Jean Monnet Chair Papers

[8]

The Economics and Politics of Deregulation

Roger G. Noll

The European Policy Unit at the
European University Institute

Jean Monnet Chair Papers

Noll, *The Economics and Politics of Deregulation*

The Jean Monnet Chair

The Jean Monnet Chair was created in 1988 by decision of the Academic Council of the European University Institute, with the financial support of the European Community. The aim of this initiative was to promote studies and discussion on the problems, internal and external, of European Union following the Single European Act, by associating renowned academics and personalities from the political and economic world to the teaching and research activities of the Institute in Florence.

The Economics and Politics of Deregulation

Roger G. Noll

1991

The European Policy Unit at the
European University Institute

Printed in Italy in October 1991
European University Institute
Badia Fiesolana
I-50016 San Domenico (FI)
Italy

Lecture #1 – Theories of Regulation*

The twenty years between 1970 and 1990 witnessed massive changes in the relationship between business and government in all advanced industrialized democracies. In the U.S., two roughly parallel but opposite trends occurred: a substantial scaling back of economic regulation (the control of prices, profits and entry in infrastructural industries) accompanied by a dramatic expansion of environmental, health and safety regulation. In Japan, strong centralized planning combined with nationalized operation of infrastructural industries slowly gave way to liberalization, privatization, and competition.

In Europe, pursuit of economic integration within the European Economic Communities led to a movement to reduce regulatory barriers to intra-European trade, but the less ambitious notion of "harmonization" gave rise to a more ambitious set of policies paralleling developments in Japan and the U.S. In infrastructural industries, European integration has meant not just coordination among national monopolies, but increasing liberalization of hitherto monopolized and heavily regulated markets. And harmonization alone does not explain the growing role of the Commission of the European Communities as a regulator in the environmental, health and safety field. Indeed, the Single Europe Act, through such measures as new Articles 130 R, S and T of the Treaty of Rome, strongly institutionalizes regulatory functions in the EEC when the multinational organization is more effective at attaining the goal of protection of the environment and public health, regardless of the connectedness of the ensuing Regulations and Directives to facilitating trade among the Member States.

Many explanations have been put forth for both the rise of protective regulation and the fall of economic regulation, but the most common explanations for each are mutually inconsistent. That is, economic deregulation is said to be associated with a new "conservative ideology", such as Reaganism in the U.S. or Thatcherism in the U.K., whereas environmental, worker and consumer protectionism is associated with growing support for state intervention to protect the public from abuses by managers of the institutions of economic production (usually capitalistic, but some-

* The four lectures comprising this Jean Monnet Paper were given as seminars by Roger G. Noll, within the framework of the Jean Monnet Chair of the European Policy Unit, to staff and researchers of the European University Institute in May of 1991.

times "out of control" public companies). Another common explanation, a softer version of the concept of ideological shift, is the power of new ideas developed during the 1960s. Within economics, the new idea, arising from extensive empirical studies of economic regulation, was that economic regulation was a sham: rather than protecting consumers from monopolies, it harmed them by creating inefficient providers of service who used their monopoly position to feather their own nests. Elsewhere, the new idea was scientific information about the harms arising from industrialization through environmental degradation (e.g., Rachel Carson's *Silent Spring*) or through unsafe products and workplaces (e.g., Ralph Nader's *Unsafe at Any Speed*).

The difficulty with these accounts is that they are superficial. Ideological conflict between interventionism and liberalism is as old as democracy, and the rise of environmental and safety regulation falsifies the notion that somehow, around 1970, liberalism gained an upper hand. Moreover, the "new ideas" of the 1960s were not very new, and in any case were often conflicting. The advocates of economic deregulation won only a partial victory, for in no country was the degree of economic deregulation as extensive as the economic research concluded was warranted. In other regulatory arenas, economic advice was largely ignored, especially in the environmental field. Likewise, the advocates of protective regulation surely did not win a resounding victory, for the measures adopted fell far short of their goals, and in many instances the institutional mechanisms for achieving even the scaled-back goals proved ineffective. Indeed, by the early 1980s both economists and advocates of stricter measures to protect the environment and public health could agree that much of the effect of regulation in this area was to protect established industries (see Bruce Ackerman and William Hassler, *Clean Air/Dirty Coal).*

To understand the development of regulatory policy during the reform period requires a comprehensive theory of how and why government policies change. Policy reform, like such prosaic acts as purchasing a product, hiring a worker, or casting a vote, is the result of decisions by individuals, who presumably bring to bear on all of these decisions more or less the same capabilities, values and purposiveness. "Why regulatory reform", therefore, is likely to be answered in much the same way as "Why any policy change", or even "Why any purposive social action?"

Scholars in almost all social scientific disciplines have contributed some part of the answer to these questions. Here I focus on one category of conceptual models to explain the public sector: liberal methodological individualism. This category of social explanation adopts the view that the appropriate unit of observation is a person. Social actions can be understood and evaluated solely by examining the decisions of individuals and the effects of actions on those individuals.

The normative component of liberal methodological individualism is that a state of a society (including a nation) is deemed better or worse solely on the basis of the effects it has on the members of the society, ac-

cording to the values that the members themselves hold. Thus, normative criteria based on the values of only some members (e.g. monarchy, theocracy) or on serving the interests of institutions (e.g. the state) are precluded. The positive, or scientific, component of this approach regards individual members of society as causes, rather than effects, of societal changes. Policy is the product of interactions, sometimes conflictual, among members of society, and the appropriate unit of analysis for understanding, predicting and perhaps even controlling policy is the individuals who seek to affect it.

Liberal methodological individualism is not without critics, but it does constitute the core of much of contemporary social science, especially in economics and psychology, but also to a limited extent in political science, sociology and philosophy. As an economist who dabbles in these other fields, it is the approach that I am competent to take. Hence, the point of this essay is to lay out what liberal theory has to say about policy change, with applications to regulatory reform.

Normative Theories of Government

Although the aim of my analysis is to develop a positive (or causal) theory of regulatory policy change, a useful starting place is normative theory, which asks what policy-makers ought to do, using liberal principles to evaluate the alternatives. Of course, normative theories of what the state ought to do can also be regarded as theories of what the state actually will do, assuming that all relevant individuals (the ones who make decisions that affect policy) apply the principles of a normative theory. Plato's philosopher-king, Burke's other-regarding public servant, and Weber's scientific bureaucrat provide examples of the linkages between normative and positive theories.

Welfare Economics

Because regulatory policy is predominantly an instrument of economic policy in that it is applied to economic actions such as production, transaction, and consumption, the logical starting point for a review of normative theories is welfare economics. The underlying premises of welfare economics are, first, that people evaluate alternative actions in a normatively meaningful way, and second, that these actions can meaningfully be aggregated across individuals. The evaluation principle requires that outcomes be expressed in a common unit of value. That is, one can compare three bananas to three oranges – and to three tons of air pollution, three sprained thumbs from a hazardous workplace device, and three deaths from cancer from a toxic chemical.

Conventionally, the unit of comparison is money, owing primarily to the ubiquity of transactions in money as a means of exchanging things of

value (goods, time and effort, even political support). For every social activity, whether eating a pear, listening to an opera, or setting a regulatory standard, a person is regarded as having a "willingness to pay" (WTP) for that action. The WTP is the maximum amount that a person could be induced to part with to secure the social action. Because this amount could also be used to do many other things (including buy other goods and services), it has normative meaning in that it provides a quantitative measure of the amount of sacrifice of other valuable economic goods and services a person is willing to make to achieve the end in question. Of course, the person may not actually have to pay the WTP to obtain this end. The WTP for a cold beer on a hot day may substantially exceed its price, generating a net positive benefit (WTP minus price) for the person who buys it. Net WTP, or "surplus", then becomes the measure of the increment to welfare arising from the social action. It measures the extent to which a person experiences more benefit than cost.

The premise of welfare economics is that society should take actions that maximize the net WTP of all of its members. For example, benefit-cost analysis seeks to ascertain the size of a public program that maximizes the difference between benefits and costs, where both benefits and costs are calculated using the WTP principle. Benefits are the WTP for the program, while costs are the WTP for using the resources consumed by the program in the best possible alternative way.

An important issue in welfare economics is how to deal with the fact that income strongly influences a person's WTP. One approach takes the view that income distribution should be ignored when evaluating a policy. Income redistribution should be decided as a separate matter, and then once society has concluded how egalitarian it will be, it can move on to deciding the size and scope of the public sector. The other approach takes the view that income distribution and other social policies are inseparable (because the act of having a public sector alters income distribution), so that it ought to be taken into account when evaluating programs. The latter approach leads to weighting the WTP of the poor more heavily.

A cornerstone of welfare economics – its "First Theorem" – is that if markets are perfectly competitive, they allocate goods and services in a manner that maximizes net WTP. Perfect competition means three things: large numbers of buyers and sellers so that no one is the captive of anyone else (and hence no one has monopoly power); perfect information, meaning that everyone is equally informed about prices, product qualities, and the consequences of economic actions; and complete internalization, meaning that in any economic transaction, all of the benefits and costs are experienced by the transacting parties without spillover effects on others.

The First Theorem of Welfare Economics provides the "market failure" theory of regulatory policy. Because perfect competition produces the best possible outcome (given that one is satisfied with income distribution), intervention can improve matters only if the conditions of per-

fect competition are not satisfied. Thus, regulation (if its costs are sufficiently low so as not to offset its benefits) is justified if a market cannot sustain many sellers (natural monopoly), if production or consumption produces third-party effects (environmental degradation), or if some participants in a market are inadequately informed (consumer and worker protection). The last category, of course, turns on regulation being a superior alternative to simply informing people, letting them sue for damages if injured, or otherwise relying on some policy other than regulation to solve the problem.

The regulatory reform era could be explained by the market failure theory if, in the 1960s, events transpired that changed the extent of market failure. If some monopolies stopped being natural, or became otherwise sufficiently less powerful that the amount that they could extract from their customers was less than the cost of regulating them, economic deregulation would ensue. If some consumers and workers became substantially less capable of evaluating products and workplaces, most plausibly because hazards became more numerous and complicated due to growth and technological change, safety regulation might pass the threshold whereby its benefits exceeded its costs. And, if economic growth increased pollution, or if new scientific knowledge increased the best estimates of the consequences of pollution, more vigorous environmental regulation would be enacted.

Ideology and ideas theories of policy change are not in principle incompatible with the welfare economics of market failure. Ideology can be interpreted as the way members of society evaluate both economic and political outcomes. If the children of the 1960s placed higher values on, say, the natural environment in relation to prosaic consumption activities, the WTP for pollution abatement would increase, thereby increasing the optimal scale and scope of environmental policy. Ideas, in the form of new knowledge about how the world works but without an ideological aspect, have the same effect. The idea of a pending ecological catastrophe (e.g. ozone depletion), when presented with evidentiary documentation, raises the expected cost of ozone-depleting activities, and hence the magnitude of the market failure associated with them, and hence the optimal effort to control them. Ideas and ideology provide a way to connect changes in policy to changes in the knowledge base and evaluation principles that underpin WTP, which in turn determines the net benefits of public policies.

Of course, the source of ideas and even ideology can be a discipline other than economics, with its focus on economic valuations through transactions (or, at least, in principle transactions for things that as a practical matter cannot be traded or for which trading is prohibited). For example, implicit in ecological policy analysis is an ideology (personal evaluation mechanism) that gives great weight to maximizing natural genetic diversity on earth. It prescribes putting forth extensive effort to preserve endangered species, and even genetic variation within each

species, to maximize the survival probability of the present array of living organisms in the face of environmental change or the mutation of a more effective predator (whether a virus or a Tyranosaurus Rex). Likewise, regulatory policy has a strong engineering component, and one can easily deduce implicit normative theories from engineering practice. For example, resource and environment policy can be governed by the desire to maximize engineering efficiency (maximize output per unit of resource use, or maximize the extent to which technical economies of scale are captured), ignoring the organizational efficiency or other sources of economic value that might be ignored by this objective. Or, from moral philosophy and ethics, one might ascribe value primarily to the process by which public decisions are made: whether they maximally preserve the dignity and autonomy of those who are affected by policy, and serve to promote justice and fairness as ends rather than as means to maximizing the net output of the public sector.

The commonality among the normative theories thusfar described is that they use evaluation mechanisms that are capable of assigning values to alternative actions so that one can identify the best action. Whereas economic valuation is rooted in actual decisions (market transactions), the principle behind welfare economics is that the same evaluation method can be applied where transactions are rare or nonexistent, with the same normative content ascribed to them. Likewise, the other valuation techniques presume that one can give concrete meaning to more and less, better and worse. Energy-efficiency and technical economies of scale are clearly such notions, as are the number of species and the extent of genetic variation within a species. Dignitary values and procedural justice are more elusive, but their adherents still consider alternative actions as producing more or less in ways that can be observed.

From the perspective of welfare economics, none of these approaches is necessarily hostile to normative economic policy analysis, although in practice they often are. Each can be regarded as simply standing behind someone's WTP. An engineer, therefore, is simply giving greater weight to resource costs (or capturing scale economies) than someone who seeks to minimize economic costs. Thus, the engineering WTP may be peculiar, but it does not raise fundamental problems for finding the policy that maximizes WTP. But problems arise if the engineer insists that the WTPs of others (who are not using engineering evaluation techniques) be ignored. The principle of "*de gustibus no est disputandum*" admits the peculiar tastes of the engineer, but it accords no special status to any particular method of evaluating alternatives. In general, this principle tends to be bad news for engineering evaluation, for research indicates that people are more prone to undervalue than to overvalue technical efficiency, relative to a standard of economic benefits and costs. Consequently, engineering and economic analyses often are in conflict, with engineering analyses usually being more supportive of both economic and protective regulations.

The preceding highlights the difference between welfare economics and these other normative theories. Welfare economics adopts the position that values are derived by aggregation over all members of a society. This allows engineers and ecologists to hold their particular normative views, but it does not allow their WTPs to be given more weight.

Democratic Theory

A fundamentally different approach is taken by normative democratic theory in political philosophy. The relevant normative index in democratic theory is the vote, not WTP. The policy that is normatively compelling is the one that garners the biggest majority in a democratic process. The more idealistic version (associated with populism in the United States) assumes that voters are sufficiently informed to have reasoned assessments of policies. Hence, the best outcomes are attained through direct democracy whereby voters make policy decisions. The initiative and referendum, the Athenian republic, and the New England town meeting are examples of institutions seeking to bring to practice the idealistic version of democratic theory.

The more practical version, underpinning the institutions of representative democracy and a professional, scientific bureaucracy (e.g. the French écoles and the Japanese "samurai bureaucrat" from the University of Tokyo), imagines democracy as a system in which voters not only are unable to make informed policy judgements, but also know that they lack such capabilities, and so delegate policy decisions to elected officials and civil servants. The latter, in turn, implement what is best, but are led to do so by seeking ratification of their values by the electorate. As a positive theory, normative democratic theory turns on political competition among candidates and the ability of voters to ascertain the character of these candidates, for otherwise majority support could not be normatively compelling. Thus, normative democratic theory has an important commonality with normative economic theory in that the former assumes something like a perfectly competitive political system. Just as consumers in a competitive market deal with many buyers and so are not faced with an absence of effective choice, so voters have sufficiently many political choices to permit voting to carry definitive information about the general character of public policy. And just as consumers in perfect competition know as much about a product as its producer, and so cannot be victimized by fraud or negligence, so do voters know enough about candidates so that they can ascertain which one will pursue the best policy.

Democratic theory has little to say about why regulatory reform emerged, other than that somehow it was all for the best. The sources of the values that led to reform are not the focus of analysis. Instead, policy-making officials did what they did because, retrospectively, voters would approve in subsequent elections. The difficulty with democratic theory, then, is that it tells us very little about where to look for a source of

change in political support for regulatory institutions. But democratic theory can be merged with welfare economics (or any other theory of the source of values) to be more predictive, for it then can tell us why votes might change. Specifically, if the marriage is between democratic theory and welfare economics, votes change because a majority of voters has a lower WTP for the status quo of regulatory policy. Voter WTPs can change for the same basic reasons as before: the facts have changed (e.g. the natural monopoly has disappeared or pollution has gotten worse), the source of values has changed (due to ideological shifts), or scientific knowledge has improved to give new insights about cause-effect relations (e.g., pollution is more harmful than we thought it was, or economic regulation reduces prices by less than we thought it did).

The normative theories based on aggregating numerical scores (or other continuous concepts of better and worse) do have a distinctly different set of predictions and normative conclusions than democratic theory has. The former theories ascribe change to aggregate values, whereas the latter ascribe change to majority votes. Only if citizens are perfect altruists (citizens regard every other person as equal in importance to themselves) do the two theories produce the same conclusions, and in this case all votes are unanimous after people realize which policy produces the greatest overall net benefits. As normative theory, then, democratic theory implies that a change is better only if it is evaluated as preferred by a majority, whereas welfare economics concludes that a policy is better if the gains to the winners numerically exceed the costs to the losers, even if there are many more losers (with tiny individual losses) than winners (with very large individual gains). As a positive theory, democratic theory argues that most people must have benefitted, whereas welfare economics reaches no such conclusion. Likewise, democratic theory does not necessarily conclude that the net benefits were positive for the entire society, just that they were positive for most.

Relevance of Normative Theories

The shortcomings of normative theories as positive predictors of change are apparent, and will not be discussed at length. Suffice to say that welfare economics as a positive theory of government is easily rejected for both pre-reform and post-reform regulatory policy. Regulation was not efficient before reform, and was not efficient afterwards. Likewise, the test of majority benefit is failed, for many policy reforms leave undone that which would have been economically beneficial. The remaining lectures will explore this point in more detail for telecommunications, airline and environmental regulation. For now, this conclusion will simply be asserted, with reference to these lectures and to the large mountain of confirming research (summarized in a forthcoming paper by Paul Joskow and myself).

The important point to note here is that, as of the beginning of the regulatory reform period in 1970, normative theories constituted pretty much the complete arsenal of analytical tools available to the scholar who sought to try to understand policy change. Indeed, the wide disparity between the teachings of normative economic theory and the reality of regulatory policy played a significant role in motivating new positive theories of policy that were developed in parallel with the regulatory reform era. Nonetheless, normative economics and domestic politics certainly played a significant role in the process. A significant puzzle is how and why they did.

Positive Political Economics and Policy Change

The term positive political economics refers to the application of microeconomic (or rational actor) theory to the study of politics. The question it seeks to answer is what policies will emerge from a political system in which people behave in a manner that is analogous to their behaviour in economic settings (production and transaction).

The simplest imaginable microeconomic political theory is the analogue to perfect competition in economics, in which numerous buyers and sellers meet to make transactions, but in this case to buy and sell policies. One can imagine that each actor has a WTP for various policies, and that citizens with a positive WTP for an action will simply buy votes from people with negative WTPs until either the former have acquired an electoral majority or the latter remain in the majority after the WTPs of the former have been exhausted. On each dimension of policy, outcomes are determined in the same way as in the market in that the winning policy is the one that maximizes net WTP. Indeed, if society applies a unanimity rule for voting (rather than majority rule), bidding for policies guarantees "Pareto improving" policy change – that is, all policy changes will leave everyone at least as well off as before the change.

For many obvious reasons, the "perfectly competitive republic" does not and can not exist. To begin, transacting votes is illegal, although not necessarily unknown, in all democratic societies. But this problem is less of an obstacle than it may at first appear to be, for citizens can always make "logrolls" whereby proponents of different policies agree to support the policies each advocates. Logrolling converts the policies themselves (e.g. the government budget) into the means for making transactions. All that is ruled out is direct cash payments for votes.

Assuming voting transactions are possible, the process of making all the necessary bargains is obviously impractical. Not only would citizens spend most of their waking hours striking bargains for policy, they would face massive collective action and coordination problems. Public policy changes almost always have collective effects, simultaneously affecting a large number of citizens. The groups of proponents and opponents, be-

fore bargaining could begin, would need to know the WTPs of all their members in order to begin negotiations to determine whether the positive or negative WTPs carried the day. Then, they would have to decide how to share the burden and benefit of the financial transfer from the winners to the losers.

Interest-Group Theory

The collective action problem in the hypothetical perfectly competitive republic has given rise to the first important contribution of positive political economic theory: the principle of organization of interests for effective political action. Because effective political organization is time-consuming and costly, a group must expect that the effect of organizing will be to produce policy benefits that exceed its organization costs. In order to mobilize for political action, a group faces two kinds of costs: a cost per member associated with forming the group and making a decision, and a fixed cost for gathering the information that the group needs to make an informed decision. The first can be written as f(N) and the second as F, and the total costs of an organization can be written as:

$$C = Nf(N) + F.$$

Generally, the first type of cost tends to be higher as a group becomes larger: as more and more people are added, it costs more per person to negotiate a common policy and to recruit an additional member. Mathematically, this implies that $f'(N)>0$.

To find joining a political organization worthwhile, a person must expect more costs than benefits. If B is the benefit of joining, which is derived from the policy change the group will produce, on average a group member will join only if:

$$B > f(N) + F/N.$$

From this simple expression, several important conclusions can be derived.

