Exit, Voice, and Loyalty

Responses to Decline in Firms, Organizations, and States

Albert O. Hirschman

1970

Harvard University Press
Cambridge, Massachusetts
and London, England

To Eugenio Colorni (1909–1944),
who taught me about small ideas
and how they may grow

20 19 18 17 16 15 14 13 12 11
Library of Congress Catalog Card Number: 77-99517
ISBN 0-674-27660-4
Printed in the United States of America

1
Introduction and Doctrinal Background

Under any economic, social, or political system, individuals, business firms, and organizations in general are subject to lapses from efficient, rational, law-abiding, virtuous, or otherwise functional behavior. No matter how well a society's basic institutions are devised, failures of some actors to live up to the behavior which is expected of them are bound to occur, if only for all kinds of accidental reasons. Each society learns to live with a certain amount of such dysfunctional or mis-behavior; but lest the misbehavior feed on itself and lead to general decay, society must be able to marshal from within itself forces which will make as many of the faltering actors as possible revert to the behavior required for its proper functioning. This book undertakes initially a reconnaissance of these forces as they operate in the economy; the concepts to be developed will, however, be found to be applicable not only to economic operators such as business firms, but to a wide variety of noneconomic organizations and situations.

While moralists and political scientists have been much concerned with rescuing individuals from immoral behavior, societies from corruption, and governments from decay, economists have paid little attention to *repairable lapses* of economic actors. There are two reasons for this neglect. First, in economics one assumes either fully and undeviatingly rational behavior or, at the very least, an *unchanging level* of rationality on the part of the economic actors. Deterioration of a firm's performance may result from an adverse shift in supply and demand conditions while the willingness and ability of the firm to maximize profits (or growth rate or whatever) are unimpaired; but it could also reflect some "loss of maximizing aptitude or energy" with supply and demand factors being un-

changed. The latter interpretation would immediately raise the question how the firm's maximizing energy can be brought back up to par. But the usual interpretation is the former one; and in that case, the reversibility of changes in objective supply and demand conditions is much more in doubt. In other words, economists have typically assumed that a firm that falls behind (or gets ahead) does so *"for a good reason"*; the concept—central to this book—of a random and more or less easily "repairable lapse" has been alien to their reasoning.

The second cause of the economist's unconcern about lapses is related to the first. In the traditional model of the competitive economy, recovery from any lapse is not really essential. As one firm loses out in the competitive struggle, its market share is taken up and its factors are hired by others, including newcomers; in the upshot, total resources may well be better allocated. With this picture in mind, the economist can afford to watch lapses of any one of *his* patients (such as business firms) with far greater equanimity than either the moralist who is convinced of the intrinsic worth of every one of *his* patients (individuals) or the political scientist whose patient (the state) is unique and irreplaceable.

Having accounted for the economist's unconcern we can immediately question its justification: for the image of the economy as a fully competitive system where changes in the fortunes of individual firms are exclusively caused by basic shifts of comparative advantage is surely a defective representation of the real world. In the first place, there are the well-known, large realms of monopoly, oligopoly, and monopolistic competition: deterioration in performance of firms operating in that part of the economy could result in more or less permanent *pockets* of inefficiency and neglect; it must obviously be viewed with an alarm approaching that of the political scientist who sees his polity's integrity being threatened by strife, corruption, or boredom. But even where vigorous competition prevails, unconcern with the possibility of restoring temporarily laggard firms to vigor is hardly justified. Precisely in sectors where there are large numbers of firms competing with one another in similar conditions, declines in the fortunes of individual firms are just as likely to be due to random, subjective factors that are reversible or remediable as to permanent adverse shifts in cost and demand conditions. In these circumstances, mechanisms of recuperation would play a most useful role in avoiding social losses as well as human hardship.

At this point, it will be interjected that such a mechanism of recuperation is readily available through competition itself. Is not competition supposed to keep a firm "on its toes"? And if the firm has already slipped, isn't it the experience of declining revenue and the threat of extinction through competition that will cause its managers to make a major effort to bring performance back up to where it should be?

There can be no doubt that competition is one major mechanism of recuperation. It will here be argued, however (1) that the implications of this particular function of competition have not been adequately spelled out and (2) that a major alternative mechanism can come into play either when the competitive mechanism is unavailable or as a complement to it.

Enter "Exit" and "Voice"

The argument to be presented starts with the firm producing saleable outputs for customers; but it will be found to be largely—and, at times, principally—applicable to organizations (such as voluntary associations, trade unions, or political parties) that provide services to their members without direct monetary counterpart. The per-

formance of a firm or an organization is assumed to be subject to deterioration for unspecified, random causes which are neither so compelling nor so durable as to prevent a return to previous performance levels, provided managers direct their attention and energy to that task. The deterioration in performance is reflected most typically and generally, that is, for both firms and other organizations, in an absolute or comparative deterioration of the *quality* of the product or service provided.[1] Management then finds out about its failings via two alternative routes:

(1) Some customers stop buying the firm's products or some members leave the organization: this is the *exit option.* As a result, revenues drop, membership declines, and management is impelled to search for ways and means to correct whatever faults have led to exit.

(2) The firm's customers or the organization's members express their dissatisfaction directly to management or to some other authority to which management is subordinate or through general protest addressed to anyone who cares to listen: this is the *voice option.* As a result, management once again engages in a search for the causes and possible cures of customers' and members' dissatisfaction.

The remainder of this book is largely devoted to the

1. For business firms operating in situations of monopoly or monopolistic competition, performance deterioriation can also be reflected in cost and resulting price increases or in a combination of quality drops and price increases. On the other hand, changes in either price or quality are ruled out when both are rigidly dictated by a perfectly competitive market; in this admittedly unrealistic situation, deterioration can manifest itself only via increases in cost which, with price and quality unchanged, will lead straightaway to a decline in net revenue. Under perfect competition, then, managers learn about their failings directly and exclusively from financial evidence generated within the firm, without any intermediation on the part of the customers who remain totally unaware of the firm's troubles. It is perhaps because the whole range of phenomena here described has no place in the perfectly competitive model that it has not been paid attention to by economists.

comparative analysis of these two options and to their interplay. I will investigate questions such as: Under what conditions will the exit option prevail over the voice option and vice versa? What is the comparative efficiency of the two options as mechanisms of recuperation? In what situations do both options come into play jointly? What institutions could serve to perfect each of the two options as mechanisms of recuperation? Are institutions perfecting the exit option compatible with those designed to improve the working of the voice option?

Latitude for Deterioration, and Slack in Economic Thought

Before setting out to answer some of these questions, I shall now step back briefly and indicate how I conceive the subject of this book to be related to economic and social science thought around us.

Talking with students of animal behavior (at the Center for Advanced Study in the Behavioral Sciences) about the social organization of primates I learnt about the smoothness and efficiency with which leadership succession, a problem human societies have found so intractable, was handled in certain baboon bands. Here is how the process is described for a typical band of Hamadryas baboons lorded over by one male leader:

Sub-adult males steal very young females from their mothers and attend them with every semblance of solicitous maternal care. The young female is rigorously controlled, and repeated retrieval trains her not to go away . . . At this stage there is no sexual behaviour, the female being yet two to three years from child bearing . . . As these young interlopers mature and the overlord ages, the younger animal starts initiating group movements although the direction of eventual movement is dependent upon the older animal's choice. A highly complex relation-

ship develops between the two animals which, by paying close attention to one another and by reciprocal "notification," cooperate in governing group movement. Old males retain command of group direction but gradually relinquish sexual control over their females to the younger male animal . . . It seems that eventually old males resign entirely from their original reproduction units but retain great influence within the band as a whole, and young males refer to them continuously particularly before developing the direction of march.[2]

Compare this marvel of gradualness and continuity with the violent ups and downs to which human societies have always been subject as "bad" government followed upon "good," and as strong or wise or good leaders were succeeded by weaklings, fools, or criminals.

The reason for which humans have failed to develop a finely built social process assuring continuity and steady quality in leadership is probably that they did not have to. Most human societies are marked by the existence of a surplus above subsistence. The counterpart of this surplus is society's ability to take considerable deterioration in its stride. A lower level of performance, which would mean disaster for baboons, merely causes discomfort, at least initially, to humans.

The wide latitude human societies have for deterioration is the inevitable counterpart of man's increasing productivity and control over his environment. Occasional decline as well as prolonged mediocrity—in relation to achievable performance levels—must be counted among the many penalties of progress. A priori it would seem futile, therefore, to look for social arrangements that

2. John Hurrell Crook, "The Socio-Ecology of Primates," in J. H. Crook, ed., *Social Behaviour in Animals and Man* (to be published by Academic Press, London). The passage quoted summarizes research by Hans Kummer, "Social Organization of Hamadryas Baboons," *Bibliotheca Primatologica*, no. 6 (Basle: S. Karger, 1968).

would wholly eliminate any sort of deterioration of polities and of their various constituent entities. Because of the surplus and the resulting latitude, any homeostatic controls with which human societies might be equipped are bound to be rough.

Recognition of this unpleasant truth has been impeded by a recurring utopian dream: that economic progress, while increasing the surplus above subsistence, will also bring with it disciplines and sanctions of such severity as to rule out any backsliding that may be due, for example, to faulty political processes. In the eighteenth century the expansion of commerce and of industry was sometimes hailed not so much because of the increase in well-being that it would make possible, but because it would bring with it powerful restraints on the willfulness of the prince and thereby reduce and perhaps eliminate the system's latitude for deterioration. One characteristic passage from Sir James Steuart's *Inquiry into the Principles of Political Oeconomy* (1767) will suffice to make the point:

How hurtful soever the natural and immediate effects of political revolutions may have been formerly, when the mechanism of government was more simple than at present, they are now brought under such restrictions, by the complicated system of modern oeconomy, that the evil which might otherwise result from them may be guarded against with ease . . .

The power of a modern prince, let it be by the constitution of his kingdom ever so absolute, immediately becomes limited so soon as he establishes the plan of oeconomy . . . If his authority formerly resembled the solidity and force of the wedge (which may indifferently be made use of, for splitting of timber, stones and other hard bodies, and which may be thrown aside and taken up again at pleasure), it will at length come to resemble the delicacy of the watch, which is good for no other purpose than to mark the progression of time, and which is immediately destroyed, if put to any other use, or touched with any but

the gentlest hand . . . modern oeconomy, therefore, is the most effectual bridle ever was invented against the folly of despotism.[3]

This noble hope echoes nearly two hundred years later in the writings of a Latin American intellectual similarly predicting, against all likelihood, that economic progress and latitude for deterioration will be negatively, rather than positively, correlated:

[In the pre-coffee era, policy makers] are lyrical and romantic because they cannot yet defer to a product whose output is constantly on the increase. It is a time of childhood and play. Coffee will bring maturity and seriousness. It will not permit Colombians to continue playing fast and loose with the national economy. The ideological absolutism will disappear and the epoch of moderation and sobriety will dawn . . . Coffee is incompatible with anarchy.[4]

History has cruelly disappointed the expectations of both Sir James Steuart and Nieto Arteta that economic growth and technical progress would erect secure barriers against "despotism," "anarchy," and irresponsible behavior in general. Yet their line of thought is hardly extinct. It is, in fact, not unrelated to today's widespread belief that a major war is unthinkable and therefore impossible in the nuclear age.

