*

. . . and the failure of policy

Milton and Rose D. Friedman

The Depression that started in the U.S. in mid-1929 was a catastrophe of unprecedented dimensions for the U.S.: by 1933, the dollar income of the nation had been halved, total output had been cut by a third, and one of every four potential workers was recorded as unemployed. It was no less a catastrophe for the world. The spread of the Depression to other countries brought lower output, higher unemployment, hunger, and misery everywhere. In Germany, the Depression helped Adolf Hitler rise to power and paved the way for World War II. In Japan, it strengthened the hold of the military clique dedicated to creating a Greater East Asia co-prosperity sphere. In China, the aftermath of the Depression destroyed the monetary system, weakened the ability of the Nationalist government to resist the Japanese and then the Communists, and fostered the final hyperinflation that sealed the doom of the Chiang Kai-shek regime and elevated Mao to power.

In the realm of ideas, the Depression persuaded the public at large that Karl Marx was right in condemning capitalism as a fundamentally unstable system given to ever more serious crises. It converted the public to the view that had earlier gained increasing acceptance among the intellectuals — that government had to play a more active role; that it should intervene actively to offset instability generated by private enterprise; that it should become a balance wheel promoting stability and assuring the security of its citizens. The change in the public's perception of the role of the market, on the one hand, and of the government, on the other, was a major catalyst for the rapid growth of government, and particularly central government, from that day to this.

The Depression produced an equally drastic change in professional economic opinion. It shattered the long-held belief, which had been strengthened during the 1920's, that monetary policy was a potent instrument for promoting economic stability. Opinion shifted almost to the opposite extreme, that "money does not matter." To fill the gap left by the apparent collapse of the reigning theory, the most brilliant economist of the twentieth century, John Maynard Keynes, offered an alternative theory, launching the Keynesian revolution, which not only captured the economics profession, but provided both an appealing justification and a prescription for extensive government intervention.

Both shifts — in the opinion of the public and of the economics profession — arose from a misunderstanding of what had actually happened. *We now know, as a few knew then, that the Depression reflected a failure of government, not of private enterprise.* And it reflected a failure of government in an area in which the government had long been assigned responsibility — "to coin money, regulate the value thereof, and of foreign coin." The Federal Reserve System, the key monetary authority at the time, imposed a crushing burden on the economy. Its policies produced or facilitated a decline in the quantity of money by one-third from 1929 to 1933. Established in 1913 in response to the panic of 1907, precisely in order to prevent similar episodes, it stood idly by while over one-third of the commercial banks of the nation went out of existence. It presided over a banking panic far more extensive and damaging than any that had ever occurred earlier, ending up succumbing to its own ineptness by closing its own doors for a week during the so-called banking holiday of March 1933.

At the time, and for a considerable period thereafter, the bank failures and subsequent banking panic were interpreted by many knowledgeable ob-

* Excerpted from FREE TO CHOOSE: *A Personal Statement*, by Milton Friedman and Rose D. Friedman, to be published by Harcourt Brace Jovanovich, Inc. Copyright © 1979 by Milton Friedman and Rose D. Friedman.

servers as having occurred despite the best efforts of the Federal Reserve to ease monetary conditions and to expand the money supply. Only much later did research demonstrate beyond doubt that the facts were quite different. At all times from 1929 to 1933, the Federal Reserve had the power to prevent any decline in the quantity of money, indeed, to expand the money supply to any desired extent. Throughout the Depression, there were persons within the System, as well as outside, calling for the Fed to take the needed action. It was conflict within the System, inertia, drift, and incompetence, not impotence that produced the disastrous failure of monetary policy.

On the scientific side, we now know that the Depression, far from showing that "money does not matter," was a tragic testimonial to the importance of money. Of course, many factors other than monetary policy affected the detailed course of the Depression and help to explain its severity and duration. But it is literally inconceivable that the Depression could have lasted as long as it did or have been as severe as it was if the Fed had acted early to prevent a decline in the quantity of money.