First, a citizen must pass a minimum threshold of caring about a policy (that is, B must be sufficiently large) or the citizen will not perceive sufficient motivation to participate in political action. Second, the difference between B and organization costs per person constitutes an adjusted WTP for political action. That is, after organizing, a citizen can contribute time, money and other resources to political action of the following amount:

$$WTP = B - (f(N) + F/N).$$

Thus, citizens who care more about a policy have more resources available to affect it, while citizens having less than a minimum interest will not be organized at all.

Third, if F is small compared to f(N) – that is, it costs more to organize and to decide what to do than it does to gather the necessary information for a decision – small groups are advantaged over large ones. The reason is that, because f(N) is increasing in N, so too, is C/N increasing in N whenever F is small. Put another way, suppose that the aggregated economic interest of two groups is the same. Then the smaller group has, first, a larger B per person, and second, a lower cost per person. Hence, the smaller group is more likely to be organized, and if both are organized, the smaller group will have a larger net WTP for effective political participation. Consequently, in large, representative democracies, small groups with a high per capita stake in a policy should have greater influence over that policy than large groups with a small per capita stake, even if the total stake of the larger group is greater.

Applied to regulatory policy, the organization of interest theory predicts dominance by people who have the most per person to gain or lose from regulation. For the most part, this group consists of participants in regulated industries. A worker, manager, or major stockholder in a regulated company depends on regulated activities for a substantial fraction (if not all) of income. Customers rarely spend very much of their income on any particular good or service, and generally industries have far more customers than employees and stockholders. Hence, the supply side is advantaged relative to the demand side in a regulated market.

In some cases, the greater political weight of supply interests can be partly offset. For example, a dispersed, atomistically competitive industry may sell to a very concentrated one comprised of large firms. One would not expect such a market to be regulated, but if it were regulated, the advantage would go to the buyers. In the United States, an example of this form of regulation was the control of natural gas prices at the wellhead. Natural gas is produced by a very large number of wells that are owned by a very large number of companies and even individuals, but it is sold primarily to a few pipelines and large utilities. The latter succeeded in obtaining economic regulation of natural gas in the 1950s, with the result being prices that were below market-clearing levels. This policy enabled gas utilities to set a lower price, resulting in higher profits and more effective competition against alternative sources of energy. Eventually, however, keeping prices suppressed by regulation reduced the incentive to drill wells, and a natural gas shortage ensued. Twenty-five years later, natural gas was deregulated to solve the shortage problem.

A more common circumstance is for some demand-side interests to have larger stakes in regulation than other consumers, leading to a situation in which some buyers are effectively organized and others are not. The predicted result is that an organized group would be charged lower

prices and receive better service than an unorganized group. Two examples illustrate the point.

In transportation, cities with transport terminals (rail stations, airports) are already organized, but travellers generally are not. A single city has no particular interest in participating in setting detailed prices as long as the price proceeding affects all equally; however, if a transportation carrier seeks to cancel service to a community, the city will care deeply. In U.S. airline, rail and truck regulation before 1980, transport companies had to receive approval from regulators to abandon a terminal. Even if the service was woefully uneconomic, the regulators would respond to the protestations of local governments and refuse to grant permission. Instead, they would seek to raise prices to all customers to offset the losses from the uneconomic service.

In telecommunications, large bulk users of services can find telecommunications to be a substantial cost. In the U.S., although by-passing the telecommunications system for any purpose was generally banned, in 1959 bulk users succeeded in obtaining permission to build private communications systems. Soon thereafter, the monopoly telecommunications provider obtained permission to offer special low prices to bulk users to keep them on the network. In the 1960s, the focus was on "private lines" – dedicated long distance links between cities, usually connecting the multiple facilities of a large corporation. Later, the telephone companies developed Centrex, a service whereby the multiple telephones in an office use a central office switch of the telephone company (rather than the customer's own switch) to make calls within the office from one extension to another. Centrex lines were priced at about one-third of regular lines, a far greater discount than the cost difference. The telecommunications examples illustrate how large users obtained a better deal through regulation than smaller users, as predicted by the interest-group organization model.

In addition to demand-side groups, suppliers of the regulated entity also may be effectively organized to influence regulatory policy. Regulated industries often buy inputs from other industries that sell all or most of their production to regulated firms. These industries may be even more concentrated, and have even greater stakes in regulatory policy, than the regulated firms. For example, the manufacturers of aircraft, railroad cars, telecommunications switches, and electric generation equipment are all in very concentrated industries that deal only with regulated companies. Interest-group theory predicts that these companies will succeed in obtaining the official blessing of regulation to become special favorites of regulated firms. Not surprisingly, regulated companies in infrastructural industries often deal exclusively with national champion suppliers of major inputs, no matter how high the price and low the quality of the champion's product. More subtly, these manufacturers also benefit if the structure of regulation leads to excess capacity, such as by imposing excessive reliability requirements or by providing service to areas

where it is not used. In the U.S. excess capacity was a widespread phenomenon among regulated industries prior to reform.

The other important supplier interest that is likely to be effectively organized is union labor. A labor union, because it has already become effectively organized for another purpose, can easily extend the domain of its activities to include political activity. High on the list of priorities of a union in a regulated industry is the nature of regulatory policy. Like equipment suppliers, unions prefer excess capacity, plus above-market wages. Regulation gives unions an advantage, because firms cannot enter regulated markets without the approval of the regulators. Hence high union wages will not attract lower-wage firms (employing another union or non-union labor) if regulators cooperate. Thus, interest-group theory predicts that employees of regulated firms will receive higher wages than employees performing the same tasks in other companies, and that regulators will assist unions by denying entry to lower-wage firms. These predictions, too, are consistent with U.S. experience in the era of regulatory reform, for in all cases deregulation has led to both entry by non-union firms and reductions in real wages for employees of the formerly regulated firms.

The interest-group theory of regulatory policy is broadly consistent with many of the facts regarding regulation, but it is incomplete in several important ways. First, by itself it does not predict that regulated industries will be particularly inefficient. All organized interests share the objective of having the regulated firm extract as much as it can from unorganized groups (as long as the amount extracted is not sufficient to motivate them to become organized). Thus, the theory predicts that regulation will be used to redistribute income to organized groups, but not that regulated firms will have inefficient price structures or production. One would expect, for example, extensive use of elaborate pricing systems that contribute to efficiency while extracting substantial excess revenues from customers (peak-load pricing, two-part tariffs, etc.). One would also expect efficiency in the design and operation of the system to minimize real resource costs (but not, of course, the income of people in the regulated firm or in dedicated supply industries). In practice, an important fact of deregulated industries is they were substantially reorganized to improve their efficiency. Clear examples are the development of the "hub-spoke" route structure in airlines after deregulation, and the massive restructuring of AT&T's manufacturing activities after divestiture.

More significantly, interest-group theory does a poor job of predicting both economic deregulation and protectionist regulatory expansion. In most cases, deregulation has caused substantial financial losses in deregulated industries. Because the participants in the industry did seem to be the principal beneficiaries of regulation, the decision to deregulate can hardly be said to have been excessively influenced by them. Likewise, manufacturing industries and the energy sector have paid the cost of envi-

ronmental regulation, even though they opposed the stringency of the policy forcefully. Subsequently, established firms did manage to obtain a system of regulation that favored them in comparison with newly established production facilities, but in most cases these benefits have been small compared to the costs of complying with more stringent standards. At best, interest-group theory explains a bias in the regulatory process, but neither its initiation nor its demise.

The most important lesson from interest-group theory is to identify a potentially serious cost of regulation – that the process is excessively sensitive to organized economic interests. This bias may have been a purpose of a regulatory policy, and its elimination an objective of deregulation. But it is also plausible that interest-group bias is sometimes accurately perceived by political actors as a necessary evil of regulatory policies that are enacted for other reasons, or that the protectionist feature is designed into the institution to form a coalition of support between the regulated industry and others seeking regulation for an entirely different set of reasons. To illustrate, suppose that neither environmentalists nor labor unions could command a majority in the U.S. congress in the 1970s, and that unions wanted "plant closing" protection in a declining manufacturing sector while environmentalists wanted reduced emissions. The Democratic Party, including both groups within its coalition, could achieve both objectives somewhat imperfectly and incompletely by using emissions control regulation as a means to slow the restructuring of the economy. Only this half loaf for each could command a majority, and so, arguably, it was adopted.

The hypothesis of a labor-environment coalition is, of course, an ad hoc explanation. No reasons are apparent for this particular coalition to form, rather than others that also might form a majority to combine environmental regulation with some other policy objective. Interest groups need to command a majority in Congress to achieve legislative objectives, so the question remains how these majorities are formed, and why some groups are included but others are not. Labor unions surely were not part of the coalition adopting economic deregulation. Why did they lose in one field but win in another during the same period? And, if the labor/environmental coalition explains the regulatory policy of the 1970-1990 period, why, simultaneously, did the same coalition fail to stop free trade, and especially the Canada-U.S. Free Trade Agreement and, in 1991, the "fast track" process for extending the free trade zone to Mexico? Quite obviously, the coalition account leaves much to be explained. Nonetheless, this conjectural explanation illustrates how the coalitional version of interest-group theory can provide a richer model of the origins of policy change.

Pathologies of Majority Rule

If, as is usually the case, no single interest (or point of view) commands a majority concerning regulatory policy, an extremely important question is how winning policy coalitions form to initiate and maintain policies. The transactional arguments based on the principle of willingness-to-pay assume that the winning policy is the one that maximizes net WTP. The key assumption of this theory is that the transactional approach to WTP maximization has a unique, stable outcome – an equilibrium policy that will not change unless the underlying preferences (WTPs) of citizens change.

Unfortunately, the assumption that an equilibrium exists in such a decision-making process is untenable. Kenneth Arrow's Nobel Prize in economics was awarded in part for his path-breaking treatise, *Social Choice and Individual Values*. In this work, Arrow proved the "Impossibility Theorem" – that unless citizens have the same preferences, no social decision process can simultaneously exhibit a few simple (and desirable) properties, including economic efficiency, a unique, stable equilibrium, and the absence of a dictator. Thus, majority rule voting, in particular, is unstable – indeed, chaotic, in that a society can move from literally any technically feasible policy to another by a sequence of majority-rule votes. Applied to coalition formation, the implications of this work are that many coalitions can emerge and that if one does emerge, it will be unstable. Hence, policy change not only cannot be predicted from knowledge of the strength of interest groups, but it is random.

The chaos of social choice theory is obviously incorrect as a prediction about public policy, for no society – not even nations with unstable, perpetually changing governments – displays rapid, random policy shifts. But stability in real-world policy does not disprove the theory. Chaos theory is based on the assumption that all decisions are made by majority rule in a very simple voting process in which anyone can propose any measure at any time. In practice, all real-world voting systems are far more complicated than this. Indeed, legislatures typically have very complex systems of rules governing who can speak, when measures can be introduced and amended, and how the agenda of a session will unfold. Social choice theory provides an explanation for why the simple concept of voting requires so complex an institutional structure. In particular, institutional complexity can make policy more stable – institutions are, among other things, a means for overcoming the chaos of unstructured collective decision making.

A simple example drawn from the procedures of the U.S. congress (and most other legislatures) illustrates the point. Suppose a legislative body is considering a bill (B). Before debate starts, the legislature knows that one of its members will attempt to change the bill by proposing an amendment (A). Indeed, in the congress, members must inform the Rules Committee of their intention to amend a pending bill before debate be-

gins. Assume that S refers to the status quo, prior to any action being taken by the legislature. Thus, the legislature is choosing among B, A, and S. Finally, assume that the legislature divides into three groups of equal size. One group prefers the amended bill (A), but would rather have the original bill (B) than keep the status quo (S). The second group prefers B to S, but prefers S to A. The last group wants to retain S, but prefers A to B. These preferences can be represented as follows, where ">" is used to indicate preference:

I: A>B>S
II: B>S>A
III: S>A>B.

Notice that in a simple majority vote, A defeats B by a vote of two to one; B defeats S by a similar vote; and S defeats A by the same margin; therefore, these alternatives illustrate the famous Condorcet paradox of intransitivity in majority-rule voting among individuals having transitive preferences. In a legislature without any rules other than majority dominance, policy would be chaotic, because each of these policies could be defeated by one of the others. Alternative B would be adopted by defeating the status quo; then A would be introduced and would defeat B; and then S would be reintroduced, defeating A.

In nearly all legislatures, the chaotic circumstance described above could not occur because of the agenda rule of the decision-making body. Specifically, in any legislative debate, the last vote taken is almost always a contest between the bill as amended versus the status quo. That is, rather than having a vote after each measure is proposed, an original proposal is first "perfected" by amendment during debate, and then the ultimate decision is whether the perfected (amended) proposal is regarded as superior to the status quo by a majority. Moreover, once defeated, a measure cannot be reintroduced during the same session of the legislature. In the preceding example, these rules produce the following agenda:

Step 1: Vote between B and A
Step 2: Vote between S and the winner of B against A.

In this sequence of votes, A defeats B by a vote of two to one, but S defeats A (the amended bill) as well, so the status quo is preserved.

A second stability-enhancing feature of almost all legislatures is the committee structure. For example, in the U.S. congress (both House and Senate) and the European Parliament, all bills are referred to a committee. Except under unusual circumstances, the entire chamber of the legislature will not consider a measure until the committee issues its "report" – that is, reports a measure to the entire chamber which has received majority support in the committee. Thus, the committee has a "gatekeeper" function in that only measures it prefers to the status quo are considered

by the entire body. In addition, in the U.S. congress committees are further advantaged in that the chair of the committee is normally accorded the right to respond first to any amendment, and immediately to propose a further amendment to any amendment. Finally, in the U.S., with two equal legislative bodies, the committee responsible for a measure usually is appointed to the "conference committee" of the two chambers, which resolves differences in the measures passed in each body. No measure is reported from the conference committee unless a majority from each chamber approves, giving the original committee further assurance that nothing will be enacted unless it commands a majority on the committee.

Status Quo Bias

The effect of committee structure is to bias policy in favor of the status quo. A change in policy must command not only a majority in the entire body, but a majority in committee as well. If the committee is not essentially a mirror of the entire body, the requirement for satisfying two majorities will be a more difficult hurdle than satisfying only a majority of the entire chamber. Normal practice is to appoint committees that are not fully representative of the entire legislative body. Instead, parties normally appoint their representatives on committees, and they can be expected to do so with an eye towards the policy interests of their members. A legislator from a district that cares a great deal about one particular policy will normally find its representative appointed to the corresponding committee, for by doing so that member's party is assisting its own prospects for holding the seat in the next election. As a result, a committee tends to be populated by representatives whose constituents are unusually interested in the committee's policy jurisdiction: e.g., farm constituencies tend to be over-represented on the agriculture committee. The effect is to give representatives whose constituents care most about a policy an institutionalized veto over policy changes.

The implications of these observations for regulatory policy are as follows. First, like all policies, regulatory policy ought to be difficult to change through writing new legislation. Moreover, if regulatory agencies manage to change policy without violating their old statutory empowerment from the legislature, passing new legislation to re-establish the old status quo will also be difficult, if not impossible. Second, if regulatory policy was created in part to provide benefits to the regulated industry, the committee structure of the legislature can be expected to have institutionalized industry's interest in preserving the policy. If the preferences of other interests change so that regulation can no longer command a majority, it may still persist because the legislative committee exercises its gatekeeping veto to prevent reform from being considered by the entire body. Third, if the entire body cares enough about the preceding state of affairs to change its own rules, reform can be facilitated by removing a

committee's gatekeeping function (such as by transferring jurisdiction to another committee).

The preceding argument can easily be illustrated by a simple diagram. Suppose policy is the choice of a position on a line (such as, for example, the allowable emissions from an automobile), as in Figure 1.1. Let L represent the best policy choice from the perspective of the entire legislature, C be the best policy choice from the perspective of the committee, and A be the best policy choice from the agency's perspective. Finally, let S represent the status quo ante in which emissions are not controlled. (Technically, both L and C represent the position of the median voter in each group.) Assume that each political actor prefers a policy as near as possible to the best possible policy.

Figure 1.1: Policy Relationships between Agencies and the Legislature

The relative positions of the actors reflect the view that committee members tend to be advocates of the policy in their jurisdiction, and that agencies, because they do not personally face conflicts between their policies and other policies, tend to be even more favorably disposed to their policy than the committee. In the situation depicted in Figure 1.1, the committee would propose C to the legislature. Depending on the rules of the legislature, the bill that was enacted would lie somewhere between C and L. For example, the legislature might amend C by substituting L. Alternatively, the committee might be able to force a simple yes-no vote on C without amendment. Because C is preferred to S by the floor (C is closer to L), C would then pass. In the former case, if the agency in fact implements A, the committee will not act to "correct" the agency's policy, because the committee prefers A to L. In the latter case, because the legislature enacts the committee's most preferred policy, the committee will try to correct an agency that implements A. Thus, by giving committees more power, the legislature in this case actually will obtain a policy outcome nearer to its preferred one! This case is not general, however, for it depends on the particular assumptions about the positions of S, L, C and A.

Finally, suppose that the position of the legislature's optimum changes. If the actual policy outcome is not the old L, but something nearer C, the shift in L will have no effect. If L moves closer to S (the legislature wants to relax the standards), the committee will simply not report a new bill. If L shifts towards C, the committee will report a new bill only if the new position of L is nearer to C than to the current policy. If the legislature gives committees considerable power so that originally a position near C

is enacted, no change in L is likely to result in new legislation unless, of course, the legislature decides to reorganize itself in order to disenfranchise the committee.

The significance of this example is that it illustrates an important characteristic of the structure of decision making. Creating powerful, specialized committees and implementing agencies does help to stabilize policy, but it does so by making policy less responsive to the democratic process. The task of legislatures is to create a structure that finds a balance between the chaotic implications of responsiveness and the undemocratic implications of structure-induced stability.

Incomplete Information

Legislative committees, as well as implementing agencies, cope with another serious problem of policy making: the informational requirements of devising a good policy. In the regulatory field, reasonably effective policy requires detailed knowledge about the technology and economics of regulated companies. Part of the job of a legislative committee and an implementing bureaucracy is to gather the information necessary for rational public policy. But the need for information raises a more profound problem for collective decision-making processes. The theoretical arguments developed thusfar have assumed that people who make policy decisions are relatively well informed. Obviously, if agencies and committees have the task of becoming informed as part of the process of making decisions, it necessarily must be the case that neither voters nor the remaining members of the legislature are sufficiently well informed to make these policy choices on their own. Indeed, the problem is even worse than this. In majority-rule voting within large bodies, a single vote is unlikely to be decisive. Consequently, a single voter has very little incentive to put forth time, effort and money to become informed before casting a vote. Mathematically, if P is the probability that a single vote will be decisive, B is the benefit of an informed vote, and C is the cost of becoming informed, then a voter will bear the costs of information only if the benefits exceed the costs:

$$P*B>C.$$

If the number of people casting votes is more than a few, P is a very tiny number. Even for the U.S. Senate, which contains 100 members, P is very small – far less than .01 – whereas for the entire electorate in even a small nation, a single voter's P is essentially zero. This means that voters have essentially no incentive to be informed about policy, or even about the policies advocated by their representative, and that likewise representatives have little incentive to be informed, in part because their constituents will not be likely to know whether they cast informed votes. The result is a phenomenon called "rational ignorance:" because votes have

little power, rational behaviour by a voter is to make little or no investment in information about how to vote.

Committees overcome the rational ignorance problem because they have sufficiently few members that each vote is likely to be important a reasonably high fraction of the time. Thus, members of committees have an incentive to collect information about the policies in their jurisdiction. But the superior information possessed by committee members (and by agencies that implement policies) creates the problem of "hidden action" arising from "hidden information". That is, agencies and committees can withhold information from others, and to some extent take policy actions that are not observed by others, to pursue their own objectives – objectives that may not be consistent with the preferences of either other members of the legislature or their own constituents.

One consequence of rational ignorance is that after legislation is passed, the policy that is implemented can depart from the expectations and intentions of the voters and legislators who supported it without their even knowing the change has taken place. And, if committees are unrepresentative in that they are composed of members whose constituents have the greatest personal stake in a policy, the direction of bias will be toward the interests of these preference outliers.

The most important implication of the preceding argument is that information imperfections bias policy in favor of organized interests, for the latter have organized in part to collect the information necessary to advance their preferred policies The structure of a legislature gives committees the advantage of superior information. Most likely, the committee will offset to some extent the informational advantages of organized interests because unorganized constituents will carry some weight in the committee's own policy preferences. But the committee will still use its informational advantage over voters and the rest of the legislature to produce a policy that is closer to its own preferences than would arise if everyone were perfectly informed.

Administrative Procedures and Political Control

The primary weapon available to legislatures in controlling hidden information and hidden action is to specify the informational requirements of a policy decision by an implementing agency. The means of specifying informational requirements is administrative law: the formal statement of the powers of an agency and the procedures it must use to implement a policy.