The common assumption of these constructs is simply stated: while technical progress increases society's surplus above subsistence it also introduces a mechanism of the utmost complexity and delicacy, so that certain types of social misbehavior which previously had unfortunate

3. (Chicago: University of Chicago Press, 1966), I, 277, 278–279.

4. Luis Eduardo Nieto Arteta, *El café en la sociedad colombiana* (Bogotá: Breviarios de orientación colombiana, 1958), pp. 34–35. This posthumously published essay was written in 1947, only a year before the outbreak of the sanguinary civil disturbances known as *la violencia*, just as Sir James Steuart wrote about the definitive conquest of despotism not long before the rise of Napoleon.

but tolerable consequences would now be so clearly disastrous that they will be more securely barred than before.

As a result society is, and then again it is not, in a surplus situation: it is producing a surplus, but is not at liberty *not* to produce it or to produce less of it than is possible; in effect, social behavior is as simply and as rigidly prescribed and constrained as it is in a no-surplus, bare subsistence situation.

The economist cannot fail to note the similarity of the situation with the model of perfect competition. For this model contains the same basic paradox: society as a whole produces a comfortable and perhaps steadily increasing surplus, but every individual firm considered in isolation is barely getting by, so that a single false step will be its undoing. As a result, everyone is constantly made to perform at the top of his form and society as a whole is operating on its—forever expanding—"production frontier," with economically useful resources fully occupied. This image of a relentlessly *taut economy* has held a privileged place in economic analysis, even when perfect competition was recognized as a purely theoretical construct with little reality-content.

These various observations add up to a syndrome, namely, to man's fundamentally ambivalent attitude toward his ability to produce a surplus: he likes surplus but is fearful of paying its price. While unwilling to give up progress he hankers after the simple rigid constraints on behavior that governed him when he, like all other creatures, was totally absorbed by the need to satisfy his most basic drives. Who knows but that this hankering is at the root of the paradise myth! It seems plausible, indeed, that the *rise* of man above the narrowly constrained condition of all other living creatures was frequently sensed, though it can hardly ever have been avowed, as a *fall;* and a radical but basically simple act of the imagination may well have metamorphosed this condition which one was really

yearning for into its exact opposite, the Garden of Eden.[5]

But we must leave paradise and return to social thought, for there is another side to our story. The simple idea that the ability to produce a surplus above subsistence makes it possible and indeed likely that occasionally less than the maximum producible surplus will be produced has not gone wholly unnoticed. In fact, next to the traditional model of the permanently *taut* economy, elements of a theory of the *slack* economy begin to be available. I am not referring now to unemployment and depression economics —the slack associated with these phenomena results from malfunctions at the macroeconomic level which frustrate firms and individuals in their supposedly undiminished zeal to maximize profits and satisfaction. Nor is the question of slack involved in the dispute about what it is that business firms, and particularly the large corporations, really do maximize: profits, growth, market shares, community goodwill, or some composite functions of such objectives. The assumption underlying this dispute is that, whatever it is that firms do, they do it the best they can even though the criterion for "best" performance is becoming rather murky. Finally I am not concerned with the large body of writings showing that the actions of conscientiously maximizing private producers and consumers may fail to produce a *social* optimum, because of the existence of monopolistic elements and externalities.

5. Samuel Johnson intimated this thought in his fable about the Happy Valley of Abyssinia. When Prince Rasselas first analyzes the discontent he feels in the paradiselike valley, he compares his condition to that of some grazing goats in the following terms: "What makes the difference between man and all the rest of the animal creation? Every beast that strays beside me has the same corporal necessities with myself; he is hungry and crops the grass, he is thirsty and drinks the stream, his thirst and hunger are appeased, he is satisfied and sleeps; he rises again and is hungry, he is again fed and is at rest. I am hungry or thirsty like him, but when thirst or hunger cease, I am not at rest; I am like him pained with want, but am not, like him, satisfied with fulness." (Samuel Johnson, *Rasselas*, II.)

Here again the difference between actual and potential output is not due to some "failure of nerve" at the microeconomic level. But of late there has been increasing attention to the possibility of just such a failure.

A seminal contribution in this area was H. A. Simon's suggestion that firms are normally aiming at no more than a "satisfactory" rather than at the highest possible rate of profits.[6] This notion was given considerable underpinning in 1963 by Richard Cyert and James March, who in their book *A Behavioral Theory of the Firm*[7] introduced the concept of "organizational slack." At about the same time, Gary Becker showed that some of the basic and empirically well-tested microeconomic theorems (for example, that market demand curves for individual commodities are negatively inclined) are consistent with a wide range of irrational and inefficient behavior on the part of consumers and producers even though these theorems had originally been derived on the assumption of undeviating rationality.[8] The importance of slack was later affirmed in a particularly sweeping form by Harvey Leibenstein.[9] Finally, in a widely discussed polemical essay, Professor M. M. Postan has recently contended that Britain's economic ailments are better understood by focusing on microeconomic slack than on any mistaken macroeconomic policies. He writes:

6. H. A. Simon, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69:98–118 (1952). An early, completely forgotten empirical work with a related theme has the significant title *The Triumph of Mediocrity in Business*, by Horace Secrist, published in 1933 by the Bureau of Business Research, Northwestern University. The book contains an elaborate statistical demonstration that, over a period of time, initially high-performing firms will on the average show deterioration while the initial low performers will exhibit improvement.

7. Richard M. Cyert and James G. March, *Behavioral Theory of the Firm* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963).

8. Gary S. Becker, "Irrational Behavior and Economic Theory," *Journal of Political Economy*, 52:1–13 (February 1962).

9. Harvey Leibenstein, "Allocative Efficiency versus X-Efficiency," *American Economic Review*, 56:392–415 (June 1966).

For many (perhaps most) of these . . . ailments the morbid causes will be found not in the malfunctioning of the life processes in the body economic, such as the low rate of savings, or the high level of prices, or the insufficient allocation of national resources to research and development, but in specific failures of its individual cells—management, design, salesmanship, or the behavior of groups of labor.[10]

I feel considerable kinship with this group of writings for I had adopted a similar position in dealing with the problem of development. The basic proposition of *The Strategy of Economic Development* (1958) was that "development depends not so much on finding optimal combinations for given resources and factors of production as on calling forth and enlisting for development purposes resources and abilities that are hidden, scattered or badly utilized." [11] And the term slack actually came under my pen when I summarized later on the essential argument of that book in an article co-authored with C. E. Lindblom:

At any one point of time, an economy's resources are not to be considered as rigidly fixed in amount, and more resources or factors of production will come into play if development is marked by sectoral imbalances that galvanize private entrepreneurs or public authorities into action . . . The crucial, but plausible, assumption here is that there is some "slack" in the economy; and that additional investment, hours of work, productivity, and decision making can be squeezed out of it by pressure mechanisms.[12]

Various reasons have been invoked for explaining slack. Leibenstein's emphasis is on the uncertainties surrounding the production function and on the nonmarketability

10. M. M. Postan, "A Plague of Economists?" *Encounter* (January 1968), p. 44.
11. (New Haven: Yale University Press, 1958), p. 5.
12. "Economic Development, Research and Development, Policy Making: Some Converging Views," *Behavioral Science*, 7:211–212 (April 1962).

of managerial and other skills. Cyert and March refer primarily to the bargaining process that takes place among the various parties whose (shaky) coalition is required for factors to be hired and for output to be produced and marketed. I stressed rather similarly the existence of obstacles to entrepreneurial and cooperative behavior needed for the making of development decisions.

Those who have found that the individual economic operators and, as a result, the economy are ordinarily far from doing as well as they might, can be expected to react to their shocking discovery along two principal lines. The immediate and most obvious reaction is a determined search for ways and means to take up the slack, to retrieve the ideal of the taut economy. As long as the pressures of competition do not seem to be sufficient, the pressures of adversity will be invoked.[13] Frequent changes in the environment, forcing the firm to be "on its toes," will be identified as one way of inducing performance closer to the firm's potential.[14] Insofar as innovation is concerned, the inducing and focusing virtues of strikes and war have been stressed.[15] My own search concentrated on pressure mechanisms such as intersectoral and intrasectoral imbalances and on production processes that exact high penalties for poor performance or do not tolerate it at all.[16] Finally, the advocates of social revolution have contributed to this line of thought: one of their most seductive arguments has long been that only revolutionary changes can tap and liberate the abundant but dormant, repressed, or alienated energies of the people.[17]

13. See Leibenstein, "Allocative Efficiency versus X-Efficiency."
14. Charles P. Bonini, "Simulation of Information and Decision Systems in the Firm" (unpub. diss. Carnegie Institute of Technology, 1962).
15. Nathan Rosenberg, "The Direction of Technological Change: Inducement Mechanisms and Focusing Devices," *Economic Development and Cultural Change*, 18 (October 1969).
16. Hirschman, *Strategy*, chs. 5–8.
17. See, for example, Paul Baran, *The Political Economy of Growth* (New York: Monthly Review Press, 1957).

Quite a different reaction to the discovery of slack occurs when the discoverer asks himself, after having got over his initial shock, whether slack may not after all be a good thing, a blessing in disguise. The idea that slack fulfills some important, if unintended or latent, functions was put forward by Cyert and March, who point out that it permits firms to ride out adverse market or other developments. During such bad times slack acts like a reserve that can be called upon: excess costs will be cut, innovations that were already within one's grasp will at last be introduced, more aggressive sales behavior that had been shunned will now be engaged in, and so on. Slack in the political system has been rationalized in a very similar manner. The discovery that citizens do not normally use more than a fraction of their political resources came originally as a surprise and disappointment to political scientists who had been brought up to believe that democracy requires for its functioning the fullest possible participation of all citizens. But soon enough a degree of apathy was found to have some compensating advantages in as much as it contributes to the stability and flexibility of a political system and provides for "reserves" of political resources which can be thrown into the battle in crisis situations.[18]

The immediate response to the discovery of slack has thus been either to assert the rationality of a certain level of slack or to look for ways of extirpating excessive levels by invoking exogenous forces such as adversity, imbalances, revolution, and so on. Both these approaches look at slack as a gap of a given magnitude between actual and potential performance of individuals, firms, and organizations. This book takes a further, more radical step in recognizing the importance and pervasiveness of slack. It assumes not only that slack has somehow come into the

18. See below, pp. 31–32.

world and exists in given amounts, but that it is *continuously being generated* as a result of some sort of entropy characteristic of human, surplus-producing societies. "There's a slacker born every minute," could be its motto. Firms and other organizations are conceived to be permanently and randomly subject to decline and decay, that is, to a gradual loss of rationality, efficiency, and surplus-producing energy, no matter how well the institutional framework within which they function is designed.

This radical pessimism, which views decay as an ever-present force constantly on the attack, generates its own cure: for as long as decay, while always conspicuous in some areas, is hardly in undisputed command everywhere and at all times, it is likely that the very process of decline activates certain counterforces.