This conclusion would be endorsed today by the vast majority of economists of all shades of professional and political opinion — but was not known to Keynes or most of his contemporaries.

THE ORIGIN OF THE FEDERAL RESERVE SYSTEM

On Monday, October 21, 1907, some five months after the beginning of an economic recession, the Knickerbocker Trust Company, the third largest trust company in New York, began to experience financial difficulties. The next day a "run" on the bank forced it to close (temporarily, as it turned out: it resumed business in March 1908). The closing of the Knickerbocker Trust precipitated runs on other trust companies in New York and then spread to other parts of the country — a banking "panic" was under way of the kind that had occurred every now and then during the nineteenth century.

Within a week, banks throughout the country reacted to the "panic" by "restriction of payments," i.e., the convertibility of deposits into currency — a move that was legally sanctioned in a few states, and tolerated without explicit sanction in the rest.

The restriction of payments did cut short bank failures and end the panicky runs. But it imposed serious inconvenience on business, led to a shortage of coin and currency, and stimulated wooden nickels and all sorts of other temporary substitutes for legal money — for a time it took $104 of deposits to buy $100 of currency. Together, the panic and the restriction, both directly and by forcing a decline in the quantity of money, sharply intensified the recession under way, turning it into one of the most severe that the U.S. had experienced up to that time.

The severe phase was short-lived, however. Banks resumed payments in early 1908, and a few months thereafter economic recovery got under way. The recession lasted in all only 13 months.

This dramatic episode was the key element that accounted for the establishment of the Federal Reserve System in 1913. It made some action in the monetary and banking area politically essential. In the Republican Administration of Theodore Roosevelt, a National Monetary Commission was established, headed by a prominent Republican Senator, Nelson W. Aldrich. In the Democratic Administration of Woodrow Wilson, the Commission's recommendations were rewritten and repackaged by a prominent Democratic Senator, Carter Glass, and enacted as the Federal Reserve Act of 1913.

But what do the terms "run" and "panic" and "restriction of payments" really mean? Why did they have the effects they did? And how did the Federal Reserve Act propose to prevent similar episodes?

A run on a bank is simply an attempt by many of its depositors simultaneously to "withdraw" their deposits in cash. It arises from a fear that the bank will fail. It represents an attempt by everyone to get "his" money out before it is all gone.

One bank alone can meet a "run" by borrowing from other banks, or by asking its borrowers to repay their loans — which they may be able to do by withdrawing cash from other banks. But if a bank run spreads, all banks together obviously cannot meet the run in this way — there simply is not enough currency in bank vaults to satisfy the demands of all depositors.

Moreover, any attempt to meet a widespread run by drawing down vault cash — unless it succeeds promptly in restoring confidence and ends the run so the cash is redeposited — will force a much larger reduction in deposits. On the average in 1907, there were $8 of deposits for every $1 of cash in the vaults of banks. For every $1 transferred from the vaults of banks to the mattresses of depositors, deposits had to go down by roughly $8. That is why a run — hoarding of cash by the public — tends to reduce the total money supply.

It is also why we call cash, or its equivalent, "high-powered money." It is also why a run, if not checked, causes such distress. Individual banks, seeking to get cash to meet the demands of their depositors, try to get their borrowers to repay loans, or refuse to renew loans or to extend additional loans — but the borrowers as a whole have nowhere to turn, so banks fail and businesses fail.

How can a "panic" be stopped once it is under way, or better yet, how can it be prevented from starting? One way to stop a panic is the method adopted in 1907: a concerted restriction of payments by the banks. Banks agreed with one another that they would not pay cash on demand to depositors. They stayed open for business by accepting checks on themselves and other banks as "deposits," and settling among themselves only "through the clearing house." That is, they operated through bookkeeping entries — a primitive version of the cashless society of the future that so many expect to develop. Under this system, banks might and did still fail because they were "unsound" banks, but they did not fail simply because they could not promptly convert their perfectly sound assets into cash.