The use of administrative law to control policy is made most apparent by a specific example: the regulation of potentially hazardous chemicals. In most advanced societies, one category of chemicals – pharmaceuticals – is regulated especially stringently. In the United States, the statute guiding the Food and Drug Administration requires that a company seeking to market a new drug must prove that the drug is both safe and effective ac-

cording to rigorous scientific testing standards. Consequently, obtaining a license for a new drug is expensive and time consuming, so that relatively few new drug discoveries proceed through the testing and licensing process. By contrast, hazardous chemicals that are used in industrial processes are regulated by very different procedures. In the United States, the regulator – the Environmental Protection Agency – bears the burden of proving that a chemical is unsafe (compared to the alternatives), and must do so within a few months after a company reports its intention to market it. The budget of the EPA is only sufficient for it to examine closely a handful of chemicals per year. The rest proceed to market without serious regulatory scrutiny. Obviously, the latter process gives a strong informational advantage to inventors of new chemicals, whereas the former not only eliminates the informational advantage, but forces inventors to gather more information about a new drug than they would otherwise find useful for making a decision about whether to market it. This advantages the owner of an old drug that might be replaced by a new one. It extends the time that it takes to enter the market successfully, and before entry occurs provides information to the owner of the old drug about the exact nature of the competitive threat.

Whether strict or lax rules for introducing dangerous chemicals are more desirable is not the point of the preceding example. Instead, it illustrates how procedures can allocate advantages among parties affected by a regulatory decision. In the United States, the legislation establishing a regulatory policy normally contains detailed instructions to the agency about how it must go about its business: what data it must collect, and what standards and burdens of proof it must apply. In addition, agency procedures are also governed by the Administrative Procedures Act, which specifies that agencies must provide a rational explanation for their actions, based on the evidence submitted to them by interested parties and their own staffs. Whether agencies have followed these procedures is then subject to judicial review. If a party affected by an agency does not receive a favorable decision, it can appeal the decision to the courts, claiming that the agency did not follow its legislative instructions with respect to policy objectives and procedures.

The legislature can use procedural requirements to favor an interest that is not informationally advantaged. For example, an environmental or consumer protection organization may be knowledgable about the effects of a hazard on its members, but may not be knowledgable about the costs and performance of technical solutions to the problem of reducing exposure to the hazard. If the act regulating the hazardous activity places the burden of proof on environmentalists to show that a control is worth its cost, the process will favor the producer of the hazard, who will possess "hidden information" that it can use strategically to bias the policy outcome. If, however, the burden of proof is on the producer to show that a control is unreasonable, the environmentalists are advantaged, because the

producer will have to reveal more of its private information to influence the outcome.

One potential explanation for regulatory policy change lies in the role of procedures in shaping outcomes. If new private information is discovered, it will be fed into the regulatory process only if it is beneficial to its discoverer. New public information will be fed into the process only if it adds strength to the position of one of the parties represented in the process. For example, the "ideas" account of regulatory policy change can be given more concrete explanatory power when interpreted through the biases inherent in administrative decision processes. New findings by economists about the effects of regulation can influence policy only if two conditions hold. First, the economic effects of regulation must be part of the information on which decisions must be based according to the statutes delegating authority to an agency. Second, some parties to an agency decision regard the new information as useful to them in advocating their preferred policy change. These parties can be either organized interests that participate in agency deliberations, or an internal staff that is charged with generating information that organized interests would not necessarily produce on their own. If these conditions are satisfied, regulatory policy can change without any change in the preferences of the relevant political actors in the legislature or the interests that are active participants in the political and administrative processes associated with regulatory policy.

Political Entrepreneurs

Thusfar, the analysis of regulatory policy change has focused almost entirely on the influence of organized interests. The danger in such a focus is that it is easy to interpret interest-group theory as encompassing all of political activity, rather than as simply a source of bias in outcomes. Rational choice theory predicts that organized interests will have greater weight than unorganized interests, but it does not predict that all sources of political preferences outside of political organizations will be accorded zero weight. Because organizations overcome rational ignorance and coordination problems for their members, they will cause voters to give relatively more weight to aspects of their political preferences that are represented by politically relevant organizations. But this does not imply that other concerns do not influence behaviour. Thus, if voters' preferences about a policy shift, and if to some extent decisions about voting and other forms of political participation are based on this policy, elected representatives are likely to shift their preferences in the same direction. But policy will not necessarily shift as well, due to the status quo bias in political institutions. Only if the status quo bias is overcome will any shift in desired policy affect actual policy outcomes, whether the source of the shift is organized or disorganized interests.

The role of the political entrepreneur is to orchestrate the effective representation of hitherto less powerful interests and to find a way to achieve policy change on their behalf. The theory of political organization does not closely examine the origins of an interest group. Essentially, people of like mind find each other and become organized, perhaps facilitated by the presence of other organizations that were created for another purpose, such as a company, a union, a church, or an outdoor recreation club. The members then share the cost of the organization, and benefit from the change in policy that it brings about. But this mechanism is not the only way to mobilize citizens of like interest. The other is for a political official to undertake the effort and cost of mobilization: gathering data about a policy, interpreting it for an otherwise unorganized group, communicating the message effectively, and then benefitting from individual decisions by group members to support the politician. Thus, one aspect of political entrepreneurship is to recognize an unsatisfied, unmobilized political preference, and to find a way to induce people who hold that preference to take political action on the basis of it. Of course, the latter will not occur unless the as-yet unfocused preference is potentially strong enough to push aside some people's other motives for political action.

Once the new group is mobilized, the political entrepreneur must solve the institutional problem of obtaining a change in policy that is favorable to this constituency. Applying chaos theory, the political entrepreneur holds one advantage – there must be a way to reach any policy change from the status quo. The obstacle is the status quo bias in the political system. Somehow the political entrepreneur must overcome this bias.

The easy way to upset the status quo is to make the issue so important to other politicians that they are forced to go along. Often chief executives – a president or a prime minister – obtain their objectives this way by winning an overwhelming election on the basis of some new issues. Even if these politicians do not command a solid majority in the legislature, some of their opponents go along with their proposals out of fear of electoral reprisals if they do not. Thus, regulatory policy might change because a political leader simply forces it from a strong base of popular support.

The more difficult way to upset the status quo is to find a loophole in the procedures of the political system that creates the status quo bias. In this circumstance, the entrepreneur does not have sufficient political power to command a legislative majority by simply demanding it. Instead, the entrepreneur must find a way within the existing institutional structure to take advantage of the fundamental instability of the status quo. Two examples illustrate the point. First, a president, with superior knowledge about the policy proclivities of nominees for a regulatory agency, may succeed in appointing officials who will shift policy in a way that the oversight committee cannot correct, essentially taking advantage of an arrangement of preferences as depicted in Figure 1.1 that is favor-

NO

1) So, President's WTP > committee's WTP ?! Why ?
2) A president is usually not a leader of an interest-group.
IF he is, he will probably not continue to be after his nomination.
IF he continues, the link President-i-group will be known ⇒ no 'superior knowl'

cf. Reagan & Bork

able to successful "hidden action" by the agency. Second, a legislator may find a way to "poach" on the jurisdiction of another committee, by-passing the latter's gatekeeping power, or may succeed in obtaining some new appointments to a committee with hidden preferences that are inconsistent with the committee's prior policies.

Political entrepreneurship interacts with interest-group analysis in still another interesting way. In the preceding account, the political entrepreneur creates and maintains the interest group by being the instrument for mobilizing it. Alternatively, a change in policy can cause the heretofore unorganized interest to become organized. If a policy change substantially increases the per capita stakes of the winners, they may then have sufficient incentive to mobilize themselves – to pay their own organization and political participation costs – to defend the new status quo. This ex post organization occurs when the identity of the actual winners is sufficiently uncertain before the policy change to prevent ex ante mobilization in favor of reform. For example, the people who will hold stock and be employed in firms that will profit from reform may face a small ex ante probability of being the beneficiaries of change because a large number of firms, stockholders and employees may prove to be the eventual winners. After the change this uncertainty is resolved, and the benefit of winning becomes concentrated in the ex post winners rather than diffused over a larger number of ex ante potential winners. The ex post resolution of the uncertainty can therefore cause the beneficiaries to cross over the threshold for becoming politically organized, where expected benefits from organized participation exceed the expected costs.

A political entrepreneur has an obvious incentive to favor policy changes that create their own support constituency ex post. If a policy change creates an organized support constituency, the political entrepreneur is in a position to capture support – votes, volunteers, contributions – from the new group, whereas if no ex post organization emerges, the political entrepreneur must bear the costs of mobilizing the continued support of the beneficiaries.

A source of policy change that is very similar to a political entrepreneur can be a judicial official who, in reviewing an agency decision, simply alters policy in a way that other political actors cannot effectively counteract. Referring back to Figure 1.1, if the legislature enacts policy L, but the court instructs the agency to implement another policy nearer C, both the agency and the oversight committee will gladly comply, and the legislative body as a whole will never get the opportunity to re-establish L by passing a law that reverses the court's decision. Of course, this result is far more likely to arise if the court's policy change causes the mobilization of a group that receives benefits from the change. In this case, the underlying set of political forces acting on the policy preferences of legislators will shift in favor of the policy change, thereby reducing the chance that the court's actions can be reversed in the legislature.

In like fashion, a bureaucratic official can secure a permanent policy change that is inconsistent with the intent of a statute; however, to do so, the bureaucratic official must succeed by first, taking action without political overseers knowing the decision is about to be made (otherwise, they can intervene to stop the action before its support group mobilizes), and second, obtaining approval for the change during the process of judicial review. One property of U.S. administrative law is that it serves to protect against both eventualities by requiring public notice of intended policy changes (with a period for comment and reconsideration) and by giving courts detailed standards for reviewing agency decisions. Of course, courts can still go along with the non-complying agency.

In the European Economic Community, the European Parliament plays a role that is similar to the role of the courts in the U.S. in that Commission and Council decisions are reported to the Parliament for review and possible amendment or disapproval. The advantage of this system is that it does not rely on a third party (the judiciary) to protect against policy decisions that are inconsistent with the policy preferences of the legislature. The disadvantage is that as government becomes more complicated, parliamentary review of executive decisions and proposals can become extremely time-consuming, causing some policy decisions to be reviewed only in a cursory way, or to be reviewed only by a committee of the legislature. Because committees can be expected generally to be unrepresentative of the entire chamber, the latter procedure enables committees and bureaucracies to work together to produce policies that, with full knowledge, the legislature would block.

The entering wedges for political entrepreneurship are informational imperfections in the political system and the fundamental instability of majority rule decision making. A public official can make use of superior information about some aspect of policy or about the decision-making process to cause policy to change. For the policy to stick, it must be within certain boundaries determined by the preferences of citizens, legislators, and other government officials. Within limits, policies can be changed without the possibility for effective response because of the status quo bias. These limits, in turn, can be altered by political entrepreneurship directed at voters, whereby the entrepreneur bears the costs of giving effective political voice to a previously unorganized group.

Conclusions

Rational actor theory supports a very complex view of policy change, admitting a variety of avenues for reform. Regulatory policy is almost always a complex enterprise, requiring the implicit consent of numerous government officials and a substantial effort in collecting and evaluating arcane information. Thus, the complexity of regulatory policy adds to the complexity of potential political explanations for policy change.

Obviously, regulatory reform could arise from a shift in preferences or a change in the information base of policy. Rational actor theory, by focusing on how these sources of change work their way through the political process, gives added concreteness to both ideological and informational explanations of change. In particular, these shifts need not cause policy to change if political institutions and preferences are organized in a way that protects the status quo against such shifts. By the same line of reasoning, neither changed preferences nor changed information is a necessary condition for durable policy change.

The somewhat unsatisfying conclusion is that the theoretical ideas available to us do not make strong, robust predictions about how, when and in what direction policy will change. Instead, theory serves more to structure the information about the sequence of events surrounding a policy change. Only detailed empirical investigation of a regulatory policy – its economic consequences and the institutional structure in which it was created and implemented – can ultimately shed much light on its evolution. The reforms of the 1970-90 era might have arisen for quite different reasons, connected only by similarities in how the status quo bias was overcome, not in any commonality of purpose or base of political support.

The approach to explaining policy change that has been outlined here has recently come to be called "The New Institutionalism". The term refers to a scholarly approach early in the 20th Century which viewed every facet of social behaviour as *sui generis*, depending on the detailed facts of the issue at hand. The search for a unifying theory was regarded as fruitless. In economics, institutionalism was associated with the view that every industry had to be studied separately, that differences in technology, resources and tastes made a useful general microeconomic theory of market processes infeasible. In political science, institutionalism took the view that every issue of public policy had its own special politics, and that cultural and historical differences made generalizations across nations in even a single policy area unfruitful.

The new institutionalism is obviously not compatible with the old, for it is based on a coherent model of political decision making that cuts across all policies and cultures: the rational pursuit of policy objectives by competing political actors with conflicting purposes. Its commonality with the old is that it does take institutions seriously. In the new institutionalism, institutions allocate decision power and specify decision rules, and so define the feasible range of policy change. Thus, to explain policy change one must mediate the general theory of preference-based choices through the special features of the institutions in which policy decisions are made.

In the regulatory policy domain, the new institutionalism provides a means for integrating numerous other theories of policy reform, some of which are not even based on rational choice theory. The most important contribution of the new institutionalism is that it focuses attention on the

choice of processes as a means of stabilizing policy and assuring that it will have a particular orientation – or bias. Presumably political actors know that institutions create decisional biases, and that they may face an impossible task in trying to re-establish their policy objectives if actual policy drifts from their intentions. Thus, decisions about policy structures and processes embody protections against policy drift, and contain information about the intent of those who changed a policy. To infer the reasons for a policy change, then, it is perhaps more important to examine the process by which the new policy is to be implemented than to read the rhetoric about the goals of new laws or new regulations.

Lecture #2 – Telecommunications Policy

Policy for traditional public utilities – natural gas, telecommunications, electricity, pipelines, water – is evolving rapidly, on somewhat divergent paths, throughout the world. Privatization and deregulation of at least some utility industries are being pursued in a surprisingly large number of countries outside of the OECD, including some very small nations (New Zealand) and some developing states (Mexico). A generation ago, most would have regarded such developments as unthinkable, for everywhere utilities were regarded as natural monopolies.

In the U.S., no utility industry has been deregulated, but liberalization has taken place in all except water systems, where the massive subsidization of water for agriculture has thusfar prevented a more economically rational system of water pricing and allocation. The most important reforms are in electricity and telecommunications regulation. In the former, although electric utilities still own most electric generation capacity, most new generation investment now is undertaken outside of the regulated utility structure, and brisk contract and spot markets have developed for the delivery of power to retail utilities and large industrial consumers. In telecommunications, basic access to the system and other "local" services (within a metropolitan area) have been separated from national services and manufacturing, and much of the industry has been deregulated.

This essay focuses on telecommunications, although the circumstances in this industry are in some ways similar to conditions in other utilities. The focus is on telecommunications because of its unusual rates of growth and technological progress, and because the telecommunications system is being used for an ever-widening array of services. Indeed, so many industries now either provide novel services or devices that use the network, or depend on it in their own business, that telecommunications is widely regarded as one industry that no nation can permit to be seriously inefficient. Consequently, it is somewhat surprising that so many nations have taken the risk of radically redesigning the structure of the industry and its policies.

Until the 1970s, telecommunications was widely thought to be most efficiently provided by a single, ubiquitous national carrier who owned or controlled all aspects of the network (including the equipment attached to it). Moreover, for national security reasons it was thought essential to have a single system under unified management. Many citizens and gov-

ernment officials still hold this view, making telecommunications liberalization highly controversial.

Nevertheless, in the post-War era, the commitment to a single national telecommunications system has weakened almost everywhere, and has all but disappeared in a few countries. However, among even large, economically advanced nations, substantial differences have emerged regarding the direction of reform. At one extreme, with the United States, the United Kingdom and to a lesser extent Japan as the principal examples, are nations that strive to come as close as possible to eliminating monopoly entirely, even with respect to basic access to the national network. In these countries, policy has permitted competitive entry into almost every nook and cranny of the industry, and price regulation is becoming more relaxed as greater reliance is placed on competition to control prices and profits. At the other extreme, represented by France, Germany and most other European nations, policy seeks to continue to protect the core network infrastructure as a monopoly. In these countries, reform consists of permitting competition at the periphery: customer equipment (other than the first telephone) and new information services, with the prices, profit and technical evolution of the network far more heavily controlled by government regulation or even nationalized operation.

These two approaches reflect two fundamentally different ways of thinking about the telecommunications industry, and other utilities. One paradigmatic view – the one that dominated policy until the recent era of reform – is the "humming system" view as exemplified by the engineering approach to network design. All network industries, including telecommunications, have inherent economies of scale. The source of these economies is in the flexibility in traffic flow that arises in any network. The greater is the interconnectedness of the system, the smaller is the network capacity that is required to handle any given pattern of use. Network economies are even more important when the demands placed on the system occur randomly, so that the system has to be "overdesigned" to cope with the unusual circumstance in which an especially heavy load is placed on it. Network interconnectedness minimizes the incremental cost of protecting against system failure when extreme demands arise.

The second conceptual model of networks emphasizes flexibility and diversity. For example, information services using other technologies (broadcasting, publications) exhibit extensive product differentiation, and substantial entry and exit by suppliers. In part, this diversity arises from the variety of different uses to which customers put information services. Moreover, these uses change radically as tastes and production technologies in information-using industries change. In addition, the underlying technology of telecommunications is itself rapidly evolving, gaining the capability to perform more and more functions. Meanwhile, the hardware costs of telecommunications are falling, and constitute a declining share of the industry's total costs. Hence, technological economies of scale are

becoming less important, whereas knowledge of customer requirements, effective marketing, and transfer of technical information among technologists, managers and customers are becoming more valuable. These latter attributes of an information-services provider require organizational flexibility and favor a variety of specialized organizations. Organizational diseconomies of scale in very large monopoly telecommunications carriers, therefore, may swamp the economies of scale in the network.

In the United States, the second vision of the industry has certainly been more influential since the late 1960s than before, or than it has been in other nations. Thus, it seems superficially plausible that telecommunications reform can be traced to the development of an ideological commitment to competitive, unregulated markets, or the "new idea" of the second view of the industry. Unfortunately, history belies this interpretation. The structure of the telecommunications industry has been controversial since its birth with the deployment of the first commercial telegraph systems early in the Nineteenth Century. Moreover, the events that caused reform in the U.S. to go substantially beyond reforms in Europe were something of a fluke. The U.S. did not adopt its current policy because a president and/or a congress formally decided to pursue it. Instead, a series of incremental reforms, some of which were unexpected and to some degree politically unsupported, were imposed on the system. Advocates of the status quo ante could not reverse these changes, although on several occasions they tried and came quite close.

Historical Roots

Since the industry began, numerous businesses in the U.S. have sought to participate in the telecommunications industry. Shortly after Marconi built the first commercial telegraph link in the United States, competition developed to connect the rest of the country to telegraph service. During more quiescent times, the competitors would merge or recognize geographic market segmentation, but in other periods they would compete fiercely.

Early Structural Controversies

In the mid-1870s, two different versions of a telephone were developed and patented by different companies, Bell Telephone and Western Union, the dominant telegraph firm. Rather than engage in protracted competition and patent infringement suits, the two companies reached a singularly one-sided accord. Bell would develop the telephone as a monopoly, and Western Union would retain a near monopoly in the telegraph, with neither encroaching on the business of the other. Twenty years later, this arrangement had plunged Western Union into economic decline, and it was

acquired by Bell, which renamed itself the American Telephone and Telegraph System.

Bell's monopoly in telephones remained only as long as its patents were in force. By the mid-1890s, other companies began to enter the local telephone business, competing directly with the local Bell System. Bell sought to thwart this entry by refusing to interconnect with its competitors – that is, a customer of a non-Bell company could neither call or be called by a Bell customer. But denial of interconnection did not work, and by the turn of the century Bell's nationwide market share was below fifty percent. In many cities, customers had their choice of telephone companies, and where competition existed, prices were lower.

Early in the Twentieth Century, Bell succeeded in developing a substantially improved long distance technology, which was protected by its patents. Customers wanted better long distance service, but it was made available only to subscribers of a Bell local system. The other companies counterattacked by interconnecting with each other, but the advantage of Bell in long distance could not be overcome. Customers began to switch to AT&T service, and Bell began a rapid move to acquire as many local competitors as possible.

The AT&T acquisition wave led to the first major antitrust interventions against the company, one seeking divestiture of newly acquired Western Union to preserve competition in telegraphy, and the other responding to the ongoing demise of local competition. The government succeeded in the first attempt, and Western Union was again made an independent company – a largely meaningless gesture in the new technological world created by telephony. With respect to the second, negotiations between AT&T and the federal government led to a settlement in 1913, the Kingsbury Agreement, that enshrined the "humming system" view of telephony. Bell agreed to allow itself to be regulated, and not to acquire noncompetitive local systems. It also agreed to let other local telephone companies interconnect to its long distance system. In return, Bell was allowed to keep its long distance monopoly and to acquire monopoly status in areas where it already supplied local service competitively. The upshot of the agreement was that the Bell System secured one hundred percent of the long distance market and more than eighty percent of local customers for most of the rest of the century.