Exit and Voice as Impersonations of Economics and Politics

In examining the nature and strength of these endogenous forces of recovery, our inquiry bifurcates, as already explained. Its breakup into the two contrasting, though not mutually exclusive, categories of exit and voice would be suspiciously neat if it did not faithfully reflect a more fundamental schism: that between economics and politics. Exit belongs to the former realm, voice to the latter. The customer who, dissatisfied with the product of one firm, shifts to that of another, uses the market to defend his welfare or to improve his position; and he also sets in motion market forces which may induce recovery on the part of the firm that has declined in comparative performance. This is the sort of mechanism economics thrives on. It is neat—one either exits or one does not; it is impersonal —any face-to-face confrontation between customer and

firm with its imponderable and unpredictable elements is avoided and success and failure of the organization are communicated to it by a set of statistics; and it is indirect —any recovery on the part of the declining firm comes by courtesy of the Invisible Hand, as an unintended by-product of the customer's decision to shift. In all these respects, voice is just the opposite of exit. It is a far more "messy" concept because it can be graduated, all the way from faint grumbling to violent protest; it implies articulation of one's critical opinions rather than a private, "secret" vote in the anonymity of a supermarket; and finally, it is direct and straightforward rather than roundabout. Voice is political action par excellence.

The economist tends naturally to think that his mechanism is far more efficient and is in fact the only one to be taken seriously. A particularly good illustration of this bias appears in a well-known essay by Milton Friedman which advocates the introduction of the market mechanism into public education. The essence of the Friedman proposal is the distribution of special-purpose vouchers to parents of school-age children; with these vouchers the parents could buy educational services that would be supplied in competition by private enterprise. In justifying this scheme he says:

> Parents could express their views about schools *directly,* by withdrawing their children from one school and sending them to another, to a much greater extent than is now possible. In general they can now take this step only by changing their place of residence. *For the rest, they can express their views only through cumbrous political channels.*[19]

19. "The Role of Government in Education," in Robert A. Solo, ed., *Economics and the Public Interest* (New Brunswick, N.J.: Rutgers University Press, 1955), p. 129. A revised form of this essay was included in Friedman's *Capitalism and Freedom* (Chicago: University of Chicago Press, 1962) as ch. 6 and the cited passage appears unchanged on p. 91. The italics are mine.

I am not interested here in discussing the merits of the Friedman proposal.[20] Rather, I am citing the above passage as a near perfect example of the economist's bias in favor of exit and against voice. In the first place, Friedman considers withdrawal or exit as the "direct" way of expressing one's unfavorable views of an organization. A person less well trained in economics might naïvely suggest that the direct way of expressing views is to express them! Secondly, the decision to voice one's views and efforts to make them prevail are contemptuously referred to by Friedman as a resort to "cumbrous political channels." But what else is the political, and indeed the democratic, process than the digging, the use, and hopefully the slow improvement of these very channels?

In a whole gamut of human institutions, from the state to the family, voice, however "cumbrous," is all their members normally have to work with. Significantly, one major, if problem-plagued, effort presently underway toward better public schools in the large cities is to make them more responsive to their members: decentralization has been advocated and undertaken as a means of making the channels of communication between members and management in the public school systems less "cumbrous" than heretofore.

But the economist is by no means alone in having a blindspot, a "trained incapacity" (as Veblen called it) for perceiving the usefulness of one of our two mechanisms. In fact, in the political realm exit has fared much worse than has voice in the realm of economics. Rather than as merely ineffective or "cumbrous," exit has often been branded as *criminal,* for it has been labeled desertion, defection, and treason.

Clearly, passions and preconceptions must be reduced

20. For a good discussion see Henry M. Levin, "The Failure of the Public Schools and the Free Market Remedy," *The Urban Review,* 2:32–37 (June 1968).

on both sides if advantage is to be taken of an exceptional opportunity to observe how a typical market mechanism and a typical nonmarket, political mechanism work side by side, possibly in harmony and mutual support, possibly also in such a fashion that one gets into the other's way and undercuts its effectiveness.

A close look at this interplay between market and nonmarket forces will reveal the usefulness of certain tools of economic analysis for the understanding of political phenomena, *and vice versa*. Even more important, the analysis of this interplay will lead to a more complete understanding of social processes than can be afforded by economic or political analysis in isolation. From this point of view, this book can be viewed as the application to a new field of an argument on which much of *The Strategy of Economic Development* was based:

> Tradition seems to require that economists argue forever about the question whether, in any disequilibrium situation, *market forces acting alone* are likely to restore equilibrium. Now this is certainly an interesting question. But as social scientists we surely must address ourselves also to the broader question: is the disequilibrium situation likely to be corrected at all, by market or nonmarket forces, or by both acting jointly? *It is our contention that nonmarket forces are not necessarily less "automatic" than market forces.*[21]

I was concerned here with disturbances of equilibrium and the return to it. Kenneth Arrow has argued along very similar lines for movements from less-than-optimal to optimal states:

> I propose here the view that, when the market fails to achieve an optimal state, society will, to some extent at least, recognize the gap, and nonmarket social institutions

21. Hirschman, *Strategy*, p. 63. Italics in the original.

> will arise attempting to bridge it . . . this process is not necessarily conscious.[22]

These views do not imply, as both Arrow and I immediately hastened to add, that any disequilibrium or nonoptimal state whatever will be eliminated by some combination of market and nonmarket forces. Nor do they exclude the possibility that the two sets of forces could work at cross-purposes. But they leave room for a conjunction—which could quite possibly be inadequate—of these two forces, whereas both laissez-faire and interventionist doctrines have looked at market and nonmarket forces in a strictly Manichaean way, it being understood that the laissez-faire advocate's forces of good are the interventionist's forces of evil and vice versa.

A final point. Exit and voice, that is, market and nonmarket forces, that is, economic and political mechanisms, have been introduced as two principal actors of strictly equal rank and importance. In developing my play on that basis I hope to demonstrate to political scientists the usefulness of economic concepts *and to economists the usefulness of political concepts*. This reciprocity has been lacking in recent interdisciplinary work as economists have claimed that concepts developed for the purpose of analyzing phenomena of scarcity and resource allocation can be successfully used for explaining political phenomena as diverse as power, democracy, and nationalism. They have thus succeeded in occupying large portions of the neighboring discipline while political scientists—whose inferiority complex vis-à-vis the tool-rich economist is equaled only by that of the economist vis-à-vis the physicist—have shown themselves quite eager to be colonized and have often actively joined the invaders. Perhaps it

22. "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53:947 (December 1963).

takes an economist to reawaken feelings of identity and pride among our oppressed colleagues and to give them a sense of confidence that their concepts too have not only *grandeur*, but *rayonnement* as well? I like to think that this could be a by-product of the present essay.

2
Exit

The availability to consumers of the exit option, and their frequent resort to it, are characteristic of "normal" (non-perfect) competition, where the firm has competitors but enjoys some latitude as both price-maker and quality-maker—and therefore, in the latter capacity, also as a quality-spoiler. As already mentioned, the exit option is widely held to be uniquely powerful: by inflicting revenue losses on delinquent management, exit is expected to induce that "wonderful concentration of the mind" akin to the one Samuel Johnson attributed to the prospect of being hanged.

Nevertheless the precise modus operandi of the exit option has not received much attention, to judge from a determined though inevitably fragmentary search of the vast literature on competition.[1] Most authors are content with general references to its "pressures" and "disciplines."

Insofar as the apologetic literature is concerned, this neglect of what could be considered one of the principal virtues of the "free enterprise system" may be particularly surprising; but some of the reasons for it have already been suggested. Those who celebrate the invigorating qualities of competition are loath to concede that the system could fail for even a single moment to make everybody perform at his peak form; should such a failure nevertheless occur in the case of some firm, that firm must *ipso facto* be assumed to be mortally sick and to be ready to leave the stage while some vigorous newcomer is presumably waiting in the wings to take its place. This "view of the American economy . . . as a biological process in which the old and the senile are continually being replaced by the young and the vigorous," as Galbraith puts it mock-

1. Which was carried out by David S. French.

4

A Special Difficulty in Combining Exit and Voice

The groundwork has now been laid for telling the reader about the empirical observation that was mentioned in the Preface as the origin of this essay. In a recent book, I tried to explain why the Nigerian railways had performed so poorly in the face of competition from trucks, even for such long-haul, bulky cargo as peanuts (which are grown in Northern Nigeria, some eight hundred miles from the ports of Lagos and Port d'Harcourt). Specific economic, socio-political, and organizational reasons could be found for the exceptional ability of the trucks to get the better of the railroads in the Nigerian environment; but having done so I still had to account for the prolonged incapacity of the railroad administration to correct some of its more glaring inefficiencies, *in spite of active competition,* and proposed the following explanation:

> The presence of a ready alternative to rail transport makes it less, rather than more, likely that the weaknesses of the railways will be fought rather than indulged. With truck and bus transportation available, a deterioration in rail service is not nearly so serious a matter as if the railways held a monopoly for long-distance transport—it can be lived with for a long time without arousing strong public pressures for the basic and politically difficult or even explosive reforms in administration and management that would be required. This may be the reason public enterprise, not only in Nigeria but in many other countries, has strangely been at its weakest in sectors such as transportation and education where it is subjected to competition: instead of stimulating improved or top performance, the presence of a ready and satisfactory substitute for the services public enterprise offers merely deprives it of a precious feedback mechanism that operates at its best when the customers are securely locked in. For the management

> of public enterprise, always fairly confident that it will not be let down by the national treasury, may be less sensitive to the loss of revenue due to the switch of customers to a competing mode than to the protests of an aroused public that has a vital stake in the service, has no alternative, and will therefore "raise hell." [1]

In Nigeria, then, I had encountered a situation where the combination of exit and voice was particularly noxious for any recovery: exit did not have its usual attention-focusing effect because the loss of revenue was not a matter of the utmost gravity for management, while voice did not work as long as the most aroused and therefore the potentially most vocal customers were the first ones to abandon the railroads for the trucks. It is particularly this last phenomenon that must be looked at more closely, for if it has any generality, then the chances that voice will ever act in conjunction with exit would be poor and voice would be an effective recuperation mechanism only in conditions of full monopoly "when the customers are securely locked in."

As a preliminary to generalizing about this sort of situation, another example, closer to home, may be helpful. If public and private schools somewhere in the United States are substituted in the story for the railroads and lorries of Nigeria, a rather similar result follows. Suppose at some point, for whatever reason, the public schools deteriorate. Thereupon, increasing numbers of quality-education-conscious parents will send their children to private schools.[2] This "exit" may occasion some impulse toward an improvement of the public schools; but here again this im-

1. *Development Projects Observed* (Washington: Brookings Institution, 1967), pp. 146–147.

2. Private schools being costly and income distribution unequal, the public schools will of course be deserted primarily by the wealthier parents. Nevertheless, the willingness to make a financial sacrifice for the sake of improving the children's education differs widely within a given income class, especially at intermediate levels of income. In its pure form, the phenomenon here described is best visualized for a school district with many middle-class parents for whom

pulse is far less significant than the loss to the public schools of those member-customers who would be most motivated and determined to put up a fight against the deterioration if they did not have the alternative of the private schools.