That is a rather drastic way but it worked. As time passed, panic subsided, confidence in banks was restored, the banks resumed payment, and shortly thereafter the recession came to an end and recovery followed.

Another way to stop a panic is to enable sound banks to convert their assets into cash rapidly, not at the expense of other banks but through the creation of additional cash — to provide an emergency printing press, as it were. In principle, if that way worked, it would prevent even the temporary disruptions produced by the restriction of payments. That was the way embodied in the Federal Reserve Act. The 12 regional banks established by that Act, operating under the supervision of a Federal Reserve Board in Washington, were given the power in effect to print money in order that they could serve as "lenders of last resort" to the commercial banks. Initially, it was expected that they would operate mostly by direct loans ("rediscounts") to banks. Subsequently, "open-market operations" — the purchase or sale of government bonds — became the main way in which the System added to or subtracted from the amount of cash — the purchases being financed by creating new cash or its equivalent; the sales, by withdrawing cash or its equivalent from the system.

After the Federal Reserve System failed so miserably in the early 1930's to do what it had been set up to do, an effective method of preventing a panic from starting was finally adopted in 1934 — the Federal insurance of bank deposits. By giving depositors confidence that they were guaranteed against loss, it prevented the failure or financial difficulties of an unsound bank from spreading the contagion to other banks — the people in the crowded theatre were confident that it was really fireproof. Since 1934, there have been bank failures and some runs on individual banks, but no banking panics of the old style.

This method of preventing a panic had frequently been used earlier — though in a far more partial and less effective version — by the banks themselves. Time and again, when an individual bank was in financial trouble, or threatened by a run because of rumors of trouble, other banks banded together voluntarily to subscribe to a fund guaranteeing the assets of the bank in trouble. That device prevented many putative panics and cut short others. On still other occasions it failed, either because a satisfactory agreement could not be reached or because confidence was not promptly restored. We shall examine a particularly dramatic and important case of failure below.

THE ONSET OF DEPRESSION

In the popular view, the Depression started on Black Thursday, October 25, 1929, when the New York stock market collapsed, the beginning of a slide that left stock prices in 1933 at only about one-sixth their dizzying level in 1929.

The stock market crash was important, but it was a late comer. Business activity reached its peak in August 1929 and had already fallen appreciably before the crash. In fact, the crash simply reflected the emerging economic contraction. But, of course, once it occurred, it helped to deepen the contraction. It spread uncertainty among businessmen and others, who had been bemused by dazzling hopes of a new era. It dampened the willingness of both consumers and business entrepreneurs to spend, and enhanced their desire to strengthen their liquid reserves for emergencies.

These depressing effects of the stock market crash were strongly reinforced by the early fruits of the struggle for power within the Federal Reserve System. At the time of the crash itself, the New York Federal Reserve Bank, almost by conditioned reflex instilled during the era of its previous head, Benjamin Strong, immediately acted on its own to cushion the shock by purchasing government securities. But Strong *was* dead, and the Board regarded this action as smacking of insubordination. It moved rapidly to impose its discipline on New York, and New York yielded.

The result was that thereafter, the System acted very differently than it had under Strong during earlier economic recessions in the 1920's. Instead of actively expanding the money supply by more than the usual amount to offset the contraction, the System allowed the money supply actually to decline slowly throughout 1930. Compared to the collapse from late 1930 to early 1933, the decline in the stock of money up to October 1930 seems mild — a mere 2.6%. But by comparison with past episodes, it was sizable — larger than during the whole of most earlier recessions.

The combined effect was a rather severe reces-

sion. Even if the recession had come to an end in late 1930 or early 1931, as it might have done in the absence of the monetary collapse that was to ensue, it would have ranked as one of the most severe recessions on record.