One legacy of the Kingsbury Agreement was a peculiar federalist structure of regulation. Regulation of the interstate portion of the telephone system was granted to the Interstate Commerce Commission (which also regulated the railroads), while intrastate service (both local service and shorter long distance) was to be regulated by the states. Of course, the distinction was artificial, because the entire system was a single network. Hence, federal and state officials were forced to find some way of dividing the company's investments for purposes of defining jurisdictional authority. Initially, the problem was not severe, for the system was simple and, in any case, long distance was not very important.

But eventually, as the uses of the network became more complex and long distance became not only important but crucial to some new information services, the jurisdictional boundary had substantial political significance, a point to which we will return.

Within a decade, the structure created by Kingsbury was again controversial. The main point of controversy in the era between the two world wars concerned AT&T's vertical integration into equipment manufacturing. Virtually every piece of equipment in the Bell System – switches, transmission facilities, customer premises equipment, and even the wire inside a residence connecting the telephone to the network – was manufactured by one division of AT&T, then owned and rented to customers by another. Because Bell's local monopoly was not complete, other telephone companies were a potential market for equipment. And because the U.S. is such a large country, even fifteen to twenty percent of local service was enough to attract quite large, efficient companies (the non-Bell market was comparable in size to one of the largest European countries). Once these companies succeeded in selling to the independent local exchange carriers, they naturally wanted to sell to the Bell System as well. The unwillingness of AT&T to allow them to bid on Bell System procurement gave rise to the next era of controversy.

The debate over the industry structure culminated in the passage of the Communications Act of 1934. One version of the bill would have divested equipment manufacturing from AT&T, but eventually the bill that passed called upon the newly created Federal Communications Commission to study the allegations that AT&T was not an efficient manufacturer and that it unfairly denied competitors the possibility of selling to its affiliates. The FCC was to report back to congress in five years; however, by the time the report was prepared, the war had begun, and the government set aside telecommunications policy.

The 1947 Antitrust Action and Its Aftermath

Soon after the end of the war, the FCC's 1939 report began to be discussed within the government. The FCC had concluded that vertical integration was unnecessary and probably pernicious. In response, the Truman administration began the third antitrust investigation of AT&T, and ultimately filed a complaint demanding divestiture of manufacturing activities. But before the case was litigated, Eisenhower succeeded Truman as president, and immediately put in motion negotiations to terminate the antitrust proceeding. The deal that emerged appeared favorable to AT&T, but the company would soon regret it. Basically, AT&T was allowed to retain its vertically integrated telephone monopoly, but it was required to allow anyone to use its patents on semiconductors (which were invented at Bell Labs) without royalty, and not to enter any business other than telecommunications (including computers and semiconductors for use by anyone other than a telephone company).

The importance of the decree was that it kept the nation's most technologically adept electronics firm out of the two most promising new industries of the century. As a result, new firms rapidly invested in both production and research in a wide variety of microelectronic and computer technologies that were technically very similar to the underlying technology of telecommunications. These firms rapidly became as sophisticated as AT&T in these other activities. Moreover, like the independent telephone suppliers of the pre-War era, they perceived numerous opportunities for selling new equipment to either Bell System telephone companies or their customers. These developments were quite unlike those in other advanced economies, where in most cases the telephone service monopoly did not manufacture its equipment, and telephone equipment manufacturers were permitted to enter technologically related industries. Consequently, in other countries the dominant computer and microelectronics companies were not separated from, and in conflict with, the dominant manufacturers of telecommunications equipment.

The position of U.S. regulatory policy in the period immediately after World War II was that the national telecommunications system should remain a monopoly. But large businesses saw no reason why they should not be able to supply their own internal communications needs. Moreover, as time-sharing computers with remote terminals became available in the early 1960s, companies also sought to combine internal systems for telephones with their computer systems. Because Bell could not manufacture or sell computers, the computers that were connected to a company's telephone system were pieces of terminal equipment not owned by the telephone company. Why not the telephones as well? Or, why not build one's own long-distance telephone link to connect the computers and telephones in a company's various facilities, avoiding the telephone system entirely for all intracompany business throughout the nation?

Both large users of telecommunications services and the rapidly growing computer and microelectronics industries advocated relaxing the restrictions against purely private alternatives to the telephone system. Eventually this pressure led to three significant policy changes: Above 890, Carterfone, and MCI. Above 890 refers to a decision in 1959 to allow companies to build private microwave communications links for their own use. Carterfone refers to a decision a decade later to allow use of a radio telephone that was not manufactured and owned by AT&T. MCI, decided a few months after Carterfone, allowed a company to sell private long-distance telephone links to multiple businesses over a single, independent network. All three decisions brought forth still more entrants, leading to rapid growth in customer-owned equipment and lines, and to competing small, specialized carriers that sold forms of long distance connection other than standard, voice-grade, dial-up station-to-station toll calls. By the mid-1970s, AT&T had to permit its customers to own their own telephones, to construct their own private networks, and to buy dedicated point-to-point connections from alternative carriers.

Nonetheless, the liberalizations from the late 1950s to the early 1970s did not fundamentally challenge AT&T's structure. The company still bought equipment only from itself, and the vast bulk of telephone service – ordinary telephone calls, whether local or long distance – still used the AT&T network. Moreover, regulatory policy stopped short of forcing AT&T to give up these two remaining monopolies. MCI formally proposed to the FCC that it be permitted to offer ordinary long distance toll calls over its network, but the Commission refused, stating that it held firm to the policy of a single, monopoly backbone network for ordinary telephone service.

The Post-1975 Restructuring

Facing little prospects for further reform at the FCC, the competitors took their case to court. MCI appealed the FCC's ruling on competitive long distance toll service, and contacted the Department of Justice to inform the Antitrust Division about AT&T practices that MCI thought were intended to thwart competition. Meanwhile, numerous equipment manufacturers also complained to Justice, contending that AT&T was inhibiting competition in terminal devices and refusing still to buy equipment from anyone other than its own manufacturing divisions.

The FCC lost its long-distance decision to MCI in 1978. The court ruled that U.S. regulatory policy is based on the assumption that competition is preferred, and that regulation should attempt to approximate competitive conditions only when competition is not viable. MCI's willingness to compete with AT&T, therefore, should be welcomed, not opposed, unless the FCC could find evidence to support the conclusion that consumers would be harmed by MCI's entry. Note that the decision cleverly switched the burden of proof. Instead of MCI having to prove that its service would benefit consumers, AT&T and the FCC had to prove that MCI's entry would harm them. The Commission did not attempt to offer such proof, but instead simply opened the door to long distance competition.

Meanwhile, the government filed its antitrust suit against AT&T, requesting that the company be divested, separating local service, long distance and manufacturing into three separate, unrelated businesses, and further dividing local service into several companies. The case was filed despite opposition from several other federal agencies, including the Department of Defense, which for reasons of national security valued having a single, integrated long-distance carrier. In both the Nixon and the Carter administrations, other agencies attempted vigorously to convince the Antitrust Division to alter its objectives and the White House to intervene to stop or at least to change the course of the litigation, but to no avail.

AT&T turned to congress to seek reversal of both the MCI decision and the government's filing of the antitrust case. The company began a

five-year lobbying campaign in support of a new communications act that would legally enshrine the concept of a single backbone monopoly telephone network, provided by a vertically integrated entity. The proposed bill, if enacted, would have mooted the government's case and vacated the licenses of AT&T's long distance competitors. Twice the bill came perilously close to passing, actually being adopted by one house of congress at one point. Then, in 1981, most analysts believed that the election of Ronald Reagan would lead to the same fate as befell the Truman administration's antitrust attack. The Reagan administration opposed the case; however, the entire administration, from the President down to the Attorney General, was not permitted to participate in making policy in this area because of their previous associations with AT&T. Hence, a middle level official, the Assistant Attorney General for Antitrust, William Baxter, had total control of the litigation.

Although a Reagan appointee, Baxter was also a strong advocate of competition and an opponent of regulation. He saw divestiture of AT&T as a way to encourage competition and deregulate the industry. Soon after taking office, he announced that he would litigate the case to conclusion.

Unlike the Nixon, Ford and Carter administrations, the Reagan White House opposed the antitrust case, but was powerless to stop it. Hence, Baxter's superiors turned to congress, requesting legislation that would force the case to be dropped. After months of debate in a very even fight, the AT&T bill failed a final time. Within two months, AT&T gave up, offering a settlement that gave the Antitrust Division most of what it wanted. After a century, the forces for liberalization had finally won. The telephone system was going to be structurally dismembered, and nearly all of it opened to competition.

The lesson in the final events between 1975 and 1984, when divestiture finally took place, is profound. Not only did a liberalization policy never formally pass congress or enjoy open support from the president, it almost certainly never could have passed as a legislative program. In this case, two court decisions and a crusading bureaucrat whose superiors could not intervene turned the industry on its head. Of course, just as the policy never passed, so, too, the reversal of the policy also failed. In essence, whatever the FCC, the courts and the antitrust authorities managed to concoct during this era could not be reversed politically. In the end, congress was unwilling to pick a winner in a battle in which both sides were represented by very strong, well organized interests. AT&T, other local exchange companies, and state regulatory officials opposed liberalization, while numerous companies in computers, microelectronics and telecommunications equipment and some large business customers favored it.

Of course, underpinning the story is the constraint of technical and economic reality. If the advocates of change had not had a reasonable basis in believing that they could participate effectively in the industry, government officials probably would not have let them try. In one sense,

"new ideas" in the form of scientific advancement caused more companies to want to compete with AT&T in a wider variety of markets. Technology had developed at an astounding rate, and new service ideas were emerging in numerous companies besides AT&T. Nonetheless, the notion that important segments of the industry could be competitive was certainly not new. To the extent new ideas played a role, it was a political one – increasing the number of organized interests that sought to compete.

As in other instances of liberalization, the decision to permit competition, because competition was viable, expanded its own support constituency after the fact. Firms, employees, stockholders and customers of the successful entrants lobbied for the preservation of liberalization, but they were fewer and less well organized before it took place. Consequently, liberalization was difficult both to start and to stop. In this way, the ability of the FCC to take gradual liberalizing actions made possible the more radical changes that occurred later.

Policy after Divestiture

Because telecommunications liberalization was gradual, spanning more than two decades, one cannot make a clear before and after comparison between the eras of regulated monopoly and competition. Moreover, the facts are further clouded by the rapid rate of technological progress in the industry. Here I will focus on surely the most important and dramatic event, the divestiture of AT&T, and the regulatory policies that followed.

Divestiture was by far the most important reform for several reasons. First, in all segments of the telecommunications industry except customer premises equipment, AT&T had not lost its dominance during the earlier reform period. Before divestiture, it still manufactured 85 percent of telephone company equipment, still controlled over 90 percent of long distance service, and was positioned to control information services and new radio telephone technology. Second, divestiture fragmented the local exchange business. Instead of one large company with 85 percent of all local access lines, the nation had seven independent Bell System local access companies, each of which was only slightly larger than the largest predivestiture independent, General Telephone. Thus, the equipment procurement industry had been made substantially more competitive on the demand side than it was on the supply side, where AT&T was still dominant.

The Basis for Divestiture

The theory of divestiture was that entry would make the supply side competitive. No telecommunications market was made structurally more competitive by divestiture. All it sought to do was make competitive entry

possible. Thus, it relied on the belief that AT&T, as a monopoly, was sufficiently inefficient and lethargic that it could not retain is dominance without the protection of vertical integration. Fragmenting local exchange carriers by creating seven Bell Operating Companies facilitated this objective by reducing the likelihood that a cozy bilateral monopoly between service and manufacturing would survive organizational separation.

The divestiture agreement set forth rules concerning which firms could compete in which markets. The agreement first created a new geopolitical entity, the Local Access and Transportation Area, or LATA. Within these areas, local exchange companies (the Bell System companies that provided basic access) could provide ordinary telephone service. Approximately 160 LATAs were created, all considerably larger than a single local exchange, and some consisting of entire states. Thus, the Bell Operating Companies were given the right to provide long distance service inside LATAs, which accounts for approximately 25 percent of all long distance revenues.

Telephone calls crossing a LATA boundary were required to use a long distance carrier. All long distance carriers were to be provided with "equal access" to customers, meaning that the carrier designated by a customer would automatically be the one used when the customer dialed a long distance number. To comply with equal access, the BOCs were required to spend billions of dollars in new switching equipment, creating a huge market for digital switches in the U.S. The equal access provision, although adopted to facilitate long distance competition, also served to encourage competition in equipment manufacturing by creating a very large surge in demand immediately after the Bell Operating Companies were opened to competitive suppliers.

The local exchange carriers were prohibited to enter a variety of businesses, notably equipment manufacturing, interLATA long distance, and information services (recently, the information services prohibition has been relaxed). Thus, the idea of the decree was to focus BOCs on providing the backbone system of local access and to let others supply the equipment for building it, the equipment for customers, the services (other than ordinary telephone service) that would be provided, and the long distance network connecting local systems.

The rationale for the line-of-business restrictions was that without them local exchange carriers could use their monopoly in basic access to advantage themselves unduly in other markets. Regulation could not protect against self-dealing. When technology is evolving rapidly, numerous small technical decisions must be made, and these can determine whose proprietary equipment and software works best in the network. Regulators can never hope to second guess these decisions, and even if they try, the primary effect would be to slow and to distort technological progress.

Implicit in this view was a belief that the value of diversity and competition exceeded the economies of network integration. In defending itself at the FCC and in court, AT&T asserted that telecommunications ex-

hibited important economies of integration: economies of scale in network components, and economies of scope in the joint design of the entire network. In practice, these economies were quite limited, and largely confined to basic access. Switches had become modular with digital technology, and afforded no further economies after a few hundred ports. Transmission had scale economies, but hardware costs had shrunk to about ten percent of the costs of long distance carriers. Hence, organizational efficiency and marketing effectiveness had become the most important determinants of the performance of a long distance carrier.

Other arguments were also advanced to support a single national monopoly. One was the pursuit of universal service: the provision of telephone service to every household. Under regulation, prices for telecommunications services had levied a heavy usage-based tax on long distance that was used to cover the costs of basic access. The result was a massive subsidy of access in small towns and rural areas. The cost of basic access is governed primarily by the distance of a customer from the local switch, and in areas with low population density these distances are much greater than in larger cities. Hence, by the time of divestiture, both business and residential access prices in small communities were far below cost, paid for primarily by prices far in excess of costs for long distance telephone calls.

The universal service argument was a red herring. Separation of long distance from local service companies did not prevent the latter from charging the former for use of the system. Moreover, before divestiture, the FCC was moving in any case to reduce the subsidy by reducing usage charges on long distance. The cross-subsidy policy was creating enormous economic inefficiencies because the commodity being taxed, long distance, had a highly elastic (price-sensitive) demand, while the commodity being subsidized, access, had an almost perfectly inelastic (price-insensitive) demand. Hence, price reductions for access had essentially no effect on the penetration of the telephone system, but significantly curtailed long distance usage. On balance, consumers would be better off paying more for access and less for long-distance. The best way to assure universal service was to use higher access fees in urban areas to offset the higher costs in small communities. In any case, how the subsidy was raised had nothing to do with the structure of the industry.

The Regulatory System

Until a few years after divestiture, the telephone system was regulated jointly by federal and state regulators using the traditional technique of rate-base regulation. The basic approach was to set a revenue requirement for the company that would cover all reasonable costs, plus a reasonable rate of profit. Because state and federal regulators shared responsibility, a key part of regulation was an arbitrary division of the costs of the system between the two jurisdictions. The agreement was that the cost of basic

access would be divided between the two on the basis of relative use. Hence, the fixed cost of a telephone access line was divided between the jurisdictions on the basis of quantity of output, and thereby implicitly was treated as a variable cost.

A second feature of the system was "residual pricing". All services other than basic access had prices set approximately to maximize their revenues. Basic access was then priced at whatever level was necessary to cover the rest of the costs of the firm.

The regulatory system had two primary effects. Most obviously, it subsidized local service in small communities. Access prices were approximately the same everywhere, but costs were far higher in rural areas. The less obvious part had to do with the effects of long distance competition on the subsidy, and it was here that long distance competition created havoc in regulatory policy.

AT&T had designed its switches so that only one long distance carrier could connect to them for ordinary dial-up long distance. A competitor entered by simply becoming local subscribers. Customers would dial the competitor's local number, reach its nearest switch, and then dial a long distance number and a billing number. Because the AT&T component of the call was entirely local, AT&T collected none of the surcharge on long distance that was paying for rural local service. Moreover, because of residual pricing, every penny of shortfall in long distance surcharge went straight to an increase in basic monthly access fees. Hence, competitors could charge a much lower long distance price than AT&T, not only taking away AT&T's customers but forcing AT&T to raise its other prices. Before divestiture, none of these effects were very substantial, for the long distance carriers had a very low market share, but the rapid growth of competition constituted a serious threat to the entire system of price regulation.

A common view was that only the perversities in the system encouraged entry, but this was clearly not the case. When the competitors were offered equal access (and an equal surtax on their calls), they chose it rather than the old system, for most customers did not want the trouble of dialing all the numbers that were required to use a competitive long distance carrier. In addition, for technical reasons, the old system substantially degraded the quality of the competitor's long distance service.

Not surprisingly, the effect of entry divided the regulators. State authorities set the basic access price that the FCC's pro-competitive policy was forcing upward. The FCC, meanwhile, reaped the political benefit of declining long distance prices arising from both technological progress and competition. State regulators, therefore, joined AT&T in opposing competition.

Divestiture caused the positions of the divested parts of AT&T to divide as had their regulators. Suddenly the new AT&T, consisting of equipment and long distance, did not want to subsidize its former part-

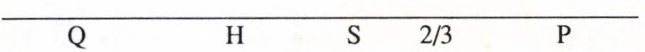
ners, the operating companies. By dividing the company, divestiture gave momentum to reform of the pricing system.

The primary price reform sought by the FCC was to eliminate the subsidy of access by long distance and other usage-based charges in the federal jurisdiction. (Federal and state regulators still dispute the boundary of jurisdictional authority, with both claiming responsibility for the interconnection prices for information service providers.) By the time of divestiture, usage-based long distance charges were paying for 26 percent of the costs of local access. The FCC's proposal was gradually to replace these charges with a fixed monthly fee for basic access to long distance. State regulators and local exchange carriers, believing that customers would not heed the distinction between the federal and state monthly charge and so would blame them for what appeared to be an increase in monthly subscriber charges, opposed the plan, and appealed to congress.

As with the previous AT&T legislative proposal to forestall competition, state regulators and local exchange carriers came very close to victory. The House of Representatives overwhelmingly passed a bill eliminating the monthly federal access charge to residential customers, and the bill went on to the Senate. Just before the Senate was to act, the FCC began trimming its sails, eventually proposing to cut the residential access charge in half, and to phase it in over a longer period. After taking these actions, the Senate defeated the House bill by a margin of one vote.

These events can be more easily understood with the aid of a diagram depicting the positions of the parties, as shown in Figure 2.1. The line represents the amount of an access charge,

Figure 2.1: The Politics of Access Charges



with the right-hand side corresponding to higher prices. The status quo (Q) is a zero basic access fee for the federal share of local exchange costs. The position of the FCC and the president (P) was for shifting all of the federal share for fixed local access costs to a monthly charge. The House position (H), reflecting the position of its Democratic majority, favored monthly access charges only for business. The Senate position (S), reflecting that body's Republican majority, was more sympathetic to monthly access charges proposed by an administration of the same party, but it feared political reprisals from constituents when a highly visible fixed monthly charge was substituted for a less visible long-distance usage fee. Hence, the Senate preferred H to P, although it was not very close to either position.

Once the House passed a bill representing its most desired policy outcome, the Senate would go along with the House rather than allow the

FCC proposal to take effect. But the FCC favored the Senate's most preferred outcome, S, to H, and sought to forestall legislation by adopting S. (In fact, the FCC could do even better by adopting the point 2/3, the preferred policy of the member whose vote is needed to sustain a presidential veto of a bill enacting any access charge policy between 2/3 and Q; however, this action would require that the president bear all of the political costs of a monthly access charge and also conflict openly with the Republican leadership in the Senate, so that an agreement to settle for S is easy to understand.)

Although settling for less than its preferred position, the FCC nonetheless profoundly changed the structure of telephone price regulation. Eventually, all of the fixed costs of local access probably will be collected through a fixed monthly charge, but even in the short run most of these costs would not continue to be covered by a surcharge on long distance.

Post-Reform Performance

Liberalization of telecommunications policy had very little effect on the overall performance of the industry until the late 1970s. Until customer equipment was liberalized and long-distance competitors were permitted to offer ordinary toll service, competition was too limited and too small a fraction of the industry for its effects to be detected in overall industry data. Systematic data on telephone prices have been collected since 1935, and they reveal that from 1935 until 1980, the price index for all services rose about two percentage points per year less rapidly than the overall consumer price index. Basic monthly service for residential users also rose less rapidly than inflation.