In the preceding examples, insensitivity to exit is exhibited by public agencies that can draw on a variety of financial resources outside and independent of sales revenue. But situations in which exit is the predominant reaction to decline while voice might be more efficacious in arresting it can also be observed in the sphere of private business enterprise. The relation between corporate management and the stockholders is a case in point. When the management of a corporation deteriorates, the first reaction of the best-informed stockholders is to look around for the stock of better-managed companies. In thus orienting themselves toward exit, rather than toward voice, investors are said to follow the Wall Street rule that "if you do not like the management you should sell your stock." According to a well-known manual this rule "results in perpetuating bad management and bad policies." Naturally it is not so much the Wall Street rule that is at fault as the ready availability of alternative investment opportunities in the stock market which makes any resort to voice rather than to exit unthinkable for any but the most committed stockholder.[3]

---

the decision to send the children to private school is a significant, yet tolerable burden.

3. The passages in quotes are from B. Graham and D. L. Dodd, *Security Analysis*, 3d ed. (New York: McGraw-Hill, 1951), p. 616. The argument is spelled out in some detail in ch. 50, "Stockholder-Management Controversies." In the fourth edition of this work (1962), the authors return only briefly to this argument, and seem to be aware that the institutional odds are heavily stacked against any substantial success of their exhortations: "In quixotic fashion perhaps," they say wistfully, "we wanted to combat the traditional but harmful notion that if a stockholder doesn't like the way his company is run he should sell his shares, no matter how low their price may be" (p. 674).

While it is most clearly revealed in the private-public school case, one characteristic is crucial in all of the foregoing situations: those customers who care *most* about the quality of the product and who, therefore, are those who would be the most active, reliable, and creative agents of voice are for that very reason also those who are apparently likely to exit first in case of deterioration.

One interest of this observation is that it could define a whole class of economic structures where a tight monopoly would be preferable, within the framework of the "slack" or "fallible" economy, to competition. But before jumping to this conclusion, we must take a closer look at the observation by translating it into the ordinary language of economic analysis.

In terms of that language, the situations just described have more than a faint odor of paradox. We all know that when the price of a commodity goes up, it is the *marginal* customer, the one with the smallest consumer surplus, the one, that is, who cares *least,* who drops out first. How is it then that with a decline in quality the opposite seems quite plausible: *Is it possible that the consumers who drop out first as price increases are not the same as those who exit first when quality declines?* [4] If this question were to be answered in the affirmative, it would be easier to understand why combining exit and voice is so troublesome in some situations.

The basic reason for our paradox lies in the still insufficiently explored role of quality (as contrasted with price) in economic life. Traditional demand analysis is overwhelmingly in terms of price and quantity, categories which have the immense advantage of being recorded, measurable, and finely divisible. Quality changes have usually been dealt with by economists and statisticians

4. Appendix C refers to this possibility as the "reversal phenomenon." The discussion in the following pages should be read in conjunction with Appendixes C and D by those who find diagrams clearer than language.

through the concept of the *equivalent* price or quantity change. An article of poor quality can often be considered to be simply less in quantity than the same article of standard quality; this is the case, for example, of the automobile tire which lasts on the average only half as long (in terms of mileage) as a high quality tire. Alternatively, poor quality can often be translated into higher costs and prices; for example, increased pilferage in the rendering of railroad freight service will result in higher insurance premiums. In the latter case, a large part of the quality deterioration can be described by the statement: "now everybody really pays more for the same railroad service than before." To the extent that this statement is correct, there would be no reason to expect the effect of quality deterioration on demand (that is, for who gets out first) to be any different from the effect of a uniform rise in price. In other words, if a quality decline can be fully expressed as an equivalent rise in price that is *uniform for all buyers* of the article, the effects on customer exit of the quality decline and of the equivalent rise in price would be identical.

The crucial point can now be made. For any one individual, a quality change can be translated into equivalent price change. But this equivalence *is frequently different for different customers of the article because appreciation of quality differs widely among them.* This is so to some extent even in the just mentioned case of automobile tires and of increased pilferage of freight sent by rail. Appreciation of the longer life of quality tires will depend on the time discount of each individual buyer. In the case of rail freight, the increase in the insurance premium fully offsets only the increase in average direct monetary costs which is occasioned to the shipper by the worsening in service. For some shippers this may be all they care about, but there will surely be others for whom the lessened reliability of rail service represents costs (in inconvenience,

reputation of their own reliability, and so forth) that cannot be fully made good through an insurance scheme. That appreciation of quality—of wine, cheese, or of education for one's children—differs widely among different groups of people is surely no great discovery. It implies, however, that a given deterioration in quality will inflict very different losses (that is, different equivalent price increases) on different customers; someone who had a very high consumer surplus before deterioration precisely because he is a connoisseur and would be willing to pay, say, twice the actual price of the article at its original quality, may drop out as a customer as soon as quality deteriorates, provided a nondeteriorated competing product is available, be it at a much higher price.

Here, then, is the rationale for our observation: in the case of "connoisseur goods"—and, as the example of education indicates, this category is by no means limited to quality wines—the consumers who drop out when quality declines are not necessarily the marginal consumers who would drop out if price increased, but may be intramarginal consumers with considerable consumer surplus; or, put more simply, the consumer who is rather insensitive to price increases is often likely to be highly sensitive to quality declines.

At the same time, consumers with a high consumer surplus are, for that very reason, those who have most to lose through a deterioration of the product's quality. Therefore, they are the ones who are most likely to make a fuss in case of deterioration until such time as they do exit. "You can actively flee, then, and you can actively stay put." This phrase of Erik Erikson [5] applies with full force to the choice that is typically made by the quality-conscious consumer or the member who cares deeply about the policies pursued by the organization to which he be-

5. *Insight and Responsibility* (New York: W. W. Norton & Co., Inc., 1964), p. 86.

longs. To make that kind of consumer and member "actively stay put" for a while should be a matter of considerable concern for many firms and organizations, and particularly for those, of course, that respond more readily to voice than to exit.

Before the varieties of consumer behavior in the case of connoisseur goods are further explored, a brief homage to the hoary concept of consumer surplus is in order, for it appears to have the useful property of measuring the potential for the exercise of influence on the part of different consumers. This potential is the counterpart of the concept's traditional content. Consumer surplus measures the gain to the consumer of being able to buy a product at its market price: the larger that gain the more likely is it that the consumer will be motivated to "do something" to have that gain safeguarded or restored. In this way it is possible to derive the chances for political action from a concept that has dwelt so far exclusively in the realm of economic theory.[6]

Evidently the nature of the available substitute has something to do with the question whether or not connoisseur goods will be rapidly forsaken, in case of deterioration, by the more quality-conscious customers. In the discussion of the exit and voice options in Chapter 3, it was assumed that the only available competing or substitute good was initially of inferior quality, but carried the same price tag. Usually, of course, many other combinations of price and quality exist: in particular, consumers may often have had some hesitation between the good they actually bought, a better-quality substitute with a higher price, and a poorer-quality substitute with a lower price. Suppose now that only the former type of substitute exists

6. For a similar transformation of a time-honored economic concept, the gain from trade, into a political category, namely the influence a trading partner may acquire in the gain-receiving country, see my *National Power and the Structure of Foreign Trade* (Berkeley: University of California Press, 1945, rev. ed. 1969), ch. 2.

and that the quality of the connoisseur good normally bought by a group of consumers deteriorates. In this case it is immediately plausible that the consumers who valued the deteriorating good most will be the first ones to decide that it is worth their while to go over to the higher-quality, higher-price substitute. If only a lower-price, lower-quality good is available, on the other hand, these highly quality-conscious consumers, even though they suffer greatly as a result of quality deterioration, will stick with it longer than their less quality-conscious colleagues. These and similar propositions can be easily proved by indifference curve analysis.[7]

Hence the rapid exit of the highly quality-conscious customers—a situation which paralyzes voice by depriving it of its principal agents—is tied to the availability of better-quality substitutes at higher prices. Such a situation has, for example, been observed in the field of housing. When general conditions in a neighborhood deteriorate, those who value most highly neighborhood qualities such as safety, cleanliness, good schools, and so forth will be the first to move out; they will search for housing in somewhat more expensive neighborhoods or in the suburbs and will be lost to the citizens' groups and community action programs that would attempt to stem and reverse the tide of deterioration. Reverting to the public-private school case, it now appears that the "lower-priced" public schools have several strikes against them in their competition with private schools: first, if and when there is a deterioration in the quality of public school education these schools will lose the children of those highly quality-conscious parents who might otherwise have fought deterioration; second, if, thereafter, quality comes to decline in the private schools, then this type of parents will keep their children there for much longer than was the case

7. See Appendix D, which also discusses in more technical terms a number of other points made in this section.

when the public schools deteriorated. Hence, when public and private schools coexist, with the quality of education in the latter being higher, deterioration will be more strenuously fought "from within" in the case of the private than in that of the public schools. And because exit is not a particularly powerful recuperation mechanism in the case of public schools—it is far more so in that of private schools which have to make ends meet—the failure of one of our two mechanisms is here compounded by the inefficiency of the other.

The relevance of the foregoing observation is greatest in certain important discontinuous choices and decisions, such as between two kinds of educational institutions or two modes of transportation.[8] If one assumes a complete and continuous array of varieties, from cheap and poor-quality to expensive and high-quality, then deterioration of any but the top and bottom variety will rapidly lead to a combination of exits: the quality-conscious consumers move to the higher-price, higher-quality products and the price-conscious ones go over to the lower-price, lower-quality varieties; the former will still tend to get out first

8. In Appendix D it is shown that the reversal phenomenon can occur only when there are at least three goods: the intermediate variety which is the one that deteriorates or whose price increases, another variety that is higher-priced and higher-quality, and a third with the opposite characteristics. In this constellation the less demanding consumers will exit first (toward the lower-priced, lower-quality good) when the *price* of the intermediate good increases, whereas the quality-conscious consumer will exit first (toward the higher-priced, higher-quality good) when *quality* decreases. Even though in the above example only two goods are made explicit, namely public and private school education, the required third alternative on the "other side" of the normally bought good would be present if there were a price increase for public education, namely, informal education at home. This would no doubt be the alternative chosen by many of the less demanding consumers if public schools ceased being free. Hence the presence of the reversal phenomenon cannot be ruled out in this case. A similar reasoning applies to other seemingly dichotomous choices: upon looking more closely, it is usually found that a third alternative exists; some inferior commodity can be found in case the price of the usually bought good increases.

when it is quality that declines rather than price that rises, but the latter will not be far behind.

The proposition that voice is likely to play a more important role in opposing deterioration of high-quality products than of lower-quality products can nevertheless be maintained for the case of a good with many varieties, if these varieties can be assumed not to be spread with equal *density* over the whole quality range. If only because of economies of scale, it is plausible that density is lower in the upper ranges of quality than in the lower and middle ranges. If this is so then deterioration of a product in the upper quality ranges has to be fairly substantial before the quality-conscious will exit and switch to the next better variety. Hence the scope for, and resort to, the voice option will be greatest in these ranges; it will be comparatively slight in the medium- and low-quality ranges.