BANKING CRISES

But the worst was yet to come. Until the fall of 1930, the contraction, though severe, had been a garden-variety recession, unmarred by banking difficulties, runs on the banks, or the like. The character of the recession then changed drastically, as a series of bank failures in the Middle West and South undermined confidence in banks and led to widespread attempts to convert deposits into currency.

The contagion finally spread to New York, the financial center of the country. The critical date is December 11, 1930, when the Bank of United States closed its doors — the largest commercial bank ever to have failed up to that time in U.S. history, and a bank moreover, that, although an ordinary commercial bank, had a name that led many at home and abroad to regard it as an official bank. Its failure was therefore a particularly serious blow to confidence.

It is something of an accident that this particular bank played such a key role. It was an accident that the Bank of United States happened to be the particular big bank in a major financial center that failed. Given the structure of the U.S. banking system, plus the policy of drift and indecision that the Federal Reserve System was following, if it had not failed when it did, some other major bank in a major financial center would sooner or later have done so, and its failure would have had similar effects on confidence in banks. But it was also an accident that the Bank of United States itself failed. It was fundamentally a perfectly sound bank. Though liquidated during the worst years of the Depression, it ended up paying off depositors 92.5¢ on the dollar. There is little doubt that if it had been able to continue as an ongoing business, no depositor would have lost a cent.

In the standard pattern of earlier crises, when rumors started to spread about the Bank of United States, efforts were made by the New York State Superintendent of Banking, the Federal Reserve Bank of New York, and the New York Clearing House Association of Banks to devise plans to save the bank through providing a guarantee fund, or merging it with other banks. Until two days before the bank closed, this effort seemed assured of success.

The effort finally failed primarily because of the particular character of the bank plus the prejudices of the banking community. The name itself, with its appeal to immigrants, was resented by other banks. Far more important, the bank was owned and managed by Jews, and served mostly the Jewish community. It was one of a handful of Jewish owned banks, in an industry that, more than almost any other, has been the preserve of the well-known and well-bred. By no accident, the final rescue plan involved merging the Bank of United States with the only other major bank in New York that was largely owned and run by Jews, plus two much smaller banks that had a similar ethnic character.

The plan failed because the New York Clearing House at the last moment withdrew from the proposed arrangement — purportedly in large part because of the anti-Semitism of some of the leading members of the banking community. At the final meeting of the bankers, Joseph A. Broderick, then the New York State Superintendent of Banking, tried but failed to get them to go along. "I said," he later testified at a court trial,

> "it [the Bank of United States] had thousands of borrowers, that it financed small merchants, especially Jewish merchants, and that its closing might and probably would result in widespread bankruptcy among those it served. I warned that its closing would result in the closing of at least ten other banks in the city and that it might even affect the savings banks. The influence of the closing might even extend outside the city, I told them.
>
> I reminded them that two or three weeks before they had rescued two of the largest private bankers of the city and had willingly put up the money needed. I recalled that only seven or eight years before that they had come to the aid of one of the biggest trust companies in New York, putting up many times the sum needed to save the Bank of United States but only after some of their heads had been knocked together.
>
> I asked them if their decision to drop the plan was still final. They told me it was. Then I warned them that they were making the most colossal mistake in the banking history of New York."[1]

For the owners and depositors of the Bank of United States, the closing was tragic. The depositors had their funds tied up for years, and never recovered all of them; two of the owners were tried in court, convicted, and served prison sentences for what everybody agreed were technical infractions of the law.

For the country as a whole, the effects were even more far-reaching. Depositors all over the country, frightened about the safety of their deposits, added to the sporadic runs that had started earlier.

1. Footnotes appear at the end of the article.

Banks failed by the droves — 352 banks in the month of December 1930 alone.