After 1978, when liberalization began to have an important effect, nominal telephone prices rose far more rapidly than ever before. The index of all telephone prices rose by 4.3 percent annually between 1978 and 1988, and the average basic residential access price (including the new FCC access charge after 1986) more than doubled, from $8.32 on January 1, 1980, to $16.66 on October 1, 1987. These data were the source of the controversy in the mid-1980s about the wisdom of liberalization, structural reform, and the shift from usage charges to monthly access fees; however, these figures do not indicate declining performance in the industry. About one-third of the price increase reflected higher real interest rates in the U.S. since the late 1970s, which significantly affected telephone prices because the industry is so capital intensive. Moreover, the industry's price performance retained roughly its relationship to overall inflation as the telephone price index rose 1.8 percentage points less rapidly than consumer prices generally. Thus, the very large rate of increase in residential access – about ten percent per year from 1980 to 1987 – was offset almost exactly by price reductions in other services,

especially interstate long distance (the service regulated by the FCC that benefitted from the federal monthly access charge), where nominal prices declined by an average of seven percent per year from 1983 (the last year before divestiture) until 1988. Intrastate long distance prices, regulated by the states and by regulators who opposed divestiture and liberalization, also fell, but not as dramatically, with nominal prices declining about one percent per year.

The figures on service prices understate the beneficial effects of liberalization for four reasons.

First, the aggregate price data were affected by the massive investment program that local carriers undertook in the mid-1980s. Carriers planned to replace analog switching with digital switching, but the equal access provision of the divestiture agreement accelerated this replacement. Thus, the capabilities and capacity of the telephone system increased substantially in the 1980s, with no dramatic effect on overall price trends.

Second, the data do not take into account the effects of divestiture and competition in customer equipment. From ordinary telephones to complex equipment for business computer communications, prices have fallen substantially while quality and variety have expanded. In most lines of customer equipment, AT&T is no longer anything more than one of many players.

Third, since the mid-1980s many new service providers have entered the industry. Literally millions of customers now use computer bulletin boards, electronic mail, electronic credit verification, airline computer reservation systems, and electronic shopping services. In principle, all of these services could have been provided by an integrated AT&T; however, prior to liberalization of the industry, these services were not developed extensively, and have not been as extensively developed in other advanced industrialized societies where liberalization has not taken place.

Fourth, separation of radio communications systems from traditional telephony has been followed by a vast expansion of services using over-the-air transmission. The most impressive development is the explosive growth of mobile communications involving cellular radio and other new forms of mobile services that are permitted to interconnect with the telephone system.

The complete picture is that since the late 1970s, the U.S. has experienced an explosive growth in communications services. Whereas people still disagree about the extent to which liberalizing reforms are responsible for these events, there can be no disagreement that segmenting the integrated national system did not seriously erode performance in either the price of ordinary telephone services or the rate of progress in equipment and new services.

The Causes of Change

The dramatic restructuring of the telecommunications sector in the U.S. constitutes something of a puzzle to explain. Certainly a major cause is advancements in technology, which reduced the importance of the hardware costs of the backbone network in the overall economics of the industry, and vastly increased the scope of uses to which the telecommunications system could be put. But technology advanced everywhere, not just the United States, so it cannot be the sole cause of liberalization and, especially, restructuring.

The role of interest groups also played an important role. U.S. telecommunications policy clearly was affected by the presence of companies such as IBM, Hewlett-Packard, ITT, and Motorola, to name a few, who were highly successful in various aspects of the electronics sector and perceived themselves as having something important to gain by liberalization in telecommunications. Here the U.S. situation differs from circumstances in some other advanced economies, especially with respect to the extent of competition in these industries. But support from potential competitors does not appear to be either necessary or sufficient for reform. In Great Britain, British Telecom and Plessy dominated the domestic industry, with little organized support for liberalization. In France, the major communications equipment supplier (Alcatel) was separate from the nationalized service supplier and the major computer firm (Bull). Moreover, when the reform era began, both manufacturers had substantial foreign ownership (ITT and Honeywell), presumably giving them less political clout. Yet France has hardly been a leader in liberalization. To the contrary, it has nationalized equipment and pursued a policy of greater centralization.

One factor that was quite different in the U.S. was a fundamental change in political representation just as the reform era began. In the mid-1960s, the U.S. Supreme Court substantially changed the method of electing both state legislators and members of the House of Representatives, ruling that the Constitution required that all legislative districts except in the Senate had to have precisely the same population in the first election after a decennial census. In many states, legislative districts had been largely unchanged for decades, so that rapidly growing urban areas did not receive proportionately more seats as their fraction of the total population increased. As a result, some urban districts were as much as three times larger than some rural districts, and over half of the House was elected from districts in which over half of the population lived in small towns or rural areas – even though only about 25 percent of the population lived in such communities.

By the early 1970s, legislative districts had been redrawn to equalize representation. The effect was to reduce the representation of rural interests. Regulatory policy in all areas, but especially utility services, had been based on the idea of universal service – making sure that everyone

was offered service at an affordable price. Because rural service is far more costly to provide in areas with low population density, regulators developed the policy of cross-subsidization. In telecommunications, long distance service was accorded a secondary priority to basic access, and so was taxed to cover the difference between price and cost in high-cost areas.

With redrawing of legislative districts, the political support for rural subsidies can be expected to decline with the decline in rural representation. Hence, the opposition to liberalization in the legislatures of the mid-1970s ought to have been substantially less than it would have been a decade earlier. Because three times key features of the liberalization program were nearly stopped by legislation (the 1976 AT&T initiative, the 1981 Reagan attempt, and the 1986 access charge controversy), it is certainly plausible that without the Supreme Court's reform of representation, the U.S. today would not have a telecommunications policy that was greatly different from policy elsewhere.

One aspect of post-reform pricing bears out this theory. Between 1983 (the last year before divestiture) and 1986, the relationship between rural and urban business access prices changed dramatically. The average state-regulated business basic access price rose approximately $6.00 per month in local exchanges serving fewer than 5,000 terminals. (A terminal is any piece of customer equipment that is connected to the network, and the population served by a local exchange is about twice the number of terminals. In large urban areas, central office switches normally serve tens of thousands of terminals, so that a locality with 5,000 or fewer terminals is almost always a small, isolated town or a rural community.) By contrast, in local areas with 1,000,000 or more terminals, corresponding to the 25 or so largest metropolitan areas and containing about half of the total population of the U.S., basic business access prices fell by $0.22 during the same period. In addition, after 1986 businesses paid the new $6.00 per month federal access charge that financed the reduction in interstate toll charges.

For residences, the change was not as dramatic: the increase in smaller communities was about $2.25, compared to an increase of $1.87 in the largest cities. When the $3.00 monthly federal access charge is included, the fair conclusion is that no perceptible change took place in the structure of residential access prices. But the increase in residential access charges caused the price of service (including the federal charge) to be approximately equal to its average cost in most cities. Rural residents paid about $2.00 a month less for service, but were still heavily subsidized. The source of the subsidy switched, however. Business customers everywhere now pay over $30.00 per month, a price that substantially exceeds all estimates of the average cost of local service in all cities exceeding about 25,000 in population. Moreover, although rural businesses still pay less, the difference between business service prices in large cities and small communities dropped by more than $6.00 per month. Thus, the

magnitude of the rural subsidy fell as a consequence of liberalization, an effect that is broadly consistent with the results expected from changes in the system of political representation.

Rural subsidization has not ended, of course, but there is no reason to expect such a dramatic result. Rural industries such as agriculture, forestry and mining constitute a large part of the economy. Agriculture, although it accounts for only about three percent of GDP in the U.S., is still among the largest industries. These industries are well-organized and so can be expected to be politically effective. Moreover, voters do not appear to be so strongly self-interested that they want to end rural subsidies, for the idea of nationwide rate-averaging for basic utility services appears to have broader support than just from rural constituencies. But the theoretical prediction is a modest one – the decline in rural representation should lead to a decline in rural subsidies – and this prediction is confirmed by the data in telephone pricing.

Conclusions

As recently as 1986, with the congressional reaction to access prices, the liberalization policies in the U.S. always seemed perilously close to reversal. The path of change had been slow, primarily because rapid change simply would not be tolerated by the political system. But at each crisis point, the restructuring and competition policies managed to survive. By the late 1980s, the performance of the industry was so high, and improving so rapidly, that a popular movement to roll back the clock was unlikely from either congress or the public at large.

Nevertheless, the U.S. experiment is not over, and still could be significantly reversed (although never to even the structure of 1980). The reform era was made possible in part by key "noncomplying" actions by bureaucrats or judges who advanced policy reform without the support of legislation or the president. The reversal of the FCC's decision prohibiting competitive entry into toll service and the success of the Assistant Attorney General for Antitrust in fending off his own president's attempt to kill divestiture are key events in the history of the new structure and policy.

The lesson is that similar surprises from judges or regulators could occur again. Noncomplying acts by bureaucrats and judges create a new status quo, and sometimes a new organized interest to defend the new policy. In the more complicated process of policy change in the legislature, noncomplying acts benefit from the status quo bias.

The most important change that could be imposed without direct political ratification would be to allow local exchange carriers to enter manufacturing, information services, and long distance. These actions would run the risk of causing a loss of competition in at least some important parts of the industry. Many U.S. officials – including the Bush Adminis-

tration FCC – favor eliminating at least some of the restrictions on business activities of local exchange carriers, although not because they openly favor eliminating competition. Instead, the position of the Bush FCC is standard neoconservative noninterventionist economic policy: let companies do what they want, in the expectation that they will then have an incentive to innovate and to provide better service.

Eventually, local exchange carriers will be permitted to reintegrate. The issue is when. Advocates of structural separation of regulated monopoly from competition favor permitting reintegration only in activities in which competition is secure from self-serving, discriminatory practices by a monopolist due to the state of technology and the market. Regulators are viewed as part of the problem, not the solution; better regulatory policy is not viewed as an effective countermeasure to monopolization of competitive markets by a reintegrated local exchange carrier. For most residential customers, this point of view implies relaxation of line-of-business restrictions only when residential users enjoy an alternative to the local wireline carrier for basic access. Cable television, cellular radio, or one of several other emerging technologies might provide alternative access, but probably not until the end of the Twentieth Century at best.

If the past is a guide, policy will steer a course between the extreme positions, following a path of gradualism. Local exchange carriers will be given more and more exceptions to the restrictions handed down by divestiture, but not sweeping ones that would threaten competition in any component of the industry. Hence, the advocates of line-of-business restrictions will win more than they lose, and reintegration probably will not take place until other technologies have a chance at competing effectively with local telephone companies for basic access service. The negative side of this policy is that gradualism fans the expectations of local exchange carriers about their eventual entry into other parts of the industry. These expectations suppress their willingness to create the technological capacity in the local network for a variety of new services and new customer equipment. For example, some Bell Operating Companies have refused to offer new forms of ISDN interconnection, which their political opponents attribute to their unwillingness to give others a first-in advantage while the local carriers are prohibited from offering these services. Of course, if new services and equipment cannot be offered, the competitive conditions perceived as necessary for relaxing restrictions on the BOCs can never be met. Meanwhile, another judge could simply vacate the restrictions, and the structural experiment would be over prematurely.

The lessons from the U.S. reforms cannot yet be conclusively drawn, and indeed may never be clear. Because technology is evolving so rapidly in the information sector, which countries and nations are luckiest in the innovation race will play a major role in deciding which country's information sector progresses most rapidly. The U.S. could win the race for

largely extraneous reasons (e.g. its strong university research capabilities in engineering and physical science). Or, the fragility of the institutional support for the U.S. experiment could cause its abrupt end before its consequences can be fully apparent.

Nevertheless, throughout the reform period, the progress of liberalization has been associated with a certain sense of technological inexorability. Individual policy decisions may have hastened or slowed the pace, but the size, fragmentation, diversity and state of technical advancement of the U.S. information sector (including microelectronics) is probably the single most important force affecting policy in the long run. The controversies regarding structure are rightfully debated with vigor, for they can not only determine the winners and losers in the current round of technological advance, they can also substantially affect how quickly the next advance will come. But in the long run, it is quite likely that the genie has been permanently set loose from the bottle: an essentially competitive, minimally regulated industry is almost certain to be the future of the U.S. industry. And, if performance since the mid-1970s is a valid basis for prediction, the U.S. will certainly be no worse off from the experiment, and conceivably will be far better off – sufficiently so that other advanced economies will be forced to follow a similar path due to competitive necessity.

Lecture #3 – Airlines

Economic deregulation has been close to complete in the entire U.S. transportation sector, and airlines are no exception. Indeed, the agency that was created to regulate the airlines in 1938, the Civil Aeronautics Board, was abolished in 1985, with a few remaining regulatory functions transferred to the Department of Transportation. Thus, nowhere can the promise and problems of deregulation be more thoroughly assessed than in transportation.

The airline industry is an especially fruitful example to study because of some interesting, unique characteristics that have raised several important policy issues during the deregulation era. For the most part, deregulation has been simpler to understand in the other transportation industries (trucking, railroads, buses, domestic water carriers). Briefly, deregulation has been an unqualified success in freight shipping. It has reduced transportation costs immensely, saved the railroad system from financial collapse, improved service quality, and increased efficiency in all parts of the sector. The losers from deregulation have been some inefficient transportation companies that simply could not compete effectively, and some categories of organized labor. Part of the cost of regulation was especially high wages and inefficient work rules that could not be sustained in a competitive industry, even a unionized one. Although employment has increased in the transportation sector, the initial effect of deregulation was loss of jobs and wage reductions in many companies.

Airline deregulation has had a similar performance, but the history of the deregulation era is more complex (and the results more controversial) for three fundamental reasons.

First, airlines are very sensitive to the business cycle and to fuel prices. Hence short-term economic fluctuations and disruptions in the world oil market cause booms and busts in the industry. Inevitably, advocates of deregulation are prone to claim undue credit for the booms (the mid-1980s, when economic growth was rapid and oil prices were falling), while opponents cannot resist the temptation to attribute the busts to deregulation (the 1979-82 and 1990 combined economic downturns and runups in oil prices).

Second, more than other transportation modes, the airline industry produces a highly differentiated product. The basic reason is that airlines primarily carry people, not goods, and people care a great deal more about the speed, comfort and convenience of travel than do cartons of freight! Because air travel times are short, a small change in the effi-

ciency of a route (how many stops?) or the frequency of service (when do I have to leave to make my 11 a.m. appointment?) can have a substantial effect on the implicit time cost of a trip. Because time is more valuable to people than to goods, the details of the elapsed time of a trip are more important, and are a significant source of product differentiation among flights and airlines. In addition, the amenities of the trip – meals, in-flight service, seating comfort – provide another potentially important source of product differentiation.

Third, aircraft investments are more lumpy, and the details of interconnection more important, than in other transportation industries. For example, a railroad has a great deal of flexibility in the length of a train, and can disassemble the cars of one train to reassemble them at an intermediate destination to form new trains to continue different journeys. Moreover, because of the differences between freight and passengers with respect to the value of time, the freight modes have more flexibility in scheduling connections at intermediate points. And, for trucks and rails, a single load (a car or truck trailer) is normally a far smaller fraction of total traffic between two points than a single flight. Consequently, coordination among flights is economically very important for airlines, giving rise to, first, an advantage for large, nationwide carriers, and second, a need for "interline" ticketing coordination. The former leads to oligopoly, and the latter can facilitate collusion, so that policy is more likely to face the potential problem of market power in airlines than in other transportation modes.

These three fundamental characteristics of the airline industry have made the deregulation era intellectually more interesting in the airline business than elsewhere. Indeed, in airlines deregulation probably encouraged rather than discouraged research on the economics of the industry, which is the opposite of the effect with respect to the other transportation modes. This research has vastly improved our understanding of the industry, providing explanations for post deregulation developments that had not been predicted by either research scholars, industry participants, or government officials. Likewise, the political events surrounding airline deregulation and subsequent calls for at least partial reregulation shed considerable light on the politics of reform.

The generalizability for other nations of the lessons in airline deregulation, and deregulation of the other transportation modes, is less obvious than in utility industries, such as telecommunications and electricity. The reason lies in the economic geography of the United States. The U.S. is an extremely large country with a very low population density, and it is substantially more integrated economically than multinational trading groups like the European Economic Community. Consequently, both the supply and the demand characteristics of transportation industries differ between the U.S. and most of the rest of the world. Moreover, a substantial part of the investment in transportation is fixed costs in terminals and rights of way, which have been developed differently in the U.S. than in

most other nations. For example, American railroads are far less important for passenger service, and far more important in freight service, than are European railroads. Truck transportation in the U.S. is heavily advantaged by the far better and more extensive U.S. highway system and by lower fuel prices and road tolls. All of these factors would make the effects of liberalization differ between the U.S. and elsewhere.

Nevertheless, transportation regulation can be expected to have qualitatively similar effects everywhere: economically irrational route structures and price systems, and a general tendency to protect inefficient firms, to encourage excessive wages and inefficient work rules, and to use criteria other than cost and performance in making major capital investments. The key issue is not whether Europe or some other part of the world is like the U.S., but whether transportation can be sufficiently competitive that the inefficiencies of a less then perfect market structure are small compared to the inefficiencies arising from a protective regulatory umbrella.

Historical Origins

The airline industry had simple and highly competitive origins. Because early commercial aircraft were small and could fly only relatively short distances, the industry began with a very large number of very small companies, each serving a small geographic region. As technology progressed, the more successful firms expanded, some serving much of the nation, but the competitive structure of the industry remained.

The Role of the Postal System

The postal system played a major role in the development of the industry. A substantial part of the revenues of the industry during its first two decades consisted of fees for carrying air mail. Consequently, as part of postal policy, the federal government grew to have a major interest in how the industry operated. These interests focused on three main issues: costs, universal service, and reliability.

The postal service paid more for air mail service than the revenues it collected from its air mail customers. These losses arose from a problem within the postal system, which has always operated at a loss and subsidized many services. In this case, the price structure also preserved the postal monopoly from direct competition by airlines. Regardless of the reasons, the payment to the airlines for air mail service was commonly termed a subsidy, and the postal service (as well as its political overseers) constantly discussed ways to cut these expenditures. One possibility was to make airlines less competitive so that they could charge more for passenger service and use the increased revenues to cut the budget of the postal system.

Universal service, an issue in all regulated sectors, meant the provision of air mail delivery to the maximum number of cities. Because aircraft carried mail, people and freight together, universal mail service meant universal airline service for other purposes. In some cases, passenger and freight demand was insufficient to justify a frequency of service that provided universal daily air mail. Hence, universal service meant a subsidy for service to small towns. Price regulation allowed this subsidy to be buried in the price structure rather than form part of the government's budget. Of course, long after postal revenues become relatively unimportant to airlines, the universal service objective had become an article of faith regarding policy toward the industry.

Reliability refers to the persistence of service. Daily mail deliveries are regarded as a necessary part of postal service, undaunted by any calamity. Aircraft flown by private firms were a potential threat to reliability. A company simply had to fly regularly, and disruptions due to bankruptcy or equipment failure were not acceptable. Consequently, the postal system wanted to deal with large, financially secure companies, not the "fly-by-night" small companies, often owned and flown by former stunt pilots, that characterized the early industry.

The Creation of Airline Regulation

The instigating event for airline regulation was the Great Depression. As in other industries, the Great Depression – by definition – was associated with large price reductions below long-run average cost, and financial failures within the airline business. The industry sought help in the form of price regulation. Its position was strengthened by several much-publicized deaths in aircraft accidents during the 1930s, including the death of a senator. Officials from the larger companies blamed the deaths on the financial pressures facing small carriers, who, they alleged, responded to price wars by cutting maintenance, flying long hours, and otherwise shortcutting safety. The last argument was especially successful, for the view that economic competition leads to a reduction in safety persists despite a mountain of evidence to the contrary – perhaps a good example of rational ignorance concerning an important policy issue!

The argument that the airlines were "ruinously competitive" – incapable of surviving as providers of safe, reliable service in a competitive regime – carried the day, and in 1938 the industry was subjected to regulation of prices and route structure. The 1938 Civil Aviation Act created the CAB, which instituted a regulatory policy having four main features.

First, the industry was regulated according to rate of return regulation on an industry-wide scale. The CAB set a target rate of return, calculated industry-wide costs, and allowed prices to be set so as to return, on average, the target profit rate. In practice, this aspect of regulation did not produce a binding profit constraint. In general, airlines rarely succeeded

in earning the allowed rate of profit. The importance of the calculation was that it set prices based on industry average costs.

Second, the price structure was determined by a uniform pricing formula that decoupled price and cost on a particular route. In essence, the price on a flight (P) was linearly dependent on the distance of the flight (D):

$$P = K + cD,$$

where K and c are constants. Moreover, K was set too low, and c too high, in relation to costs, so that, in general, short-haul flights with relatively few passengers per day recovered less than the average cost of a passenger, whereas long-haul flights earned excess profits. The purpose, of course, was cross-subsidization of short flights from small towns to nearby large airports. A by-product of this pricing policy was that short-haul service was made artificially attractive, thereby assisting in the demise of the nation's bus and rail passenger service.