This finding permits two inferences. First, it can be related to the discussion of education which suggested that the role of voice in fending off deterioration is particularly important for a number of essential services largely defining what has come to be called the "quality of life." Hence, a disconcerting, though far from unrealistic, conclusion emerges: since, in the case of these services, resistance to deterioration requires voice and since voice will be forthcoming more readily at the upper than at the lower quality ranges, the cleavage between the quality of life at the top and at the middle or lower levels will tend to become more marked. This would be particularly the case in societies with upward social mobility. In societies which inhibit passage from one social stratum to another, resort to the voice option is automatically strengthened: everyone has a strong motivation to defend the quality of life at his own station. That cleavages between the upper and lower classes tend to widen and to become more rigid in upwardly mobile societies has become increasingly obvious;

but it has not been an easy observation to make in a culture in which it had long been taken for granted that equality of opportunity combined with upward social mobility would assure both efficiency and social justice.[9]

A rather different inference results if the assumption of a progressive thinning out of varieties at the upper end of the quality scale is brought into contact with the plausible notion that a combination of exit and voice is needed for best results. If this notion is accepted, then the recuperation mechanism may rely too much on exit at the lower end of the quality scale, *but suffer from a deficiency of exit at the upper end.* An illustration of the latter proposition will be found toward the end of the book.

9. The fallacies of this belief were laid bare in Michael Young's incisive satire *The Rise of Meritocracy* (1958, Penguin Edition 1968). See also below, pp. 108–112.

5

How Monopoly Can be Comforted by Competition *

The realization that a tight monopoly is preferable under certain circumstances to a looser arrangement in which competition is present comes hard to a Western economist. Nonetheless, the preceding argument compels recognition that a no-exit situation will be superior to a situation with some limited exit on two conditions:

(1) if exit is ineffective as a recuperation mechanism, but does succeed in draining from the firm or organization its more quality-conscious, alert, and potentially activist customer or members; and

(2) if voice could be made into an effective mechanism once these customers or members are securely locked in.

There are doubtless many situations in which the first condition applies—some additional examples will be given in this and later chapters. The second condition is a very large subject indeed: as was already pointed out, to develop "voice" within an organization is synonymous with the history of democratic control through the articulation and aggregation of opinions and interests.

By itself, the fact that the members or customers are locked in cannot therefore ensure that an effective volume of voice will be forthcoming. As will be argued below, one important way of bringing influence to bear on an organization is to threaten exit to the rival organization. But this threat cannot be made when there is no rival, so that voice is not only handicapped when exit is possible, but also, though in a quite different way, when it is not. Neverthe-

*In writing this chapter I inexcusably failed to refer to John Hicks's celebrated statement of 35 years ago: "The best of all monopoly profits is a quiet life." Had I remembered it, I would have been rather less critical about the economist's neglect of the "lazy monopolist." At the same time, I would have been able to express even more sharply the principal point of the chapter: On certain assumptions about the existence and intensity of voice, competition can afford an even quieter life than does monopoly.—A.O.H., September 30, 1971.

less, it is often possible to make probabilistic statements such as: considering the authority structure and responsiveness of organizations in a given society, and the general readiness of individuals and groups to assert their interests, it is likely that in this or that particular case, voice is going to do a more creditable job of maintaining efficiency when the customers or members are locked in than when some exit is available.[1]

Perhaps the best way of looking at the matter is to recognize that we face here a choice of two evils. Next to the traditional full-fledged monopoly whose dangers and possible abuses have long been exposed, attention should also be paid to those organizations whose monopoly powers are less complete, but who are characterized by sturdy, if undistinguished survival after exit of the more alert customers or members. Often there will be a real question which one of these two institutional varieties is the more unsatisfactory.

The point of view here adopted contrasts with the spirit

1. One may note an interesting symmetry here with the case of perfect competition. As pointed out in ch. 1, n. 1, the firm which produces for a perfectly competitive market finds out about its failings directly through increases in its costs rather than indirectly through customers' reactions because it cannot change either the price or the quality of its product. It will experience losses which will depend on the size of its lapse from efficiency. If the lapse is small, small too will be the losses and the firm will have an opportunity to recover. If one moves just a small step away from perfect competition, to a situation, that is, where the firm has some market power as a price- and quality-maker while demand remains very elastic, then one lands in a very different situation: a small lapse can produce a slightly deteriorated product which will lead to so large a loss of revenue that the firm immediately succumbs. It is now suggested that a similar situation may prevail at the other end of the spectrum. In some situations, a full monopoly may be preferable, from the point of view of the effectiveness of our recuperation mechanisms, to a monopoly just slightly hampered by competition. For this limited competition may result in revenue losses too small to alert management to its failings while it could weaken voice decisively by drawing away from the firm its most vocal customers. At both extremes of perfect competition and pure monopoly the recuperative mechanism may therefore work better than if only a *small step* were made from these extremes in the direction of market power and competitive structure, respectively.

that has long animated the concern over monopoly and the struggle against it. The monopolist has traditionally been expected to utilize to the utmost his ability to exploit the consumer and to maximize profits by restricting production. Public policies have been based primarily on this expectation. Even Galbraith, ordinarily so ready to repudiate the "conventional wisdom," takes this exploitative behavior to be the prime and perhaps only danger which must be guarded against. In his *American Capitalism* he merely pointed out that competition has become an unrealistic alternative to the monopolistic tendencies of advanced capitalist economies and extolled an alternative, already existing remedy, to wit, "countervailing power." But what if we have to worry, not only about the profit-maximizing exertions and exactions of the monopolist, but about his proneness to inefficiency, decay, and flabbiness? This may be, in the end, the more frequent danger: the monopolist sets a high price for his products not to amass super-profits, but because he is unable to keep his costs down; or, more typically, he allows the quality of the product or service he sells to deteriorate without gaining any pecuniary advantage in the process.[2]

In view of the spectacular nature of such phenomena as exploitation and profiteering, the nearly opposite failings which monopoly and market power allow, namely, laziness, flabbiness, and decay have come in for much less scrutiny. To find these problems recognized as public policy issues one has to look beyond the "Anglo-Saxon" world where economic thinking is usually carried on in terms of some maximizing or "taut economy" model. When, a few years ago, a prestigious French economic official put forward proposals for various public controls of business, he did single out incompetence and "abandon"

2. Compare the following remark of a student of Brazilian society: "The large Brazilian landholding is an evil not because it is inhuman and brutal, but because it is inefficient." Jacques Lambert, *Os dois Brasis* (Rio de Janeiro: INEP-Ministerio da Educação e Cultura, 1963), p. 120.

on the part of faltering corporate management as an important problem.[3]

Political power is very much like market power in that it permits the powerholder to indulge either his brutality or his flaccidity. But here again the dangers of abuse of power, of invasion of individuals' rights have—for very good reasons—stood in the center of attention, rather than those of maladministration and bureaucratic ineptitude. Accordingly, the original purpose of the now so widely discussed office of ombudsman was to help redress citizens' grievances against officials who had exceeded the constitutional limits of their power. Later, however, the institution experienced a "shift in its main purpose" which today "has become promotion of better administration," the correction of malpractices and the like.[4] This presumably means that the institution is now also used to correct and reprimand official *indolence* though it was originally devised for the purpose of stemming abuses of power on the part of overactive and overbearing officials.

Such versatility is admirable, but cannot be expected to be the rule. It would be surprising if every one of the safeguards against a monopolist's single-minded pursuit of profits turned out to do double duty as a cure of his propensity toward flabbiness and distraction. Exit-competition is a case in point. While of undoubted benefit in the case of the exploitative, profit-maximizing monopolist, the

3. François Bloch-Lainé, *Pour une réforme de l'entreprise* (Paris: Editions du Seuil, 1963), pp. 54–57, 76–77. "Anglo-Saxon" literature, particularly on trade unions, has paid some attention to the possible existence of "sleepy" or "lazy" monopolies. See, for example, Richard A. Lester, *As Unions Mature* (Princeton: Princeton University Press, 1958), pp. 56–60, and Lloyd G. Reynolds and Cynthia H. Taft, *The Evolution of Wage Structure* (New Haven: Yale University Press, 1956), p. 190. But the exploitative potential of the monopoly has always stood in the center of the discussion and it has been the exclusive motive for regulation and antitrust legislation.

4. Hing Yong Cheng, "The Emergence and Spread of the Ombudsman Institution," *The Annals*, special issue on "The Ombudsman or Citizen's Defender" (May 1968), p. 23.

presence of competition could do more harm than good when the main concern is to counteract the monopolist's tendency toward flaccidity and mediocrity. For, in that case, exit-competition could just fatally weaken voice along the lines of the preceding section, without creating a serious threat to the organization's survival. This was so for the Nigerian Railway Corporation because of the ease with which it could dip into the public treasury in case of deficit. But there are many other cases where competition does not restrain monopoly as it is supposed to, but *comforts and bolsters* it by unburdening it of its more troublesome customers. As a result, one can define an important and too little noticed type of monopoly-tyranny: a limited type, an oppression of the weak by the incompetent and an exploitation of the poor by the lazy which is the more durable and stifling as it is both *unambitious and escapable*. The contrast is stark indeed with totalitarian, expansionist tyrannies or the profit-maximizing, accumulation-minded monopolies which may have captured a disproportionate share of our attention.

In the economic sphere such "lazy" monopolies which "welcome competition" as a release from effort and criticism are frequently encountered when monopoly power rests on location and when mobility differs strongly from one group of local customers to another. If, as is likely, the mobile customers are those who are most sensitive to quality, their exit, caused by the poor performance of the local monopolist, permits him to persist in his comfortable mediocrity. This applies, for example, to small city or "ghetto" stores which lose their quality-conscious patrons to better stores elsewhere as well as to sluggish electric power utilities in developing countries whose more demanding customers will decide at some point that they can no longer afford the periodic breakdowns and will move out or install their own independent power supply.

The United States Post Office can serve as another

example of the lazy monopolist who thrives on the limited exit possibilities existing for its most fastidious and well-to-do customers. The availability of fast and reliable communications via telegraph and telephone makes the shortcomings of the mail service more tolerable; it also permits the Post Office to tyrannize the better over those of its customers who find exit to other communication modes impractical or too expensive.

Those who hold power in the lazy monopoly may actually have an interest in *creating* some limited opportunities for exit on the part of those whose voice might be uncomfortable. Here is a good illustration of the contrast between the profit-maximizing and the lazy monopolist: the former would engage, if he could, in discriminatory pricing so as to extract maximum revenue from its most avid customers, while the lazy monopolist would much rather price these customers out of the market entirely so as to be able to give up the strenuous and tiresome quest for excellence. For the most avid customers are not only willing to pay the highest prices, but are also likely to be most demanding and querulous, in case of any lowering of standards.[5]

Instances of such topsy-turvy (from the point of view of profit maximization) discrimination are not easy to document in economic life, in part perhaps because we have not looked for them very hard and in part simply because price discrimination in general is not easily practiced. But a closely analogous situation is familiar from politics. Latin American powerholders have long encouraged their political enemies and potential critics to remove themselves from the scene through voluntary exile. The

5. There is another way in which the lazy monopolist may be able to rid himself of the voice of these customers: he can extend *just to them* especially high-quality, "gold-plated" service. This would be discrimination with respect to quality rather than to price. The purpose, once again, is not to extract maximum revenue, but to buy "freedom to deteriorate."

right of asylum, so generously practiced by all Latin American republics, could almost be considered as a "conspiracy in restraint of voice." An even more straightforward illustration is supplied by a Colombian law that provided for paying former presidents as many U.S. dollars if they resided abroad as they would receive in Colombian pesos if they lived in their own country. With the U.S. dollar being worth from five to ten pesos while the law was in effect, the officially arranged incentive toward exit of these potential "trouble makers" was considerable.