Had the Federal Reserve System never been established, and had a similar series of runs started, there is little doubt that they would have been met as the 1907 panic was — by a restriction of payments. That would have been a more drastic measure than any actually taken in the final months of 1930 but, by cutting the vicious circle set in train by the search for liquidity, restriction would almost certainly have prevented the subsequent series of bank failures in 1931, 1932, and 1933, just as restriction in 1907 quickly ended bank suspensions then. Indeed, the Bank of United States itself might have been able to reopen, as the Knickerbocker Trust Company did in 1908. The panic over, confidence restored, economic recovery would very likely have begun in early 1931, just as it did in early 1908.

As it was, the existence of the Reserve System prevented this drastic therapeutic measure: directly, by reducing the concern of the stronger banks, who, mistakenly as it turned out, were confident that borrowing from the System offered them a reliable escape mechanism in case of difficulty; indirectly, by lulling the community as a whole, and the banking system in particular, into the belief that such drastic measures were no longer necessary now that the System was there to take care of such matters.

The System could have provided a far better solution by engaging in large-scale open market purchases, thereby providing banks with cash to meet the demands of their depositors. That would have both ended — or at least sharply reduced — the stream of bank failures and would have prevented the public's attempted conversion of deposits into currency from reducing the quantity of money. But unfortunately, the Fed's actions were hesitant and small. In the main it stood idly by, and let the crisis take its course — a pattern of behavior that was to be repeated again and again during the next two years.

It was repeated in the spring of 1931, when a second banking crisis developed. An even more perverse policy was followed in September 1931, when Britain abandoned the gold standard. The Fed reacted — after two years of severe depression — by taking the most deflationary measures in its history, imposing yet another monetary blow on a struggling economy.

In 1932, under strong pressure from Congress, the Fed finally undertook large-scale open market purchases. The favorable effects were just starting to be felt when Congress adjourned — and the Fed promptly terminated its program.

The final episode in this sorry tale was the banking panic of 1933, once again initiated by a series of bank failures, and intensified by the interregnum between Herbert Hoover and Franklin D. Roosevelt, who was elected on November 8, 1932, but not inaugurated until March 4, 1933. Herbert Hoover was unwilling to take drastic measures without the cooperation of the President-elect, and FDR was unwilling to assume any responsibility until he was inaugurated.

As panic spread in the New York financial community, the System itself was infected. The head of the New York Federal Reserve Bank tried to get President Hoover to declare a national banking holiday on his last day in office; failing in that attempt, he joined with the New York Clearing House Banks and the State Superintendent of Banking to persuade Governor Lehman of New York to declare a state banking holiday effective on March 4, 1933, the day of FDR's inauguration — the Federal Reserve Bank closing along with the commercial banks. Similar actions were taken by other governors. A nationwide holiday was finally proclaimed by President Roosevelt on March 6.

The central banking system, set up primarily to render impossible the restriction of payments by commercial banks, itself joined the commercial banks in a more widespread, complete, and economically disturbing restriction of payments than had ever been experienced in the history of the country. One can certainly sympathize with Hoover's comment in his memoirs: "I concluded [the Reserve Board] was indeed a weak reed for a nation to lean on in time of trouble."[2]

At the peak of business in mid-1929, nearly 25,000 commercial banks were in operation in the United States. By early 1933, the number had shrunk to 18,000. When the holiday was ended by President Roosevelt ten days after it began, fewer than 12,000 banks were permitted to open, and only 3,000 additional banks were later permitted to do so. All in all, therefore, roughly 10,000 out of 25,000 banks disappeared during those four years — through failure, merger, or liquidation.

The total stock of money showed an equally drastic decline. For every $3 of deposits and currency in the hands of the public in 1929, only $2 remained in 1933: a monetary collapse without precedent.