The third element of CAB regulation was entry control through route authority. The CAB focused its control of the structure of the industry on licenses to fly between two points, not on whether a firm could enter the industry generally. To sustain cross-subsidies, the CAB needed to prevent competitive entry on the routes with excess profits, and to find a means to encourage carriers to fly on the losing routes. It did this by pairing route awards: companies were rewarded for serving uneconomic routes by gaining protection from entry or the authority to enter highly profitable, long distance routes. In addition, to facilitate cross-subsidization within a single company, the CAB promoted mergers. For thirty years, it allowed no new carriers to enter, and it actively encouraged consolidation of the existing carriers.

The fourth important feature of CAB policy was that it did not attempt to control directly the quality of service, other than by being more generous in awarding new routes to carriers that promised better service. The only service constraint was through the process of route awards. In order to be granted a route that was already served by another carrier, an airline had to prove that the carriers already in the market were providing inadequate service. Usually, this meant insufficiently frequent flights. As long as a carrier served a route above some minimum requirement to keep out competition, service quality was not subject to regulatory scrutiny.

Superficially, the CAB system was well designed to serve the constituencies giving rise to regulation – the industry, small towns, and the postal service. The skewed price structure and system of route awards served the universal service objective. The overly generous target profit rate and the encouragement of mergers created a reliable, profitable industry.

Performance under Regulation

The airline industry that emerged under regulation contained basically three types of firms. First, a few very large firms served many large cities, plus numerous smaller ones. Second, a few more carriers were geographically concentrated in one region. Generally, the national carriers flew the east-west long distance routes, and the regionals flew the north-south long distance routes. In both cases, profits from long hauls were intended to subsidize uneconomic short routes. Third, a group of very small commuter carriers emerged due to a loophole in the regulatory policy. Aircraft below a certain weight limit were exempted from regulation. This policy enabled very small towns to obtain service to nearby major cities on small aircraft that passengers would normally not choose if given an alternative, due to the level of comfort, noise and safety of the aircraft. The advantage of the loophole was that it enabled some routes to be served before it was feasible to do so with large aircraft flown by a major carrier.

For the first two categories of airlines, the regulatory policy of the CAB set up a quite perverse incentive structure that produced substantial inefficiency in the industry. The first source of inefficiency was the route system that emerged from CAB entry controls. Because a political process, not costs and demand, determined route awards, the route structures of the airlines made no sense economically. Hence, airline service was more costly than it would have been in a regime in which airlines had been able to select their own routes. A *New York Times* advertisement unkowningly made the point well as it flashed the banner headline, "Why is Ozark in New York?" Ozark, a regional carrier in the midwest, had been given routes to major eastern cities. The idea was that it could use excess long-haul profits to offset losses from serving numerous small towns in Arkansas, Illinois and Missouri. Of course, Ozark was too small a presence in the eastern U.S. to provide efficient service. The answer to Ozark's question in the advertisement was that it served New York only because the CAB wanted it there, not because their eastern routes made economic sense.

The second source of inefficiency arose from the incentive effects of the pricing system. The price structure provided an incentive to degrade service quality on the short-haul routes but to provide excessive service quality on long-haul routes. For the most part, the short-haul routes lost money and were monopolies, so firms sought to provide the minimum service that would still keep them in good standing at the CAB. The longer routes were generally profitable and served by more than one carrier. Not permitted to compete in pricing, carriers could only compete in service: a large number of convenient flights, better in-flight service, a more comfortable seating arrangement. Moreover, airlines could compete in type of aircraft, always providing the fastest, most up-to-date equip-

ment. Thus, regulation induced airlines to adopt jet aircraft more rapidly than was warranted by cost and demand considerations.

The effect of these incentives was to erode the cross-subsidy of the price structure. The excess profits on long-distance routes served by multiple carriers were persistently eroded by service competition, raising average costs and leading to another round of overall price increases. Passengers in long-distance routes may have flown largely empty aircraft with gourmet food and oversized seats, but they also saw air fares steadily climb as the CAB attempted to regenerate subsidies.

Regulation's Final Response

In the last few years of active CAB regulation, the agency made several attempts to use more detailed regulation to control the inefficiencies arising in the system. The most important were to change the system for establishing overall revenue requirements, and to attempt to control some aspects of service. On the revenue side, the CAB established revenue requirements based on an assumption about capacity utilization. The target revenue requirement was calculated by assuming that carriers would achieve, a target rate of capacity utilization that was higher than the industry was then experiencing. Presumably, this would curtail service competition. On the service side, the CAB made several futile attempts to curtail competition. One was formally to adopt rules regarding in-flight services. For example, the agency adopted a rule carefully defining a sandwich in response to the outbreak of a "sandwich war" in which competitors served ever more elaborate sandwiches to entice customers. Another CAB action encouraged agreements among carriers on flight frequencies, paralleling the controls used in international markets.

All of these attempts failed, largely because airlines could figure out ways to compete at a more rapid pace than the CAB could write regulations. Moreover, defining "optimal" capacity utilization rates, let alone sandwich composition, was a regulatory morass, placing the CAB in the unenviable position of becoming the chief operating officer of every airline.

Meanwhile, some quite disturbing facts were becoming known. Several large states deregulated their intrastate airlines, which were beyond the jurisdiction of the CAB if they did not fly interstate routes. Prices were substantially lower in these markets than in comparable interstate markets. Even unregulated short-haul routes often had lower prices than the supposedly subsidized interstate short-haul routes. In addition, an airline that existed only on paper – World Airways – applied for authority to fly 3000 mile coast-to-coast routes for $100 per ticket when the going regulated price was over $400, and took out newspaper advertisements to announce its intentions. The CAB responded by doing nothing, not even scheduling a hearing.

The Sources of Reform

Advocates of the "ideas" account of regulatory policy change point to two important prereform events. The first is the emergence of an economics research literature that was highly critical of airline regulation. Throughout the 1960s and into the 1970s, a stream of studies found that airline regulation protected carriers from competition, reduced efficiency, and pushed up ticket prices. The second is grousing inside the staff of the CAB. Economic regulation requires a staff of economists, and by the late 1960s the CAB economists had largely brought the profession's view of airline regulation inside the agency. The economics staff continually bombarded the Board with analyses of the perversities of their policies, and leaked damaging information to their colleagues outside the agency.

The shortcoming of the "ideas" account is simply that the timing is wrong. Airline deregulation began in 1977. Before then, "reforms" had largely been for the purpose of trying to preserve the essence of the old system, or to make changes that were in the interests of the carriers, such as allowing some fare discounting so that carriers could engage in more effective price discrimination. Yet the economics profession in and out of the agency agreed before the last round of regulatory policies was adopted that the desirable change was greater competition. In essence, the agency was doing its best to preserve as much of the old structure as it could after the "idea" of substantial deregulation was fully accepted by economists. At best, the economics research contributed to the momentum for reform that developed later.

The "political entrepreneur" account does a better job of explaining the origins of reform. In this case, the political entrepreneur was Senator Edward Kennedy of Massachusetts. Recall that a political entrepreneur does two things: finds a way to organize a new political interest in favor of a policy change, then orchestrates a way to overcome the status quo bias in the political system so as to take advantage of the fact that all policies are inherently unstable. By doing both, the entrepreneur captures the political support of the new interest.

In 1973, Kennedy became interested in running for president. One problem he faced was an image of being too liberal, causing him to receive little support from parts of the business community that might be willing to support a Democratic presidential candidate. To overcome this image, Kennedy decided to respond to complaints from some businesses that regulation was too burdensome and ought to be reduced. In searching for an industry in which to advocate regulatory reform, he settled on airlines, in part because advisers from Harvard University informed him of the recent sorry performance of the industry and the CAB. Consequently, Kennedy decided to launch an attack on airline regulation.

A major problem Kennedy faced was that he was not a member of the Senate committee responsible for airline regulation. Hence, if he introduced legislation focusing on the airlines or even attempted to investigate

them for the purpose of legislating, jurisdiction would be assigned to a committee that supported the existing structure and had gatekeeper rights to any reform proposal. Kennedy was, however, chairman of the committee in congress that was responsible for the general area of administrative law – the procedures governing agencies in making decisions and courts in reviewing these decisions. Thus, he announced his intention to hold hearings on the administrative law of regulatory policy, with the purpose of reforming some aspects of administrative processes. He further announced that the first agency to be examined would be the CAB, and that the first witnesses would be from the agency, who would inform his committee about administrative problems that they were encountering. Of course, the remainder of the hearings was in fact a detailed review of the CAB's policies. And, because Kennedy was a nationally visible politician running for president, the hearings received widespread media attention.

Kennedy's principal rival for the Democratic presidential nomination, and the eventual winner, was Jimmy Carter. Carter sought to capture the same business constituency Kennedy was attempting to reach by advocating airline deregulation. In fact, a main theme of the Carter campaign was the necessity of having an "outsider" as president to reform a bloated Washington bureaucracy. Carter immediately signed on to Kennedy's advocacy of airline deregulation, partly to deny the issue to his foe, but partly because the CAB was a good example of Carter's more general point about Washington.

Carter made good his campaign rhetoric by his appointments to the CAB. Immediately after taking office, Carter succeeded in appointing to the agency economists who favored deregulation, and they immediately proceeded to pursue a gradual path of liberalizing reforms. Thus, the third stage in reform is an illustration of "agency noncompliance" as a means of reform. The CAB proceeded to change radically the nature of airline regulation without the slightest hint from congress (other than Kennedy's hearings which produced no legislation, just publicity) that it wanted reform. The CAB immediately introduced downward rate flexibility, and changed the rules for route awards. Henceforth, the burden of proof would be shifted: instead of the potential entrant having to prove that the established carriers were providing inadequate service, the incumbents had to prove that entry would harm service. The latter, while essentially impossible to prove, is actually consistent with the "ruinous competition" argument that gave rise to regulation. If competition would be ruinous, let those who fear it make the case, rather than requiring that those who want to compete must prove that competition is not ruinous. Finally, the agency began to allow new carriers into the industry.

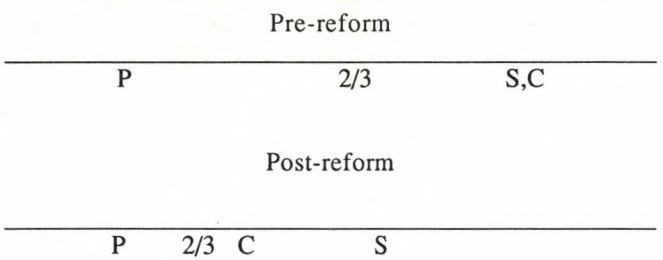
By these actions, the CAB adopted a new status quo, and the status quo bias began to work in its favor. The relevant congressional oversight committees were distressed, and immediately began the process of trying to stop the reforms. By the time the process could begin, however, a year had passed. By this time, deregulation had created some winners, and an

important aspect of the interest-group account of policy change became apparent. Before reform, the employees, managers and stockholders of the entrants and incumbents that would benefit from deregulation did not know who they were. If they were in the industry, all they knew was that, on balance, regulation helped the industry, so they supported it. If they were not in the industry before reform, they were not organized because they did not realize that they would be drawn to the industry by post-deregulation events. But once reform took place and the winners were identified, they stood ready as an organized group to oppose retrenchment.

By the time the committees overseeing airline regulation could organize a counterattack, the industry had become split over the issue of reform. Some major carriers had actually prospered under competition because they were more efficient. Some cities had gained new service, and so supported the changes. Aircraft manufacturers, who profited from the incentive in the old system to buy new aircraft before they were economically justified, also profited from the entry of new carriers and the expansion of old ones, primarily because demand for air travel proved to be price-elastic: price reductions substantially increased ticket sales and hence the demand by airlines for aircraft. As a result, manufactures, who were expected to defend regulation, took no public position on the efforts to re-establish the old system. Thus, the retrenchment effort failed.

The preceding events can be illustrated by the theory of structure-induced stability. Figure 3.1 depicts as a single dimension the extent of regulation of airlines, with more regulation associated with positions on the right side of the line. The figure shows the positions of the relevant actors in the prereform and postreform periods, which are divided roughly at 1977: S is the status quo ante (regulation), C is the ideal point of congress (reflecting the mean position of industry), P is the position of the president and the CAB, and 2/3 is the position that is "veto-proof" – the most pro-deregulation position that could not be overturned by congress after a presidential veto of a bill reregulating the industry. The difference between the two lines simply reflects the difference in organized interest representation in congress before and after the initial reforms.

Figure 3.1: Airline Regulation Policy



In the pre-reform period, Carter wanted almost total deregulation, but to adopt point P would invite legislative reversal. If the President vetoed the bill, he would lose the override vote. Hence, the CAB adopted a gradual deregulation policy that could withstand a legislative attack because the president's veto could be sustained. Thus, a policy roughly at the "2/3" point was adopted. By the time legislation was passed in 1980, the new status quo (the position "2/3" in the top half of the figure) had become more regulatory than even congress desired. Hence, congress passed, and the president signed, a bill abolishing the CAB and all but a few aspects of regulation. Congress still fell short of the president's ideal, because it insisted that the transition to deregulation take five more years, that the Department of Transportation (not the Department of Justice) have jurisdiction over airline mergers (assuring a more pro-industry policy), and that the possibility of subsidies of service to small communities be retained to guarantee universal service. But these differences were small compared to the differences between congress and the president only two years earlier.

Consequences of Reform

On nearly all measures, the performance of the airline industry has been substantially better after deregulation than before, but controversial subsequent policy actions have eroded some of the benefits of deregulation. Nonetheless, from the perspective of airline customers as a whole, the overall effect of deregulation is definitely beneficial.

The most useful basis of comparison for airline prices is the divergence of actual prices from the prices that would have been charged had the industry still used the old CAB price formulas. According to this comparison, the average ticket price was about 25 percent lower ten years after deregulation than it would have been had airlines still used the old formula. In fact, the price effects of deregulation are probably greater than

this, for service competition probably would have continued to cause upward readjustments in the formula.

In addition, entry by new airlines and expansion of established ones have caused a substantial increase in the number of flights that are flown. The increase in flights is true for all communities, large and small, and averages about sixty percent above the number of flights just before deregulation. Moreover, the increase in flights has generally been accompanied by an increase in choice of carrier. As late as 1979, on sixty percent of all trips passengers had either no choice of carrier or a choice between only two. By 1988, sixty percent of the trips offered a choice of three or more carriers.

The least expected change in the industry has been the complete rearrangement of the route system when carriers have a free choice of routes. The two most important changes are in commuter service and the design of the route structure of an airline.

First, all major carriers have either entered the commuter airline business or become affiliated with a commuter carrier. Commuter flights are now scheduled to connect to regular, long distance flights at the nearest major airport, and the effect has been a substantial increase in commuter traffic. Many small communities have experienced a profound change in service: a downgrading from regular airline service to commuter service, but a substantial increase in the number of daily flights. The net effect is to make more cities easily accessible in a one-stop connecting route.

Second, regular airline routes have been altered to the "hub-spoke" system. Instead of flying primarily east-west or north-south routes in a cross-hatched route structure, airlines now concentrate operations in a small number of cities. Almost all service begins with a flight from a hub, with the only exceptions being very long distance flights between the nation's largest cities. This structure enables airlines to concentrate aircraft maintenance and crew bases in large centers, facilitating substitution of aircraft when unplanned repairs are needed and of personnel in health emergencies. It also facilitates easy interconnection of flights with commuters or between spoke routes, because the airline has so many flights originating from the hub. As a result, many more trips require one stop and a change of plane at a hub, and many fewer trips are multistop flights on a single aircraft. In the past, passengers had been reluctant to change aircraft because of the danger that flight delays would cause a missed connection; however, hubbing reduces this concern by concentrating so many flights out of the point of connection. The effect of the new route structure is a substantial reduction in airline costs, accompanied by a slight increase in the average duration of a trip because more flights now require connections.

The negative consequences of deregulation flow from the changes in operating practices. A few cities have lost service in that actual flight frequencies have declined, but the vast majority of cities – even small ones – have experienced increased service. The primary negative effect on ser-

vice has been that some cities now have commuter flights, using smaller, less comfortable, and slightly less safe aircraft, rather than regular service using jets. In addition, in a few cases the hub-spoke structure has created local airline monopolies. Hubs tend to be located in the larger cities; however, "large" can mean a city as small as Salt Lake City, Utah, Memphis, Tennessee, or Dayton, Ohio. These cities are simply too small to support more than one hub, and if a carrier decides to form a hub in a relatively small city, its flight frequency is so large that other carriers are unlikely to provide it with much service out of their own hubs. In addition, some larger cities have become monopolized because a single hubbing carrier has been permitted to control nearly all the gates and landing times. Thus, one carrier controls nearly all flights in a few major cities, such as St. Louis, Missouri, Pittsburgh, Pennsylvania, and Detroit, Michigan. In cities dominated by a single hubbing carrier, prices are substantially higher, enough so as to offset the overall reductions from deregulation.

A commonly expressed concern is that airline deregulation has caused a deterioration of safety. In two senses this claim has some basis. Cities that lost regular service to a commuter service have a slight decline in safety because smaller aircraft are less safe. In addition, increased flight frequency has made the airways more congested around large airports, although the effect thusfar has been to increase the perception of danger through close encounters between aircraft. Midair collisions have not increased. Indeed, all measures of flight safety show a continuation of the improving trend that has taken place throughout the history of the industry. Deaths, injuries, and accidents per passenger mile continue to decline after deregulation as rapidly as they did before.

Nonetheless, airline safety may become an issue in the future. Safety, of course, was not deregulated, and no one ever seriously proposed that it should be. The Federal Aviation Administration regulates safety, and has for decades. It also operates the air traffic control system. The problem lies in the fact that the FAA was no larger in 1990 than it was in 1980, even though there was a sixty percent increase in flights. Thus, the frequency of aircraft inspection has declined, and the number of controllers per take-off and landing has fallen. Perhaps the FAA was excessively large in 1980, so that a reduction in the intensity of safety regulation will have no long-run effect. Nonetheless, at some point safety regulatory capacity will have to expand to match the growth of traffic.

Reagan Era Problems

The differences between airline deregulation during the Carter era (which actually persisted well into the 1980s because Carter appointees were holdovers at the CAB) and the Reagan era are instructive, for they point to the effects of the overall ideology of deregulation on the way policy is carried out. The Carter era deregulation was essentially populist

in origin, defended on the grounds of the benefits to consumers. The Reagan era policy was the neoconservative form of deregulation, focusing primarily on the simpler principle of removing government constraints on business. The second term of the Reagan Administration, after policy was moved from the CAB to the Department of Transportation, brought forth several turns in policy that would not have taken place under a populist regime.

The most important was merger policy. The Reagan administration approved essentially every airline merger that was proposed to it. All of the monopolized hubs in larger cities were created by mergers, whereby two carriers that had competed out of the city were allowed to merge without spinning off to competitors enough gates and landing rights to prevent monopolization of the market.

In addition, the Reagan administration failed to come to grips with the problem of airline computer reservation systems. Briefly, to succeed in the market a CRS must list nearly all flights in the nation. Consequently, large, national carriers have a distinct advantage in constructing systems. The major national carriers created CRS in the late 1970s, and then when almost all travel agencies were automated, used their CRS presence to reduce competition in both airlines and CRS. Presently they do so by charging high booking fees to other airlines, accounting for nearly half of the profit margin for the average ticket. The early entrants also created a high entry barrier for other CRS by creating contracts with travel agents that made it very costly for agents to switch – or even to use part time – another CRS. Thus, two large CRS – organized by the two largest airlines, American and United – managed to use these systems to reduce the extent to which airline deregulation produced effective competition.

Several solutions have been proposed: divestiture (with a prohibition against vertical integration of airlines into CRS), a single regulated joint-venture CRS for the entire industry, a prohibition of carrier booking fees, and elimination of contract terms with travel agents that retard CRS competition. Regulation of CRS was retained after the demise of the CAB, and is located in the Department of Transportation, but thusfar the agency has refused to take meaningful action. By contrast, in Canada and Europe CRS rules provide better safeguards against using a CRS to disadvantage competitors or encourage cartelization.

Conclusions

Airline deregulation is not without problems, caused in part by other policy failures, notably merger and CRS policies. But these problems should not be allowed to deflect attention from the fact that deregulation is clearly a success in terms of prices, flight frequencies, and the efficiency of the airline industry.

The politics of airline regulation are interesting as well. They demonstrate interesting cases of political entrepreneurship by Edward Kennedy and Jimmy Carter, and then the rearrangement of interest groups due to the fact of deregulation. They also demonstrate that within the spectrum of reforms that would be regarded as deregulatory, a clear difference emerges between the neoconservative and populist version. Unfortunately, some abuses arising from the former regime have given rise to new calls for reregulation.