Even without such special incentives, exit for disgruntled or defeated politicians has always been easier in some countries than in others. The following comparison between politics in Japan and in Latin America supplies another illustration of the corroding influence exit can have on vigorous and constructive political processes via voice:

> The isolation of Japan set rigid boundaries to the possibilities of political opposition. The absence of easy opportunities for tolerable exile was a powerful teacher of the virtues of compromise. The Argentinean newspaper editor in danger of arrest or assassination could slip across the river to Montevideo and still find himself a home, amid familiar sounds and faces and familiar books, easily able to find friends and a new job. (Nowadays, perhaps, he would arrange a refuge in one of the mushrooming international organizations beforehand.) But to all but a tiny fraction of Japanese only one place has ever been home.[6]

In this view, Japan gained an advantage from being a "no-exit" polity while the ever-beckoning opportunity to exit that was characteristic of Hispano-American societies contributed perhaps as much to the factionalism and *personalismo* typical of their politics as the Spanish national character, the *machismo* cult, and similar conventionally given reasons.

6. R. P. Dore, "Latin America and Japan Compared," in John J. Johnson, ed., *Continuity and Change in Latin America* (Stanford: Stanford University Press, 1964), p. 238.

6

On Spatial Duopoly and the Dynamics of Two-Party Systems

The situations which have been analyzed up to now have as their point of departure a clear-cut deterioration in the performance of a firm or organization. The exit and voice options are reactions to this deterioration and, under certain conditions, will arrest and reverse it. Consumers were portrayed as being more or less sensitive to a change in quality, but they *all* experienced the change as either positive or negative. This assumption can and will now be dropped. In this respect, quality and price are once again revealed as totally different phenomena: a decline in the price of a commodity is good news for *all* consumers just as a rise in price means a loss in real income for *all,* but one and the same change in quality may make the commodity more appreciated by some consumers while others find it less to their taste than before. This is also the case, of course, for shifts in the positions of political parties and other organizations.

When firms and organizations have this possibility of changing quality in such a way as to please some while displeasing others, the question arises as to the quality which they are most likely to select. The economist's answer is that the firm will select that point on the quality scale which will maximize its profits.[1] This routine reflex does not really solve our problem, however; for if a firm both loses and gains customers by a given quality change (while costs remain unchanged) the criterion of profit maximization may not yield a unique solution at all. Or suppose that the firm is a monopolist, which does not actually lose or gain customers as it varies the quality of its

1. For simplicity's sake, it may be assumed that the quality changes in question do not affect costs.

product, but causes, through such variations, happiness and unhappiness in different groups of its customers. To make such situations determinate, it is plausible to introduce another criterion: in addition to maximizing profits, the firm will tend to minimize discontent of its customers, for the highly rational purpose of earning goodwill or reducing hostility in the community of which it is a part.[2] With this criterion in operation, the firm is in general likely to select a point in the middle of the quality range along which its profits are maximized. Suppose we have two categories of customers of a monopolistic firm, the *A*-fiends who would deplore any shift along a linear scale from quality *A* to *B* and the *B*-fiends who would hail it. The discontent-minimizing firm is then likely to select the midpoint between *A* and *B*,[3] provided the intensity of discontent of both *A*- and *B*-fiends is identical. If the discontent of the *A*-fiends as quality moves away from *A* is far stronger and more vocal than the corresponding discontent of *B*-fiends, then the firm is likely to select a quality that is considerably closer to *A* than to *B*.

The concept of voice has just made its appearance and has made it possible to introduce determinacy into the problem of quality-selection by the firm. Instead of inter-

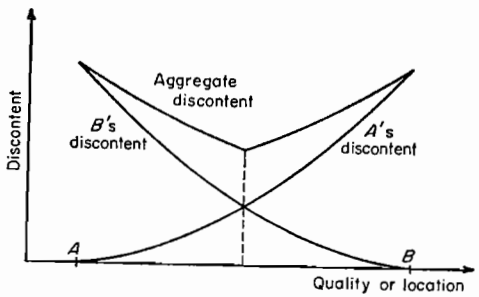
2. It is of course possible to equate this concern with profit maximization "in the long run."

3. If the frequency distribution of consumers' tastes has the normal shape, it is obvious that a discontent-minimizing firm will select the midpoint. Even when consumers' tastes are distributed with equal density along the *A-B* scale, discontent, when assumed to be proportional to the distance between actual and desired quality, would be minimized in the same fashion. This was shown long ago for the special case in which the *A-B* scale represents physical distance in a linear market (see n. 6, below). Location of the firm at some point of the scale then stands for "quality"; any change in this quality is obviously agreeable to some consumers and disagreeable to others and the cost of transportation inflicted by the firm's location on different consumers is the measure of their discontent (provided the marginal utility of money is constant). If a bimodal distribution of consumers' tastes is assumed, as was done in the text, a further condition must be imposed if it is to be concluded that the midpoint will be chosen. This is the plausible idea that discontent

preting the firm's decision to minimize discontent as a sovereign act on which it has decided out of enlightened self-interest,[4] it would have been possible and perhaps more realistic to say that, in selecting the middle of the quality range, the firm is simply responding to voice—or, rather, to customers' voices which have been assumed to be pulling the firm in opposite directions. But if voice plays so decisive a role when profit maximization provides no policy guidance to the firm, it will hardly be disregarded entirely when profit maximization points to some specific point along the quality range. In other words, the concern with voice (that is, with minimizing hostility and discontent) can be expected to qualify the concern with maximum profits. Should profit maximization conflict with

---

rises more than proportionately with deviation of the actual from the preferred quality. The discontent functions would then have the following shape:



The aggregate discontent of *A*-fiends and *B*-fiends is again minimized by selecting the midpoint between *A* and *B*.

4. Or out of pure altruism, as is assumed in Otto A. Davis and Melvin Hinch, "A Mathematical Model of Policy Formulation in a Democratic Society," in J. L. Bernd, ed., *Mathematical Applications in Political Science* (Dallas: Arnold Foundation, 1966), II, 175–208. At one point in their article, the authors examine, with results similar to those in the text, how a "beneficent dictator" would minimize the citizens' "utility loss function," that is, their discontent with the policies pursued by him.

discontent-minimization, there will be some compromise or trade-off between these two objectives.

A situation in which such a conflict and trade-off are particularly likely can be constructed as follows: Suppose that, of the two categories of consumers, *A*-fiends and *B*-fiends, the former have no alternative to turn to if quality of the firm's output moves toward *B*, while *B*-fiends' demand is so highly elastic that they will desert the firm in rapidly increasing numbers as quality shifts from *B* to *A*. In this situation a firm that single-mindedly maximizes profits will produce at point *B* of the quality range, while one that minimizes discontent would rather produce at point *A*. At *A*, the *A*-fiends would be happy while the *B*-fiends would all have taken their business elsewhere; they might not be kindly disposed toward the firm that has disappointed them, but they have cut themselves off through exit from all or much of the influence they might exert. In any event, the fact that they found a substitute so easily makes it likely that their welfare loss was not unduly high. If the firm produces at *B*, on the contrary, the *A*-fiends would still be with it, but presumably in a state of serious and vocal discontent. In this situation, a firm that is at all sensitive to voice will withdraw some distance from the point in the quality range at which it could achieve maximum profits. Note therefore that a firm is particularly likely to be deflected from the point of maximum profits when the consumers which are made unhappy by the firm producing at that point are in the position of having "nowhere else to go." This is a result that contradicts, or at least qualifies, the conventional ideas about the "powerful consumer." His power is usually believed to originate in the fact that he can take his business elsewhere and can thus "punish" the firm which does not pay heed to his preferences, but we see now that another kind of power resides in the consumer who *cannot* take his business elsewhere and who has therefore the maximum incentive to

cajole, threaten, and otherwise induce the firm to pay attention to his needs and tastes.[5]

The preceding discussion has a direct bearing on topics in economic and political thought which have a long and distinguished genealogy. Some forty years ago, Harold Hotelling published a celebrated article[6] which pioneered in a number of fields: duopoly, location theory, and the dynamics of two-party systems. His argument has been elaborated and qualified by later writers, but his basic points have not met with a direct challenge. Hotelling's principal idea can be summarized quite briefly. Customers or, in the political variant of the model, voters are assumed to be evenly distributed along a finite linear scale from *A* to *B*, or from Left to Right. Suppose that initially two firms (or two parties) have divided up this linear territory among themselves by locating at the midpoints of the left and right halves. From the social point of view, this is the ideal arrangement because it minimizes transportation costs for the consumers. In the political application of the model, the same result can be obtained: by locating at the quartiles, ideological distance between the voters and the parties and, hence, voters' discontent with the parties' platforms and policies will be minimized. Now assume that one of the two firms or parties, say, the one at the left-hand side, is allowed to shift its location without cost while the other is, or is thought to be, tied down. A profit-maximizing firm or a vote-maximizing party is

5. As long as quality change was defined as a deterioration that is felt as such by *all* consumers, exit and voice were pushing the firm in the same direction. If the firm mends its ways, the ensuing recovery will therefore be a "joint product" of exit and voice and it will be difficult to disentangle and evaluate the respective contributions of each. When quality change means improvement for some and deterioration for others, the comparative strength of the two mechanisms is more easily tested because they may work, as just explained, in opposite directions. I return to this point at the end of the present chapter.

6. "Stability in Competition," *Economic Journal*, 39:41–57 (1929).

under these conditions likely to move toward the right. The reason is that as long as it makes a point of staying to the left of the tied-down firm it retains a firm hold on its far-left customers or voters, while it can snatch new customers and voters away from the right-wing firm or party by advancing into its territory. Two important conclusions follow: (1) under the assumed conditions of duopoly there will be a tendency for the two firms to move toward the middle of the scale; (2) profit- or vote-maximizing behavior leads in this fashion to socially undesirable results since goods will be made available to consumers at higher total costs (if transportation costs borne by the consumer are counted) than would obtain if the firms had remained anchored at the quartiles. In a similar way it can be argued that it is probable, but socially undesirable, for parties in a two-party system to move ever closer together.[7]

The success which this elegant model has had, particularly among political scientists, is matched only by its

7. There is one important difference: after the political contest between the two parties is decided, the winning party takes over the government for the whole country while in the case of duopoly, the two firms permanently share the market among themselves. Thus, by locating at the quartiles, the parties would minimize the public's discontent with their positions and their policies, but not with those of the government that is the outcome of the struggle between the two parties. It can be argued, however, that a two-party system implies a preference for a risky but meaningful over a meaningless choice. Put somewhat differently, the average citizen may well prefer a situation in which a party with which he identifies closely has an even chance of beating one he sharply disagrees with, over a situation in which power is always held by a middle-of-the-road party which he neither likes nor dislikes strongly. This point is overlooked by Davis and Hinch, who view the possible location of the two candidates at the quartiles as a result of a nomination process in the course of which each party's nominee is selected exclusively by the members of that party. This is quite realistic in terms of the institutions of American democracy. But the result is not necessarily objectionable from the point of view of the community as a whole, as would seem to be implied by the Davis-Hinch analysis, which sets up the hypothetical policies pursued by a "beneficent dictator" as some sort of optimal solution. See n. 4, above.

failure to predict correctly the actual course of events—a fine illustration of the Streeten-Kuhn maxim that a model is never defeated by facts, however damaging, but only by another model.[8] Not that the model was left entirely untouched. When, in the wake of the Depression and the New Deal, the Democratic and Republican parties moved farther apart ideologically, an attempt was made to reconcile the model with these events. Advantage was taken of the fact that the results of the model depended critically —as had already been pointed out by Hotelling [9]—on the assumption of zero elasticity of demand throughout the linear market. With that assumption, consumers continue to buy the product from the nearest store no matter how far is "nearest" and citizens similarly go on voting for the party closest to them. If demand is elastic, on the other hand, a firm or party would lose customers or voters at its own end of the market as it moved toward the center and this loss of business or votes would at least restrain the socially undesirable clustering tendency of the original model.[10]

The pendulum of the facts swung back in the other direction in the fifties, with the soporific calm of the Eisenhower years and the somewhat premature announcement on the part of prominent scholars that ideology was dead.