FACTS AND INTERPRETATION

These facts are not in question — though it should be stressed that they were not known or available to many contemporary observers, including John Maynard Keynes. The real issues are of interpretation. Was the monetary collapse a cause of the economic collapse or a result? Could the System have prevented the monetary collapse? Or did it happen in spite of the best

efforts of the Fed — as so many observers at the time concluded? Did the Depression start in the U.S., and spread abroad? Or did forces emanating from abroad convert what might have been a fairly mild recession in the United States into a severe one?

The System itself expressed no doubt about its role. So great is the capacity for self-justification that the Federal Reserve Board could write in its annual report for 1933, "The ability of the Federal Reserve Board to meet enormous demands for currency during the crisis demonstrated the effectiveness of the country's currency system under the Federal Reserve Act. . . . It is difficult to say what the course of the depression would have been had the Federal Reserve System not pursued a policy of liberal open market purchases." "Oh, what a tangled web we weave, when first we practice to deceive" — ourselves, in this case.

On cause and effect, there is little doubt that the monetary collapse was both. It had partly independent origins in Federal Reserve policy and unquestionably made the economic collapse far worse than it would have been; but also, once the economic collapse started, it intensified the monetary collapse. Bank loans that might have been "good" in a milder recession became "bad" loans in the severe collapse that occurred, weakening the lending banks and encouraging depositors to start a run on them. Failures of business enterprises, growing unemployment, all fostered uncertainty and fear, and a desire to convert assets into the most liquid form. "Feedback" is a pervasive feature of an economic system.

On the System's power to prevent the monetary collapse, the evidence by now is all but conclusive that it clearly had the power to do so. Defenders of the System have offered a series of excuses — but none has proved a defensible explanation of the failure of the System to perform the task its founders had established it to perform.

Moreover, the System not only had the power, it also had the knowledge required to exercise that power. In 1929, 1930, and 1931, the New York Federal Reserve Bank repeatedly urged the System to engage in open market purchases, the key action the System should have taken but did not. New York was overruled, not because persuasive evidence was presented that its proposals were not feasible, but on very different grounds, all stemming basically from the struggle for power within the System and confused, indecisive leadership by the Board in Washington. Outside the System, there were also knowledgeable voices calling for the right action. An Illinois congressman, A. J. Sabath, said, on the floor of the House, "I insist it is within the power of the Federal Reserve Board to relieve the financial and commercial distress." Some academic critics expressed similar views — including one who later became the head of one of the Federal Reserve Banks. As already noted, the only important departure from the Fed's passive policy — in 1932 — occurred under direct pressure from the Congress.[3]

On the international character of the Depression, the decisive evidence that it spread from the U.S. to the rest of the world comes from the movements of gold. In 1929, the U.S. was on a gold standard in the sense that there was an official price of gold ($20.67 per fine ounce) at which the U.S. government would buy any gold offered or sell anyone gold on demand in return for U.S. currency or its equivalent. Most other major countries were on a so-called gold-exchange standard, under which they might or might not buy and sell literal gold freely, but under which they specified an official price for gold in terms of their own currencies, and undertook to keep the price of their currency in terms of the dollar fixed at the level determined by the two official prices of gold. Under such a system, if the United States spent (or lent or gave) abroad more dollars than the recipients of those dollars wanted to spend (or lend or give) in the U.S., the difference would come back to the United States in the form of a demand for gold. The U.S. would have a net "outflow" of gold, it would "lose" gold — in the technical jargon. Conversely, it would "gain" gold.

Suppose now that the Depression had originated abroad while the U.S. economy continued, for a time, to boom. An early effect would be a decline in foreign purchases of U.S. goods and an increase in U.S. purchases of foreign goods — as the worsening economic conditions abroad reduced the cost, or increased the availability of foreign goods. The effect would be an excess of dollars spent abroad and an outflow of gold from the U.S. Such an outflow of gold would have reduced the Federal Reserve System's "gold reserves" and have induced it to take action to reduce the quantity of money. That is the way in which, in a system of fixed exchange rates, deflationary (or inflationary) pressure is transmitted from one country to another. Had this been the course of events, the Federal Reserve could correctly claim that its actions were a response to pressures coming from abroad.