If there is a lesson for other nations in the U.S. experience, it is in the restructuring of the U.S. industry after deregulation. In all but a few very large countries, international travel is the more important part of the industry. International route structures are determined through governmental negotiations, with largely irrelevant national boundaries as a major constraint. Moreover, international flight frequencies are also established through negotiation. So far, liberalization has been minimal in that a carrier's national identity continues to have a substantial effect on where and how often it can fly. If liberalization occurs in the EEC, European airlines, which already are substantially less efficient than U.S. carriers, can be expected to redesign their route structures even more dramatically than their U.S. counterparts did.

Moreover, if U.S. experience is a guide, the number of major European airlines providing continent-wide service will be only a handful – substantially fewer than one "national champion" carrier per country. Of course, U.S. experience may not be relevant because of the unjustifiable merger policy pursued by the Reagan administration; however, even without these mergers, the U.S. would still have fewer then twelve major nationwide carriers in a much larger market with a much greater demand for air travel.

Perhaps the unsettling prospects for some national carriers that are protected by the present regulatory structure explain why the EEC has been reluctant to push too hard for airline liberalization, despite the obvious fact that the current system operates under a set of principles that are completely at variance with the precepts of European economic integration. Perhaps another contributing factor is the shared desire among EEC members to protect rail passenger service from competition with airlines. In any case, the consequences are detrimental in two ways: airline service is more expensive and less convenient than it would be if carriers were free too compete in routes, flights and prices.

Lecture #4 – Air Pollution

The massive expansion of environmental, health and safety regulation in the 1970s stands in sharp contrast to the reductions in economic regulation. In the 1970s, the U.S. enacted expansive statutes regulating air pollution, water pollution, solid waste disposal, toxic chemicals, consumer product safety, occupational safety, and pesticides. By 1980, the U.S. was spending an estimated three percent of GDP on complying with these regulations.

For the most part, the expansion of protective regulation predated the period of economic liberalization, but not completely. Between 1978 and 1990, no major new environmental legislation was enacted in the U.S., and by 1978 economic deregulation was seriously underway in only the airline and telecommunications industries. Nevertheless, during the Reagan years, the administration tried to roll back protective regulatory policy, but neither the courts nor congress would go along. When liberalization was sweeping the economic regulatory domain, it had very little effect in other regulatory areas.

The focus of this essay is on the regulation of air pollution because only this area of protective regulation has been significantly touched by regulatory reform. In 1990, substantial new amendments to the Clean Air Act were passed which simultaneously increased the scope and stringency of air pollution controls yet also significantly changed the basic strategy of controlling emissions. Part of the act deals with new emissions standards for automobiles, new controls on airborne toxics, and stringent controls on power plants to curtail acid rain. But the act also calls for a new emphasis on using economic incentive methods to reduce emissions. Henceforth, emissions trading will play an important role in reducing emissions that cause acid rain, and local pollution control authorities have been given legal flexibility to adopt emissions trading and pollution taxes to achieve air quality objectives.

Traditional Regulatory Methods

Protective regulation (environmental, health and safety regulation) by the federal government began with the Pure Food and Drug Act of 1906, which sought to control adulterated and dangerous food and drugs. But this legislation was weak, and protective regulatory policies were largely

left to the states until the 1960s. The primary exception was nuclear power, but initially the regulatory authority – the Atomic Energy Commission – was also the sponsor of the technology, and safety was not regarded as a major issue.

Although the most important protective regulatory statutes were enacted in the 1970s, the trend toward greater federal control began in 1962, with the passage of new legislation for drug regulation. The new statute created a burden of proof on producers of new drugs to demonstrate that their products were safe and effective in carefully controlled clinical trials. This statute set the tone for nearly two decades of more statutes to follow. Regulation used the "command and control" approach to protecting citizens against hazards: set a technical standard that a hazard must satisfy in order to be permitted.

The 1960s arguments for federalizing protective regulation were very similar to the arguments now given to justify expanding the powers of the EEC to issue similar regulations. Separate regulatory standards in each state increase the costs of firms that try to sell products in a multijurisdictional area, and can be used by a state to disadvantage products manufactured elsewhere. Moreover, federalization captures economies of scale in information about the causes of hazards and the technical possibilities for ameliorating them. In addition, federalization prevents localities from engaging in the Faustian bargain of competing for industries by offering lax regulation. Finally, some hazards are inherently multijurisdictional in that activities in one state impose hazards in others.

Initially, the U.S. government entered the field of protective regulation in only a modest way, serving as a clearing house of technical information and as a regulator of last resort when state and local governments could not effectively cope with regional problems. Gradually, however, federal authority become far more expansive. By 1972, in all of the important areas of protective regulation, federal agencies set policy goals, wrote standards or reviewed (with the right to assert authority) state and local standards, and set the rules for products, workplaces, and emissions.

The Standards Process

The standard-setting approach to regulation had two defining features. First, it dealt separately with each source of hazard. Second, it specified exactly what the owner of a potential hazard had to do in order to comply with regulatory policy. In environmental regulation, the standard rarely specified a performance objective. Instead, it specified a production or emissions control method. The basic idea was to write standards for all important sources of pollution by surveying extant technical possibilities and telling a polluter to use the one that the regulator thought was appropriate.

The environmental regulator had two general guidelines for writing standards. The first was a target level of pollution, in most cases defined

as the level that posed no threat to public health. The second was an economic feasibility constraint. Existing facilities could not have bankrupting regulatory requirements imposed upon them unless they were a direct threat to public health. In the vast majority of cases, environmental pollution arises from the combined effects of a large number of separate facilities. Hence, no one facility is a threat to public health even if total pollution does constitute such a threat. The nonbankrupting constraint arose in the U.S. from constitutional protections against government seizure of property without just compensation, but even without the constitutional requirement, government officials are unlikely to want to cause substantial economic disruption. Moreover, the economic feasibility constraint minimized the extent to which protective regulation would reallocate economic activity among localities, and so minimized political resistance to ambitious environmental policy targets. Thus, the task of the regulator was to impose technical standards on sources that simultaneously would not impose undue economic harm and would protect public health.

Two paradigmatic views about protectionist regulatory policy gave rise to the adoption of the standard-setting approach. One was "technological optimism" – the view that most human problems will eventually yield to a relatively painless technological solution, so that the best way to solve a social problem is to develop new technology to deal with it. By setting in place a standards system, regulators would be ready to force use of the technical fix when it arrived. The other was the "bad actor" theory of hazards arising from economic activity. Much of the problem was thought to be caused by a few socially irresponsible people who were not adopting cheap, best-practice methods to curtail the hazards they created. Cleaning up pollution was in large measure a simple matter of identifying the bad actors and imposing standards on them.

In air pollution control, the traditional standards approach led to six significant types of policy actions. (1) The federal regulator, the Environmental Protection Agency, set Ambient Air Quality Standards (AAQS), using scientific studies about the relationship between pollution and health. If achieved, the AAQS would eliminate all adverse health effects from air pollution. (2) The EPA also set emissions standards for mobile sources, defined as a maximum quantity of emissions per mile. Only California, with severe smog problems from auto emissions, was permitted to set more rigorous standards, and no state was permitted less rigorous regulation. (3) EPA also set "new source performance standards" (NSPS) for stationary sources, which applied only to newly constructed or substantially modified production facilities. Because these facilities had standards imposed before they were constructed, these standards were not constrained by the nonbankruptcy constraint, and so were generally far more rigorous than standards for existing stationary sources. (4) An air quality standard of "prevention of significant deterioration" (PSD) was adopted for areas of the country that satisfied AAQS,

These areas could not allow the quality of their air to deteriorate, and were required to impose NSPS on new facilities. This policy prevented clean-air areas from stealing industry from dirty-air areas by writing lax standards for new facilities while still satisfying air quality targets. (5) EPA delegated to the states the responsibility for developing "state implementation plans" (SIP) consisting of standards for existing stationary sources that would meet AAQS. States also designated regional air quality control authorities to write and enforce these standards. EPA then reviewed the SIP, and could, if it found the plan insufficient, assert authority and impose a new plan on the state. (6) All standards were periodically reviewed to bring them up to date with technology and to make corrections based upon the air quality effects of the last round of standards. This procedure required that every few years regulators review the production and control technology of every production facility in the U.S. that was located in a region that did not attain AAQS.

Within a few years the problems with the standards approach were apparent. Initial expectations about the speed with which the U.S. would achieve AAQS proved wildly optimistic, and most major cities made only very slow progress. The NSPS and PSD rules were surely part of the problem, because they created an enormous disincentive to replace old polluting facilities with new, less polluting ones, perhaps in areas where air pollution was not a problem. Among existing stationary sources, substantial disparities emerged in the cost of compliance. Facilities that were located in polluted areas and that faced significant competition from facilities in unpolluted areas were given very lax standards, for otherwise the facilities would close. Facilities that faced no effective competition and a relatively price-insensitive demand were given rigorous standards. On a cost-per-ton basis, some facilities ended up paying 100 times as much for abatement as other facilities located nearby. Policy was certainly succeeding in assuring that air pollution control did not cause a reallocation of economic activity; however, it was also imposing substantial costs while making little progress toward achieving air quality goals.

Incentive Regulation

The concept of incentive regulation arose as a means to escape ineffective but costly standards regulation. Attacking both parts of the problem – costs and effectiveness – was important, for only then could reform capture the two main political constituencies concerned about environmental policy, business and environmentalists. The cost of environmental regulation ought to be of concern to environmentalists for three main reasons. First, in large urban areas the attainment of reasonably unpolluted air was simply not possible in any foreseeable time unless a regulatory approach could be found that was less costly than the standards system. Second, area-wide economic disruptions from environmental regulation are

minimized when regulation is efficient, in part because efficient regulation encourages the restructuring of the local economy in favor of less-polluting industries. The effect is less pollution per production worker, and less need for forcing people to relocate or become unemployed to achieve air pollution policy goals. Third, efficient regulation reduces the political resistance to environmental policy. If it is cheaper to achieve air quality goals, some sources of pollution will find it cheaper to comply than to fight the system politically, and all sources will have a smaller net willingness to pay for anti-environmental policies because they will have less to gain from their repeal.

Incentive regulation methods are market-based systems of environmental control that produce emissions control targets at least cost. Incentive methods reduce compliance costs in two ways. First, they contribute to *static efficiency* in that they lead to a situation in which the last unit of abated emissions imposes the same incremental cost at all facilities. Hence, no further cost reductions are possible by reallocating emissions among sources. Second, they contribute to *dynamic efficiency* by encouraging technological progress in emissions control technology. If under a standards system the binding limit on emissions control requirements is economic feasibility, technological progress can only harm a polluting facility. A new cost-reducing emissions control system will lead regulators simply to tighten the emissions standard until, again, the economic feasibility constraint is reached. As a result, all of the benefits of cost reduction are captured in reduced emissions. If a source is uncontrolled because current technology is economically infeasible, technological progress can lead to a lower-cost control that is economically feasible, and so imposes a new cost on the source. Under incentive regulation, a more efficient control system benefits the polluting firm.

Both of these effects arise from a key property of incentive regulation: the costs faced by a source of pollution include both abatement costs and a cost of emissions. Because companies seek to minimize total costs, placing a cost on emissions balances the incentives between emitting more pollution or installing better emissions control systems.

Emissions Taxes

One method of incentive regulation is emissions taxes, whereby firms are charged for each unit of emissions. A company will abate emissions up to the point at which the cost of further abatement exceeds the tax savings that more abatement would permit. Overall emissions targets are achieved by setting the tax high enough so that, collectively, the polluting facilities in an area minimize the sum of taxes plus abatement costs by abating enough so that overall targets are met. Because all companies face the same tax, they all abate until further abatement costs equal the tax, and so abatement is efficient: abatement responsibilities cannot be reallocated among them in a way that reduces total costs.

The primary problem with emissions taxes is that tax revenues are collected, so that the cost of the policy to a company is abatement costs plus taxes, not just abatement costs. Even if the government used emissions taxes to reduce other business taxes, many firms would end up worse off with an emissions tax system than with an inefficient standards system. Hence, no government has made extensive use of emissions taxes. In the instances in which they have been used, the purpose of the tax has been to pay for the regulatory system or to subsidize pollution abatement, rather than to create effective abatement incentives. Thus, the emissions tax approach, although attractive from the standpoint of efficiency and effectiveness, is probably politically infeasible.

Emissions Trading

The other incentive approach is emissions trading, in which regulators establish an overall emissions ceiling for a region, adopt a procedure for allocating emissions permits for this amount of total emissions among sources of pollution, and allow sources to buy and sell emissions permits. Emissions trading has exactly the same economic incentive effects as a tax: a company will abate pollution beyond its permit holdings if it can sell the permits for a greater amount than the abatement costs. Likewise, a firm facing especially high abatement costs will buy permits from a company facing lower costs.

The numerical example in Table 4.1 illustrates the point, showing the costs of abating pollution from two sources. The

Figure 4.1: Abatement Costs and Emissions Trades

COMPANY A			COMPANY B		
Amount Abated	Total Cost	Added Cost	Amount Abated	Total Cost	Added Cost
1	$ 100	$ 100	1	$ 200	$ 200
2	300	200	2	600	400
3	600	300	3	1000	400
4	1000	400	4	1500	500
5	1400	400	5	2200	700
6	1900	500	6	3000	800
7	2500	600	7	4000	1000

"Amount Abated" refers to the physical quantity of emissions per time period (usually, per day). "Total Cost" refers to the amount spent on abating the quantity on the corresponding line. "Added cost" is the increment to total cost due to the last unit of abatement. Thus, for Company A,

moving from 2 to 3 units of abatement adds $300 to total abatement costs. Suppose that the total emissions ceiling is ten units, and that five units each have been assigned to each company. Thus, if each firm has a total uncontrolled emissions of eight, then each has been given three emissions permits. Each could abate by five units, but this would be inefficient. Instead, Company B will offer to buy a unit of emissions from Company A. The former can save the added cost of $700 from increasing abatement from 4 to 5 units, whereas the latter will need to spend only $500 to increase abatement from 5 to 6. Thus, at any price of an emissions permit between $500 and $700, both companies are better off from having traded. The rest of society is also better off as well, because, at no sacrifice of environmental goals, the two companies have reduced their costs, and hence the prices for their products will be lower, and more resources will be available for producing other things.

In reality, emissions trading is more complicated than the example. To begin, costs differences among polluters are larger than the differences in the table, so that there is more to be gained from trading than this example illustrates. In addition, an air quality region contains far more than two sources of pollution. As a result, firms will probably have to devote more effort to finding trading partners than in the example, but there will also be less uncertainty about the trading price. With a large number of companies, trading will quickly establish a stable permit price, rather than arise from negotiations as in the example.

The regulatory requirements for an emissions trading system to be effective at minimizing costs and controlling pollution are straightforward, but not necessarily simple. First, regulators must establish an emissions baseline for each source. The baseline is the initial allocation of permits to a source, which it can then simply treat as a standard by reducing emissions to be consistent with the permit. Second, the regulator needs to monitor all trades in order to keep track of emissions by source in order to enforce the overall emissions ceiling for the region. Third, a mechanism must be in place for firms to demonstrate that they are actually emitting no more than the quantity of permits that they hold. Fourth, regulators still need to monitor air quality throughout the region to be sure that "hot spots" do not emerge – locations where pollution accumulation is especially heavy. In some cases, to achieve air quality goals, some sources may be forced to engage in additional abatement; however, these controls can be financed in part by selling permits to others whose emissions do not contribute to the hot spot. Fifth, regulators may want to facilitate the development of a trading system by organizing the emissions market. By establishing a regular time and place for trading, and publicizing the results, efficient operation of the market is facilitated.

Enforcement

The most challenging problem for all methods of environmental regulation is designing an effective enforcement system. Basically, the job of enforcement consists of checking facilities to be sure that their emissions match their allowances, regardless of how the allowances are determined (standards, trades, taxes). In the standards system, because standards are usually based on the use of a technology rather than actual emissions, enforcement is in some sense easier because all an inspector has to do is check whether the technology is installed and working. But this enforcement mechanism is easier only because it is not concerned with actual performance. If regulators are to succeed in curtailing total emissions under an overall pollution ceiling, they must base standards on performance and measure actual emissions. At that point, the essence of the enforcement problem is the same for all methods: determine how much is actually being emitted.

Emissions monitoring can proceed in three ways. For large production facilities, "continuous emissions monitoring" is not only technically and economically feasible, it is often already required. Trades of emissions permits create no problems here, because the permits of a company can easily be checked against emissions records. "Random intermittent monitoring" can be used when continuous monitoring is infeasible. With this approach, inspectors arrive unannounced to a facility with mobile measuring systems and monitor emissions. The results are then checked against permit holdings. The third approach is "emissions modelling," in which estimates are made of the emissions that arise when a particular technology is used and a production facility is operating. Companies keep records of hours of operations, and inspectors make random checks to be sure that controls are working and that production records are accurate.

The last system is the least accurate, but it is the one that is most easily developed from a standards system. Generally, it requires a regulatory process to ascertain the emissions performance of a control method. Hence, it requires prior approval by a regulator that a particular control method actually meets emissions ceilings. This approach can cause emissions trading to require more enforcement for sources that are monitored by emissions modelling methods. Every time a company makes a trade and changes its emissions control methods, it must go through a standards review to estimate the emissions arising from the new method. In the present system, standards reviews occur approximately every five years; in principle, a source might make trades more frequently than this, and in any case coordinating standards review and trades could make the system cumbersome. Nevertheless, the result is not a loss of enforcement or control, for trades could not be completed until standards reviews were finished. Moreover, because the process is cumbersome, companies and regulators would have a strong incentive to move quickly to one of the

other two methods, both of which are more accurate and less bureaucratic.

Emissions Trading Experience in the U. S.

Serious attempts to control air pollution on a nationwide scale were not attempted until after the passage of the Clean Air Act of 1970. Before then, only a handful of large metropolitan areas had attempted a comprehensive plan to reduce air pollution, with the most notable examples being Pittsburgh's program to reduce pollution from steel production and California's actions to reduce photochemical smog in Los Angeles. In both cities, the defects of the standards approach had not become apparent. Pittsburgh succeeded in substantially improving air quality by writing standards for a single industry – steel. By 1970, Los Angeles had attacked smog by imposing controls on a small number of large sources, primarily electric generation facilities which were required to use natural gas as a boiler fuel during the most smoggy parts of the year, and by limiting automobile emissions. By 1970, regulators knew that the first set of standards would not solve the problem, but they had not yet embarked on a systematic attempt to control all significant sources in the region. Thus, in 1970, the federal government initiated the standards approach for nationwide air pollution control.

Controlled Trading

Before a decade had passed, the standards approach was rather obviously not working as well as had been expected. In the major cities where pollution was worst, the best economically feasible standards for old sources (plus NSPS and strict auto emissions controls) had only marginally improved air quality. Soon after taking office, the Carter administration, which had emphasized the inefficiency of regulatory policy in the campaign, instituted a program of "controlled trading options." All three words of the program were important: trading because the program did allow emissions trading among sources; controlled because the method of trading was cumbersome and ruled out many types of trades; and options because local pollution control authorities were not required to adopt the system. The Carter program represented an attempt to make the adoption of incentive systems purely a question of implementation method, paired with continued support for environmental improvement. Thus, the program was intended to appeal to both environmentalists who were disappointed with progress in improving environmental quality and businesses who sought a cheaper, more flexible regulatory system.

The details of the program clearly embodied both concerns, but in so doing restricted trading possibilities so greatly that the effect of the program was small. The idea of the Carter trading program was that a group

of sources that already had standards written for them could collectively propose new standards for the entire group, but subject to four major conditions.

First, the new standards had to reduce emissions, not just reallocate them. Depending on the case and locality, these reductions varied from ten to fifty percent. Thus, there was no reason to propose trades unless the cost saving was so large that the sources could increase abatement efforts substantially and still reduce total compliance costs.

Second, the new pattern of emissions had to go through the traditional standard-setting process, usually leading not to a performance standard but a technology control standard. Because almost all standards were not expressed as a ceiling on total daily emissions, but instead as a particular method of control, the trading partners had to estimate the emissions from the old standards, estimate the emissions from the new methods of control, prove that the latter was substantially lower than the former, and convince the regulators that their evidence about performance of the two sets of standards was correct. In addition, they had to prove that the new set of standards was as easily enforced as the old. Again, rigid adherence to aspects of the standards process that were a weakness (an orientation toward control methods rather than performance) reduced the attractiveness of trading.

Third, trades were not allowed that enabled firms to avoid certain national standards rules. For example, a new facility could not trade out from under NSPS by abating elsewhere rather than by adopting the rigorous standards for new facilities. New facilities that adopted NSPS were required to reduce pollution elsewhere ("offsets") in the amount of emissions arising from the new facility after NSPS standards were met. These offset trades could take credit only for controls beyond something called "best available control technology," which was the method presently on the market that reduced emissions for the offset partner by the greatest amount. To control beyond this, the new facility usually had to redesign the production facility of the offset partner.

Fourth, trades were not permitted across pollutants. For example, photochemical smog and acid rain both have numerous components, but national air quality standards are set separately for several components. Thus, for example, a source could not offset emissions of oxides of nitrogen by reducing emissions of sulfur oxides.