8. Paul Streeten formulated this maxim in a letter to the author. The idea is persuasively developed in Thomas S. Kuhn's *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).

9. "Stability in Competition," p. 56.

10. Demand was assumed to be elastic throughout the linear market or to attain positive elasticity beyond a given range of transportation costs in the articles of Arthur Smithies ("Optimum Location in Spatial Competition," *Journal of Political Economy*, 49:423–439 [1941]) and of A. P. Lerner and H. W. Singer ("Some Notes on Duopoly and Spatial Competition," *Journal of Political Economy*, 45:145–186 [1939]), respectively. Smithies specifically presented his modification of the Hotelling model as a way of accounting for the strengthening of the ideological stances of the Democratic and Republican parties in the thirties, in contrast to the dilution of ideologies in the late twenties when Hotelling wrote.

In this atmosphere the Hotelling model was touched up once again. In a well-known work, Anthony Downs questioned the realism of the Hotelling assumption that voters are evenly distributed along the ideological spectrum, between the Left and Right.[11] If the frequency distribution of the voters along this scale has one peak toward the middle (the "middle-of-the-road") and tapers off toward both extremes, then Hotelling's clustering tendency will obviously assert itself once again. (It should be remarked that, in these conditions, the tendency would not cause the sort of social loss that it implies on the assumption of even distribution.) Thus Downs rehabilitated the Hotelling thesis, not by questioning the assumption of elastic demand through which the thesis had been qualified—to the contrary, he fully endorsed that qualification—but by counterbalancing elastic demand with the assumption of a more or less "normal" frequency distribution of the voters from Left to Right.[12]

As soon as the Hotelling model had been thus refurbished by Downs, its power to explain reality was again cast into doubt by the undisciplined vagaries of history. The selection by the Republican party of Goldwater in 1964 and, to a lesser extent, of Nixon in 1968 testified to the extreme reluctance of at least one party to conform to the Hotelling-Downs scenario. In general, evidence was increasing that the two parties were fairly consistently on opposite sides of many important issues.[13]

The concept of voice permits a more fundamental re-

11. *An Economic Theory of Democracy* (New York: Harper and Brothers, 1956), ch. 8.

12. Downs devoted much space to examining the results of other types of frequency distributions on two-party and multi-party systems. But in his discussion of two-party systems, he stresses the tendencies toward convergence and ambiguity of party positions and thus essentially bolstered the original Hotelling findings.

13. See S. M. Lipset, *Revolution and Counter-Revolution: Change and Resistance in Social Structures* (New York: Basic Books, 1968), p. 398, and literature there quoted (n. 27).

vision of the Hotelling model than was achieved in the thirties by introducing elastic demand. The fact is that Hotelling's original assumption of inelastic demand is perfectly realistic under conditions of a duopoly selling an essential good and of a well-established two-party system. It was not this assumption that was wrong or unrealistic, *but the inference that the "captive" consumer (or voter) who has "nowhere else to go" is the epitome of powerlessness.* True, he cannot exit to the other firm or party and in this way bring pressure on his own firm or party to improve its performance, but just because of that he, unlike the consumer or voter who can exit, will be maximally motivated to bring all sorts of potential influence into play so as to keep the firm or the party from doing things that are highly obnoxious to him. Hotelling's clustering tendency can therefore be countered and restrained not by substituting elastic for inelastic demand in his model, but by realizing that inelastic demand at the extremes of the linear market can spell considerable influence *via voice.*

As already outlined, voice will force the firm or the party to trade its profit-making or vote-getting objectives to some extent against the discontent-reducing objective. Such a trade-off becomes even more likely when the inevitable uncertainty about prospective sales or votes is taken into account. In other words, a party which is beleaguered by protests from disgruntled members because they dislike proposed "wishy-washy" platforms or policies will often be tempted to give in to these voices because they are very real here and now, while the benefits that are to accrue from wishy-washiness are highly conjectural.

The general conditions for the use of voice have been discussed in Chapter 2. With respect to the subject now under discussion, the matter can perhaps best be formulated as follows: for voice to function properly it is necessary that individuals possess reserves of political influence

which they can bring into play when they are sufficiently aroused. That this is generally so—that, in other words, there is considerable slack in political systems—is well recognized. "Nearly every citizen in the community has access to unused political resources" writes Robert Dahl.[14]

Clearly, Hotelling's concern about the social losses that might be caused by the clustering tendency was excessive. Those who are made unhappy by the party's wishy-washy position can influence the party through a mechanism that is none the less powerful for operating outside the market. On the other hand, there can be no guarantee that the voice mechanism will bring the party exactly back to the somewhat problematical "social optimum" which, in analogy to Hotelling's treatment of the location problem, can be defined as the point at which the sum of the ideological distances between the party and its clients is minimized. The influence of those who have nowhere else to go may well make the party overshoot that point, with disastrous consequences for its vote-gathering objectives. This was essentially what happened to the Republican party in 1964 with the selection of Goldwater as its presidential nominee.

Hardly ever was a hypothesis so cruelly contradicted by the facts as were the predictions of the Hotelling-Downs theory by the Goldwater nomination. Nevertheless, not even this event led to an outright questioning of the theory. In a searching article, three political scientists looked for the reasons for which the Republican party had so clearly failed to act as the vote-maximizer on that occasion.[15] They came close to the correct answer by focusing on the right-wing of the party and by showing that this element was far more activist than the middle-of-the-

14. *Who Governs?* (New Haven: Yale University Press, 1961), p. 309.

15. P. E. Converse, A. R. Clausen, and W. E. Miller, "Election Myths and Reality: The 1964 Election," *American Political Science Review,* 59:321–336 (June 1965).

roaders. The writing of letters to public officials or to newspapers and magazines was investigated as a particular type of intense political activity and it was shown that indeed this activity was engaged in to a wholly disproportionate extent by those right-wing Republicans who "had nowhere else to go." But the authors use these most interesting data primarily to explain the *misperceptions* of the Republican party and of their nominee with respect to the chances for victory, instead of drawing the following, much more basic conclusion: in a two-party system a party will not necessarily behave as the Hotelling-Downs vote-maximizer because those "who have nowhere else to go" are not powerless but influential.[16]

This power of those who have nowhere else to go in a two-party system has come to light in a different form with the Democratic defeat in the 1968 elections. The mobilization of the indifferent voters and the winning over of the undecided ones was seen to depend to a considerable extent on the enthusiasm which each of the parties can inspire among activist party workers and volunteers. Since the activists are far from being middle-of-the-roaders, their enthusiasm can be dampened by a party's moving to an excessively middle-of-the-road position. Hence the adoption of a platform which is designed to gain votes at the center can be counter-productive: it may damage rather than shore up the party's fortunes at the polls. With this mechanism the voice of those who have nowhere to go actually works "through the market" as it imposes

16. In the last paragraph of the article, something of this conclusion is in fact suggested by the authors: "The intense levels of political motivation which underlie the letter-writing of the ultra-conservative wing are part and parcel of the ingredients which led to a Republican convention delegation so markedly discrepant from either the rank-and-file of the Party or its customary leadership." But apart from this statement, the whole emphasis of the article is on the misperception of the party, rather than on the misjudgment of those who expect the party to conform to the Hotelling-Downs model.

decreasing and at some point negative returns on a move of the party to the center. It is as though those who are located at the end of a linear market were in charge of advertising the firm's products to those in the middle; naturally their advertising zeal is likely to go down as the firm moves its site farther and farther away from them.

In this sort of constellation traditional analysis would have no difficulty recognizing the limitations of the Hotelling-Downs model. The same goes for another qualification of the model which has already been mentioned: when sufficiently antagonized and outraged, the supposedly captive members may either "sit this one out" or even secede from the party and set up their own movement, however futile a gesture this may be. Here demand at the extreme would turn out to be elastic after all rather than totally inelastic and traditional concepts would account well enough for what is happening.[17] But the crux of the matter can now be sharply stated. These situations in which the supposedly powerless voters at the extreme manage to inflict actual vote losses on the party if it moves too far to the center are only special manifestations of the general influence and power that come with "having nowhere to go." In other words, that power exists and that influence will be brought to bear even without such direct and measurable effects on the party's vote (or the firm's profits). There are *a great many ways* in which customers, voters, and party members can impress their unhappiness on a firm or a party and make their managers highly uncom-

17. In line with the analyses of Lerner, Singer, and Smithies, in the articles cited in n. 10, above, Downs speaks in this connection of the "influence type of party" or of "blackmail parties" (*Economic Theory of Democracy*, pp. 131–132). Insofar as abstention is concerned, recent research shows that by far the principal influence on voter turnout has been the ease or cumbersomeness of the registration procedure rather than voter commitment to, or alienation from, individual candidates. See Stanley Kelley, Jr., R. E. Ayres, and W. G. Bowen, "Registration and Voting: Putting First Things First," *American Political Science Review*, 61:359–379 (June 1967).

fortable; only a few of these ways, and not necessarily the most important ones, will result in a loss of sales or votes, rather than in, say, loss of sleep by the managers.[18]

The situation which has been discussed here invites one further speculation. It has previously been pointed out that different organizations are differentially sensitive to voice and exit and that the optimal mix of voice and exit will therefore differ from one type of organization to another. For example, state enterprise which in case of a cash deficit due to revenue losses knows it can rely on the treasury is likely to be far more sensitive, at least up to a point, to voice (protests of consumers, appeals to higher authorities to replace existing management, and so forth) than to exit. This differential responsiveness has interesting consequences when the change in quality that gives rise to consumer reaction is felt as deterioration by some consumers while others sense the change as an improvement. Assume in addition that, as the quality moves in one direction, the organization exposes itself primarily to exit because the members antagonized by that move have an alternative organization to join while a move in the opposite direction will primarily activate voice of the antagonized, but "captive," consumers. It is then possible to predict the "quality path" of the firm or organization. Suppose small quality changes in the organization's performance occur constantly as a result of random events. If the organization responds more to voice than to exit, it is much more likely to correct deviations from normal quality that are obnoxious to its "captive" consumers; whereas deviations from quality that lead to exit of its noncaptive, exit-prone consumers would tend to go uncorrected for a considerable time.