Conversely, if the Depression originated in the United States, an early effect would have been a decline in U.S. purchases abroad and an increase in U.S. sales abroad, and hence an inflow of gold. This would have brought pressure on foreign countries to reduce the quantity of money and would have been the way the U.S. deflation would have been transmitted to them.

The facts are crystal clear: the U.S. gold stock

rose from August 1929 to August 1931, the first two years of the contraction — clinching evidence that the United States was in the van of the movement. Had the System followed the rules of the gold standard, it should have reacted to the inflow of gold by expanding the money supply instead of contracting it, as it actually did.

Of course, once the Depression was under way and had been transmitted to other countries, what happened then had a reflex influence on the United States — another example of the feedback that is so ubiquitous in any complex economy. The country in the vanguard of an international movement need not stay there. France, which had accumulated a large stock of gold as a result of returning to the gold standard in 1928 at an exchange rate that undervalued the franc and therefore had much leeway, at some point passed the United States and not only began to add to its gold stock but also, after late 1931, to drain gold from the United States. Its dubious reward for such leadership was that, although the U.S. economy hit bottom when it suspended gold payments in March 1933, the French economy did not hit bottom until April 1935.

CONCLUSION AND AFTERMATH

One ironic result of the inept monetary policy fashioned by the Federal Reserve Board against the advice of the New York Federal Reserve Bank was a complete victory for the Board against both New York and the other Federal Reserve Banks in the struggle for power. The myth that private enterprise, including the private banking system, had failed, and that government needed more power to counteract the alleged inherent instability of the free market, meant that the System's failure produced a political environment favorable to giving the Board greater control over the regional banks.

One symbol of the change was the transfer of the Federal Reserve Board from modest offices in the U.S. Treasury Building to a magnificent Greek temple of its own on Constitution Avenue (since supplemented by a massive additional structure).

The final seal on the shift of power was a change in the name of the Board and in the title of the head officers of the regional banks. In central bank circles, the prestigious title is governor, not president. From 1913 to 1935, the head of a regional bank was designated "governor"; the central Washington body was called "The Federal Reserve Board"; only the chairman of the Board was designated "governor"; the remaining members were simply "members of the Federal Reserve Board." The Banking Act of 1935 changed all that. The heads of the regional banks were put in their place by being designated "presidents" instead of "governors"; and the compact "Federal Reserve Board" was replaced by the cumbrous "Board of Governors of the Federal Reserve System," solely in order that each of the members of the Board could be designated a "governor."

Unfortunately, the increase in power, prestige, and trappings of office have been accompanied by no corresponding improvement in performance. Since 1935, the System has presided over — and greatly contributed to — a major recession in 1937, a wartime and immediate postwar inflation, and a roller coaster economy since, with alternate rises and falls in inflation, and decreases and increases in unemployment. Each inflationary peak and each temporary inflationary trough has been at a higher and higher level, and with a gradual increase in the average level of unemployment.

The System has not made the same mistake that it made in 1929-1933 — of permitting or fostering a monetary collapse — but it has made the opposite mistake, of fostering an unduly rapid growth in the quantity of money and so promoting inflation. In addition, it has continued, by swinging from one extreme to another, to produce not only booms but also recessions, some mild, some sharp.

In one respect, the System has remained completely consistent throughout: in blaming all problems on external influences beyond its control and taking credit for any and all favorable occurrences. It thereby continues to promote the myth that the private economy is unstable, while its behavior continues to document the reality that government is today the major source of economic instability.

[1] Milton Friedman and Anna J. Schwartz, *A Monetary History of the United States, 1867-1960* (Princeton: Princeton University Press, 1963), p. 310.

[2] *Memoirs*, p. 212.

[3] For a fuller discussion, see Friedman and Schwartz, *Monetary History*, pp. 391-419.