The reasons for the severe restrictions on trading systems were essentially political. Environmental groups were extremely skeptical of economics as a useful tool in environmental policy making. Economic benefit-cost analysis often reached negative conclusions about policy objectives that environmental groups favored. In fact, these studies had to cope with two serious sources of bias. On the cost side, polluters usually had the best estimates of abatement costs, and would attempt to convince regulators and their policy analysts that costs were likely to be higher than perhaps they might be in reality. On the benefits side, estimation methods

were primitive to nonexistent for such things as aesthetic effects or minor, non-debilitating health effects, and were normatively dubious for morbidity and mortality. Thus, the early experience of environmentalists was to regard economics as generally hostile to their policy objectives.

A second problem was that some environmentalists reacted to the failures of the standards system by turning away even further from the concept of efficiency. The combined technological optimism and bad actor views of 1970 had given way to a pessimistic belief that environmental pollution was an inherent feature of a capitalistic, mass consumption society. Hence, if emissions trading was attractive to large corporations in the manufacturing and energy sectors, then it must be detrimental to the long-term policy objectives of environmentalists.

A more common belief among environmentalists was that the failures of environmental regulation during the 1970-77 period were due primarily to the fact that Republican Richard Nixon, an ally of business, had overseen the first few years of operation of the Environmental Protection Agency. The failure of policy, then, was due to the fact that it was not conscientiously carried out. A Democratic president, giving greater weight to environmental concerns, would do better.

For all of these reasons, environmentalists opposed any significant use of incentive mechanisms in environmental policy. Because Democrats controlled congress and the presidency, environmentalists were guaranteed influence in how the new administration reformed the system it inherited in 1977. The emphasis of the new program was going to be tougher standards and tougher enforcement, with some experimenting at the margins with emissions trading.

In the ensuing years, despite the restrictions, several thousand trades took place, saving pollution sources several billions dollars. Even environmentalists agreed that trading had not undermined environmental policy goals, but instead in some cases had facilitated them. Trades not only led to small (overall) reductions in emissions, but they also gave sources a reason to explore new abatement methods so that a trade could be consummated. These new methods, then, could be required of other sources when standards were revised for old facilities. Hence, by the mid 1980s, environmentalist opposition to trading was not so intense. Moreover, the Republicans again controlled the presidency and, in addition, the Senate, so that lower-cost methods for achieving environmental policy objectives were a necessary component of any proposal for additional progress in reducing emissions.

Lead Trading

The next major environmental initiative was to remove lead from motor vehicle fuel. The EPA proposed an ambitious plan to all but eliminate lead from gasoline, but a key part of the program was a trading program whereby refineries could trade their permits to use lead during the phase-

down. The system set lead usage targets for each refinery which dropped precipitously over a few years. Each refinery could reduce the amount of lead in leaded gasoline, and the amount of leaded gasoline produced, by the target amount each year, but this quickly became inefficient. Producing small amounts of leaded fuel at a large number of facilities was economically irrational. Letting refineries trade leaded-fuel production rights concentrated production of leaded fuel in a dwindling number of facilities.

The leaded fuel program was clearly the most successful air pollution program that the EPA had ever launched. Within a few years virtually all of the nation was not only in compliance with the AAQS for lead, but was well below the maximum concentration that was permitted. Moreover, the policy objective had been accomplished almost painlessly. Trading in lead permits had been brisk, and the production and distribution of refinery products had not suffered a loss of efficiency. Of course, the problem of lead pollution was not an especially difficult one to solve: the pollutant is easy to control and to measure, and reasonably good substitutes were readily available. Nonetheless, the program did work, and it was clearly facilitated by the trading program. Trading not only reduced the transition cost to unleaded fuels, it reduced the opposition of part of the industry, small refiners. The small refiners could not efficiently reduce production of leaded fuels. They required either essentially no change or zero use of lead. Because immediate cessation of production of leaded fuels was impractical and unnecessary, a standards approach would have been hopelessly bogged down in deciding which refineries could produce the small amount of fuel that was still needed. Trading solved the problem. A small refinery could profit from its share of production of leaded gasoline by selling its permits rather than producing itself.

CFC Trading

The next event in using emissions trading was in connection with the reduction of chlorofluorocarbons, the ozone-depleting chemicals. The Montreal Protocols, an international treaty dealing with ozone depletion, requires each signatory (including the U.S. and Europe) to reduce both production and usage of CFCs by fifty percent over a decade. Fresh from the success with lead, in 1989 the U.S. instituted a production trading program to achieve its CFC production targets.

The CFC program is more complicated than the lead program, for it covers several different chemicals. These chemicals differ in reactivity (the extent to which the chemical will destroy ozone), so that the benefits of curtailing them differ. Thus, the trading program defines a permit in reactivity units, not physical quantities. Trading across chemical entities is permitted, but the quantities of the chemicals must be inversely proportional to their reactivities. As with lead, producers of chlorofluorocarbons face annual reductions in production allowances, but rather than

each gradually winding down production, the industry can capture scale economies by trading production rights.

The 1990 Clean Air Act

By far the most ambitious step toward the use of economic incentives in environmental regulation was the adoption of new air pollution legislation in 1990. The new act sets up a detailed program for trading sulfur emissions in order to control acid rain, and it allows local air pollution control authorities to use either trading or emissions taxes to achieve air quality objectives. The actual effect of the latter portion of the act is still somewhat in doubt, for, as is always the case in the U.S., the courts eventually will decide how radical a departure from the status quo the new legislation permits. The dominant legal opinion at this writing is that economic incentive methods can completely replace the entire standards system as long as the regulators can make the case that they are not losing enforcement capability. The only exception is that new facilities still probably will be required to satisfy NSPS, although even this view is disputed by some, and will no doubt be tested in court.

Acid Rain

The acid rain program is actually quite similar to the lead and CFC programs in most respects. Oxides of sulfur come in only two forms, sulfate and sulfur dioxide, and the extent to which they acidify rain is well known – just as is the reactivity of different CFC chemicals. Emissions can be easily expressed in "ton-equivalents" to sulfur dioxide for trading purposes. Moreover, nearly all sulfur is introduced into the environment by burning hydrocarbon fuels that contain sulfur: fuel oil, gasoline, or coal. Ascertaining the sulfur content of fuel is easy, and so measuring how much is retained by emissions control systems (and how much, therefore, is emitted) is a relatively simple task.

The new acid rain program makes the important leap of defining the obligations of polluters in terms of emissions quantities, not fuel consumed, the technology used to desulfurize fuel, or the method of trapping sulfur emissions. This step is a major advance by itself, for in the past standards had specified control methods rather than an emissions objective. The program simply specifies that the major sources of sulfur emissions must reduce total emissions by ten million tons per year in ten years.

The 1990 legislation provides a provisional initial allocation of emissions rights to over 200 facilities in the U.S. EPA is left the task of refining the initialization system, but in the end major sulfur sources will begin the emissions reduction period with a clear target for abatement over a ten year period. These sources can then trade permits so that the pattern of abatement among sources will be flexible, just as with lead and CFC.

A major innovation, enacted in the law, is the development of a new market mechanism for facilitating trading. Because many facilities produce sulfur emissions, trading partners might experience some difficulty in finding each other. Moreover, if trades are negotiated privately, the price may not be made public. Open pricing helps other potential traders calculate whether they, too, should buy or sell permits. To assure a regular, publicly visible market, the act sets up a "zero revenue auction" for 2.8 percent of all sulfur emissions. All sources must make available for sale 2.8 percent of their permits. Each source then submits a bid, stating how many permits it wants and how much it is willing to pay for them. Anyone may submit a bid – a new facility, or even an environmental organization that wants to reduce emissions by more than the target amount and intends not to use the permits it buys. The EPA then calculates the price at which the quantity demanded equals the amount available for auction. Successful bidders buy permits at this price, and the facilities from which the permits were taken receive the same price for the permits that they lost. Thus, the government does not keep the revenue from the auction, instead returning it to the facilities holding the permits. Firms that lose permits do so because they bid less than the ultimate market price for the permits that they held. Presumably they bid less because their abatement costs were lower than the value of their permits, for a firm can always keep all of its permits at zero net cost by bidding an infinite price for them. The firm would then both pay and receive infinite amounts, thereby breaking even but retaining its permits. By bidding a lower price equal to abatement costs, if a firm sells its permits the revenues finance the seller's additional abatement, with a profit left over.

The new acid rain trading program is not yet implemented, for the EPA must still write the details of the process. Consequently, the program can not yet be evaluated. Nevertheless, the program is an important next step for the U.S. experiment in using incentive regulation. If it works well, more widespread use of trading – with the auction method – is sure to follow.

Los Angeles Smog

Experiments by local pollution control authorities to take advantage of the new law are being designed. The nation's largest and most sophisticated local authority, in Los Angeles, is examining the feasibility of using emissions trading to deal with photochemical smog, which is caused primarily by the interactions of sunlight, water vapor, oxides of nitrogen, and reactive organic gasses (ROG). Incorporating ROG into a trading system will be a major innovation, for literally hundreds of organic chemicals need to be included. ROG emissions come from a long list of industrial processes (almost all solvents emit ROG), and an equally long list of consumer products (paint, liquid cleaning solutions, hair spray, floor wax, deodorant, etc.). Nearly all ROG emissions are uncontrolled

because they come from highly dispersed and very small sources. One approach would be to set standards for literally hundreds of thousands of products, specifying the maximum amount of each vaporizing organic chemical that each product could contain. These standards would be enforced by acquiring and testing a vast array of products. The standards approach has indeed begun for some industrial processes. For example, in the manufacturing of integrated circuits, regulatory standards specify how many times per day each production worker can dip a cotton swab into a solvent for the purpose of cleaning the newly produced chip (a classic example of an input standard that is only remotely related to achieving environmental objectives).

Under the proposed ROG trading systems, regulators will control the total amount of ROG entering the region by licensing products. Each product will specify its chemical content, so that this content times its regional use determines its emissions. Producers will be given a ceiling on total ROG content, which can be met by curtailing production, changing the chemical composition of the product, or buying permits from someone else. The enforcement mechanism is as yet not clear, but one possibility would be a system similar to the stamps used in taxing alcoholic beverages and cigarettes. The stamping can be done by regulators in a single central location, or by product distributors using metered stamping devices. Enforcement would consist of checking to determine whether product containers at points of sale and use were stamped.

The Los Angeles experiment is still in the design stage, and so remains uncertain in concept and performance. Indeed, it may not be implemented in so ambitious a form because regulators may not be willing to take the risk, or because a court may stand in the way. Nonetheless, the willingness of front-line regulatory authorities even to attempt so radical a departure from traditional standard setting stands as testimony to the ripeness of these reforms.

Conclusions

Incentive regulation in environmental policy is now being given a serious chance in the United States. As yet, a relatively small part of the nation's overall environmental regulatory system is experimenting with the use of incentives, but if the experiments continue to work as well as the early efforts, radical reform may be in the works.

A common feature of the reforms to date has been that they have received the support of at least some members of both business and environmental groups. By severely constraining the controlled trading options, the Carter administration gained the grudging assent of some environmentalists. Each successive step could be a bit more flexible and expansive because the previous system had not led to any obvious failures.

An extremely important feature of the standards process was that it froze in place the basic economic structure of the country, by both industry and geography, due to PSD and NSPS. These policies greatly benefitted established production facilities in industries that contribute significantly to air pollution by raising the costs of competitive entry. These policies remain in force under the new regime. However, the third major barrier to economic restructuring – the economic feasibility constraint on standards – has been vastly weakened. Local industries with high emissions, low abatement costs, but weak standards because they cannot pass along price increases will now either abate (by selling some emissions at a price higher than their abatement costs) or will simply close down (selling all of their permits at a price that exceeds the value of the production facility). The result will be an economically warranted restructuring, with the most polluting industries either reducing emissions or leaving areas where air pollution is the worst.

A significant political advantage of emissions trading is that it enables rationalization of air pollution control efforts without eliminating the features of the standards approach that protected established facilities. Firms can capitalize the profits from their protected position by simply selling their generous allotment of emissions permits from a regulatory system that could not impose significant economic harm on them. Now that all of the obvious, relatively inexpensive controls are in place on facilities that could afford significant abatement costs, and air quality objectives are still not met, the firms with lax standards are in an enviable position. If, for example, Los Angeles decrees an equal percentage reduction in emissions from all sources, and then allocates emissions permits to each source in this new amount, sources with lax standards will be able to sell their permits at a very handsome price to the sources that already have rigorous standards. In this way, emissions trading provides a means whereby the wealth protections in the old system do not continue to stand in the way of more effective and efficient abatement strategies.

More than other regulatory reform measures, the lessons of emissions trading in the U.S. are more transparently transferrable to other nations. Many industries contribute to core air, water and waste disposal problems, so that even in a small country, numerous production facilities can be brought into a market system. For example, the decade-long bottle controversy in the European Economic Community reflects a classic problem of standards regulation. The point of encouraging bottle reuse and recycling is to reduce solid waste disposal problems and energy costs. Attacking these problems on a product-by-product basis is certain to be wasteful and ineffective, and in the process to create enclaves of protected interests (as the European Commission has claimed regarding the Danish and German bottle control systems). Energy taxes, disposal taxes, and tradable permits continent wide for using new physical materials are all more promising approaches for achieving simultaneously objectives regarding resource use, the environment, and economic efficiency.

To claim that recent reforms of environmental regulation are the wave of the future is surely premature. As yet, no complex environmental issue has been successfully attacked using incentive regulation, simply because none has been in place long enough to produce clear performance indicators. But the incentive approach is likely to have its chance during the next few years, primarily because of widespread agreement that it is the last hope for relatively painless attainment of environmental policy objectives. If it works, the implications are far greater than the consequences of liberalization of economic regulation, for far more is at stake.

Bibliographical Note

The theoretical and empirical literature on the economics and politics of regulation is extensive and rapidly growing. The purpose of this note is to provide logical entry points for readers who seek further depth on the topics discussed in the four Jean Monnet Lectures.

By far the best place to begin for a summary of the most recent scholarship on regulatory policy is Part 5 of the *Handbook of Industrial Organization*, edited by Richard Schmalensee and Robert Willig and published by North Holland in 1990. The economic effects of regulation are summarized by two essays, "The Effects of Economic Regulation" by Paul Joskow and Nancy Rose and "The Economics of Health, Safety and Environmental Regulation" by Howard Gruenspecht and Lester Lave. The politics of regulation, as illuminated by economic models of political processes, is covered in "Economics Perspectives on the Politics of Regulation" by Roger Noll. In addition, Part 5 contains two essays dealing with price regulation: "Optimal Policies for Natural Monopolies" by Ronald Brauetigam and "Design of Regulatory Mechanisms and Institutions" by David Baron. These five papers reference essentially all of the important theoretical and empirical work covered in the four lectures and published before 1989.

Lecture 1 on the politics of regulation has its origins in the three giant classics from which the modern economic theory of politics is derived: Kenneth Arrow, *Social Choice and Individual Values*; Anthony Downs, *An Economic Theory of Democracy*; and Mancur Olson, *The Logic of Collective Action*. These books are important primarily for their methodological advance, rather than their specific conclusions. Their three main messages (that majority rule voting has no equilibrium, that political campaigns have low information content, and that organized interests are advantaged in democracies) were hardly new to political scientists; however, all three books vastly expanded and deepened understanding of these three ideas by constructing careful theoretical arguments derived from microeconomic theory.

The first influential application of the economic theory of politics to regulatory policy was George Stigler's "The Theory of Economic Regulation," *Bell Journal of Economics* (1971). Stigler and his colleagues at the University of Chicago have built their theory of regulation on the principle of interest-group organization and influence. As mostly recently explicated, the Chicago theory is argued to predict that regulatory policy redistributes wealth as interest-group theory predicts, but that it tends to do so without creating substantial inefficiencies, since inefficiency creates an incentive for bargains among groups that can enhance the wealth of all. Recent statements of the Chicago theory can be found in Gary Becker, "Public Policies, Pressure Groups, and Dead Weight Costs", *Journal of Public Economics* (1985) and Sam Peltzman, "The Economic Theory of Regulation after a Decade of Deregulation", *Brookings Paper on Eco-*

nomic activity: Microeconomic (1989). The latter examines the record in several deregulated industries to test the Chicago school model.

The other primary stream of research on the politics of regulation within the economics paradigm deals with how the institutional arrangements of government shape incentives for political action and, hence, policy outcomes. The vast majority of this work deals explicitly with the American federal system, focusing on the details of the U.S. structure. The reason is that this literature focuses on how institutions solve various collective action problems, and is written almost exclusive by Americans, who naturally choose U.S. institutions to make their points. Some examples are; on delegation to bureaucrats, Mathew McCubbins, "The Legislative Design of Regulatory Structure", *American Journal of Political Science* (1985); on the relationship between the design of bureaucratic agencies and policy outcomes, Mathew McCubbins, Roger Noll and Barry Weingast, "Structure and Process, Politics and Policy", *Virginia Law Review* (1989); and on the principles underlying the organization of legislatures for policy making, Barry Weingast and William Marshall, "The industrial Organization of Congress", *Journal of Political Economy* (1988).

Political scientists do not necessarily agree with the conclusions of economic theories of regulation and deregulation. The clearest statement of the argument that ideas and key people are responsible for reform is Martha Derthick and Paul Quirk, *The Politics of Deregulation*, and for a well executed compendium of traditional political science studies of regulation, see James O. Wilson, *The Politics of Regulation.*

The recent research literature on telecommunications policy (Lecture 2) is extremely voluminous, so that identifying a short list of key sources is somewhat arbitrary. Two interesting compendia are Stephen Bradley and Jerry Hausman, *Future Competition in Telecommunications*, and Barry Cole, *After the Break-up*. For a detailed account of the antitrust case against AT&T, see Roger Noll and Bruce Owen, "The Anticompetitive Uses of Regulation: U.S. v. AT&T", in John Kwoka and Lawrence White, *The Antitrust Revolution*. For an account of the influence of Congress on post-divestiture regulation, see John Ferejohn and Charles Shipan, "Congress and Telecommunications Policymaking", in Paula Newberg, *New Directions in Telecommunications Policy*, Vol. 1. The Cole book contains a paper by Roger Noll and Susan Smart that contains extensive data about telephone prices during the period of reform, entitled "Pricing of Telephone Services."

Airline deregulation has also been extensively studied (Lecture 3). Analysis of a wide range of issues can be found in Steven Morrison and Clifford Winston, "Enhancing the Performance of the Deregulated Transportation System", *Brookings Papers on Economic Activity: Microeconomics* (1989). Severn Borenstein has written several important papers on the post-deregulation airline industry, especially "Hubs and High Fares", *Rand Journal of Economics* (analyzing the economic impor-

tance of the new route structure); "Airlines Mergers, Airport Dominance and Market Power", *American Economic Review* (1990) (examining how lax antitrust policy affected the gains from deregulation); and, with Martin Zimmerman, "Market Incentives for Safe Commercial Airline Operation", *American Economic Review* (1988) (addressing the relationship between market competition and airline safety).

In addition, Elizabeth Bailey and Jeffrey Williams, "Sources of Economic Rent in the Deregulated Airline Industry", *Journal of Law and Economics* (1988) examines why the airline industry is unlikely to produce a perfectly competitive outcome. To date, a thorough examination of the effects of deregulation on labor in the airline industry had not been undertaken to my knowledge; however, such work has been undertaken for trucking deregulation: Nancy Rose, "Labor Rent-Sharing and Regulation: Evidence from the Trucking Industry", *Journal of Political Economy* (1987).

Environmental reform (Lecture 4), being the most recent policy innovation, has been least extensively researched. For a summary of the standards system and how it could evolve into a system of emissions trading, see Roger Noll, "The Feasibility of Tradable Emissions Permits in the U.S.", in Jorg Finsinger, *Public Sectors Economics*. A summary of how the evolution actually took place can be found in Robert Hahn, "Economic Prescriptions for Environmental Problems", *Journal of Economic Perspectives* (1989). An excellent treatment of the overall effects of all forms of regulation of automobiles – emissions, fuel economy, safety – is presented in Robert Crandall, Howard Gruenspecht, Theodore Keller and Lester Lave, *Regulating the Automobile*. The classic work outlining how economic protectionism and environmentalism formed an alliance to shape air pollution policy in the 1970s is Bruce Ackerman and William Hassler, *Clean Air/Dirty Coal*; are more formal treatment of this alliance in shaping the Clean Air Act is contained in the article by McCubbins, Noll and Weingast cited above. The most detailed test of the economic theory of regulation as applied to environmental policy is Wesley Magat, Alan Krupnick and Winston Harrington, *Rules in the Making*, which contains a wealth of statistical information about water pollutions standards in the U.S. (no similar analysis has yet been undertaken for air pollution standards). Finally, Robert Hahn and Alan McGartland provide an extensive political economic analysis of the Montreal protocols for CFC emissions in "The Political Economy and Instrument Choice", *Northwestern University Law Review* (1989).