Insofar as this situation approximates reality, it provides a rationale for the radicalization of political move-

18. See also Appendix A, last paragraph.

ment. The day-to-day policies of these movements tend to be influenced—specially when they are out of power—by their present activist members rather than by the preoccupation with losing the favor of all members and voters. Hence a shift toward the center which antagonizes the captive but activist members is likely to be resisted more strenuously than a radical shift, even though the latter might lead to exit of the noncaptive members and voters. One could conjecture that radicalization of political movements predicted by this model would assert itself the more strongly the longer the interval between elections; for electoral considerations can be expected to exert some restraining influence on the power of the captive party members. But this whole matter is further complicated by the phenomenon of organizational *loyalty*.

7

A Theory of Loyalty

As was pointed out in earlier chapters, the presence of the exit option can sharply reduce the probability that the voice option will be taken up widely and effectively. Exit was shown to drive out voice, in other words, and it began to look as though voice is likely to play an important role in organizations only on condition that exit is virtually ruled out. In a large number of organizations one of the two mechanisms is in fact wholly dominant: on the one hand, there is competitive business enterprise where performance maintenance relies heavily on exit and very little on voice; on the other, exit is ordinarily unthinkable, though not always wholly impossible, from such primordial human groupings as family, tribe, church, and state. The principal way for the individual member to register his dissatisfaction with the way things are going in these organizations is normally to make his voice heard in some fashion.[1]

As an aside, it is worth noting that, with exit either impossible or unthinkable, provision is generally made in these organizations for expelling or excommunicating the individual member in certain circumstances. Expulsion can be interpreted as an instrument—one of many—which "management" uses in these organizations to restrict resort to voice by members; a higher authority can then in turn restrict the powers of management by prohibiting expulsion, as is for example done to protect consumers when a public service is supplied in conditions of monopoly. But when exit is a wide-open option and voice is largely nonexistent, as in the relations between a firm and its customers in competitive markets, expulsion of a member or customer is a pointless affair and does not need to be specifically prohibited. One way of catching that somewhat rare bird, an organization where exit and voice both hold important roles, may be to look for groupings from which members can both exit and be expelled. Political parties and voluntary associations in general are excellent examples.

1. There is no intention here to associate absence of exit with "primitiveness." Edmund Leach has noted that many so-called primitive tribes are far from being closed societies. In his classic study *Political Systems of Highland Burma* (1954) he traced in detail the way in which members of one social system (*gumsha*) will periodically move to another (*gumlao*) and back again. Exit may be more effectively ruled out in a so-called advanced open society than among the tribes studied by Leach.

The Activation of Voice as a Function of Loyalty

A more solid understanding of the conditions favoring coexistence of exit and voice is gained by introducing the concept of *loyalty*. Clearly the presence of loyalty makes exit less likely, but does it, by the same token, give more scope to voice?

That the answer is in the positive can be made plausible by referring to the earlier discussion of voice. In Chapter 3 two principal determinants of the readiness to resort to voice when exit is possible were shown to be:

(1) the extent to which customer-members are willing to trade off the certainty of exit against the uncertainties of an improvement in the deteriorated product; and

(2) the estimate customer-members have of their ability to influence the organization.

Now the first factor is clearly related to that special attachment to an organization known as loyalty. Thus, even with a given estimate of one's influence, the likelihood of voice increases with the degree of loyalty. In addition, the two factors are far from independent. A member with a considerable attachment to a product or organization will often search for ways to make himself influential,

especially when the organization moves in what he believes is the wrong direction; conversely, a member who wields (or thinks he wields) considerable power in an organization and is therefore convinced that he can get it "back on the track" is likely to develop a strong affection for the organization in which he is powerful.[2]

As a rule, then, loyalty holds exit at bay and activates voice. It is true that, in the face of discontent with the way things are going in an organization, an individual member can remain loyal without being influential himself, but hardly without the expectation that *someone* will act or *something* will happen to improve matters. That paradigm of loyalty, "our country, right or wrong," surely makes no sense whatever if it were expected that "our" country were to continue forever to do nothing but wrong. Implicit in that phrase is the expectation that "our" country can be moved again in the right direction after doing some wrong—after all, it was preceded in Decatur's toast by "Our country! In her intercourse with foreign nations, may she always be in the right!" The possibility of influence is in fact cleverly intimated in the saying by the use of the possessive "our." This intimation of some influence and the expectation that, over a period of time, the right turns will more than balance the wrong ones, profoundly distinguishes loyalty from faith. A glance at Kierkegaard's celebrated interpretation of Abraham's setting out to sacrifice Isaac makes one realize that, in comparison

2. In terms of figure 3 of Appendix B, a person whose influence (that is, the likelihood that he will be able to achieve full quality recuperation) is correctly expressed by a point as high as V_3 will be willing to trade off the certainty of the competing product against even a little hope of recuperation for the traditional product. Thus he will choose voice. He who has little influence and knows it, on the other hand, is not likely to take kindly to such a trade-off. If he is to opt for voice rather than exit, he will normally require the certain availability of the competing product to be matched by the near-certainty of recuperation for the traditional variety.

to that act of pure faith, the most loyalist behavior retains an enormous dose of reasoned calculation.

When is loyalty functional?

The importance of loyalty from our point of view is that it can neutralize within certain limits the tendency of the most quality-conscious customers or members to be the first to exit. As has been shown in Chapter 4, this tendency deprives the faltering firm or organization of those who could best help it fight its shortcomings and its difficulties. As a result of loyalty, these potentially most influential customers and members will stay on longer than they would ordinarily, in the hope or, rather, reasoned expectation that improvement or reform can be achieved "from within." Thus loyalty, far from being irrational, can serve the socially useful purpose of preventing deterioration from becoming cumulative, as it so often does when there is no barrier to exit.

As just explained, the barrier to exit constituted by loyalty is of finite height—it can be compared to such barriers as protective tariffs. As infant industry tariffs have been justified by the need to give local industry a chance to become efficient, so a measure of loyalty to a firm or organization has the function of giving that firm or organization a chance to recuperate from a lapse in efficiency. Specific institutional barriers to exit can often be justified on the ground that they serve to stimulate voice in deteriorating, yet recuperable organizations which would be prematurely destroyed through free exit. This seems the most valid, though often not directly intended, reason for the complication of divorce procedures and for the expenditure of time, money, and nerves that they necessitate. Similarly the American labor law sets up a fairly complex and time-consuming procedure for one trade union to take

over from another as the sole certified bargaining agent at the plant level. Consequently, when workers are dissatisfied with the services of a union, they cannot switch easily and rapidly to another and are that much more likely to make an effort at revitalizing the union with which they are affiliated.

The previous discussion of the alternative between exit and voice makes it possible to say something about the conditions under which specific institutional barriers to exit, or, in their absence, the generalized, informal barrier of loyalty are particularly desirable or "functional." It was shown, for one, that in the choice between voice and exit, voice will often lose out, not necessarily because it would be less effective than exit, but because its effectiveness depends on the *discovery* of *new* ways of exerting influence and pressure toward recovery. However "easy" such a discovery may look in retrospect the chances for it are likely to be heavily discounted in *ex ante* estimates, for creativity always comes as a surprise. Loyalty then helps to redress the balance by raising the cost of exit. It thereby pushes men into the alternative, creativity-requiring course of action from which they would normally recoil and performs a function similar to the underestimate of the prospective task's difficulties. I have elsewhere described how such underestimates can act as a beneficial "Hiding Hand" in just this manner.[3] Loyalty or specific institutional barriers to exit are therefore particularly functional whenever the effective use of voice requires a great deal of social inventiveness while exit is an available, yet not wholly effective, option.

Secondly, the usefulness of loyalty depends on the closeness of the available substitute. When the outputs of two competing organizations are miles apart with respect to price or quality, there is much scope for voice to come into

3. *Development Projects Observed* (Washington: Brookings Institution, 1967), ch. 1.

play in the course of progressive deterioration of one of them before exit will assume massive proportions. Thus, loyalty is hardly needed here, whereas its role as a barrier to exit can be constructive when organizations are close substitutes so that a small deterioration of one of them will send customer-members scurrying to the other. This conclusion is a little unexpected. Expressed as a paradox, it asserts that loyalty is at its most functional when it looks most irrational, when loyalty means strong attachment to an organization that does not seem to warrant such attachment because it is so much like another one that is also available. Such seemingly irrational loyalties are often encountered, for example, in relation to clubs, football teams, and political parties. Even though it was argued in Chapter 6 that parties in a two-party system are less likely to move toward and resemble each other than has sometimes been predicted, the tendency does assert itself on occasion. The more this is so the more irrational and outright silly does stubborn party loyalty look; yet that is precisely when it is most useful. Loyalty to one's country, on the other hand, is something we could do without, since countries can ordinarily be considered to be well-differentiated products. Only as countries start to resemble each other because of the advances in communication and all-round modernization will the danger of premature and excessive exits arise, the "brain drain" being a current example. At that point, a measure of loyalty will stand us in good stead. Also, there are some countries that resemble each other a good deal because they share a common history, language, and culture; here again loyalty is needed more than in countries that stand more starkly alone as was precisely implied by the comparison between Latin America and Japan, which was cited above (Chapter 5).

Finally, what was said in Chapter 4 about the danger of losing influential customers when a higher-quality, higher-

price product is available "nearby," points to another conclusion on the comparative need for loyalty. If organizations can be ranked along a single scale in order of quality, prestige, or some other desirable characteristic, then those at the densely occupied lower end of the scale will need loyalty and cohesive ideology to a greater extent than those at the top. There is much evidence that this need is being appreciated both among various "left behind" groups of American society and, in the international arena, among the countries of the Third World. In the next chapter it will be shown that the most prestigious organizations and groups might, to the contrary, benefit from a decline in the level of loyalty they command.

The loyalist's threat of exit

Loyalty is a key concept in the battle between exit and voice not only because, as a result of it, members may be locked into their organizations a little longer and thus use the voice option with greater determination and resourcefulness than would otherwise be the case. It is helpful also because it implies the possibility of disloyalty, that is, exit. Just as it would be impossible to be good in a world without evil, so it makes no sense to speak of being loyal to a firm, a party, or an organization with an unbreakable monopoly. While loyalty postpones exit its very existence is predicated on the possibility of exit. That even the most loyal member can exit is often an important part of his bargaining power vis-à-vis the organization. The chances for voice to function effectively as a recuperation mechanism are appreciably strengthened if voice is backed up by the *threat of exit,* whether it is made openly or whether the possibility of exit is merely well understood to be an element in the situation by all concerned.

In the absence of feelings of loyalty, exit per se is essentially costless, except for the cost of gathering information about alternative products and organizations. Also, when loyalty is not present, the individual member is likely to have a low estimate of his influence on the organization, as already explained. Hence, the decision to exit will be taken and carried out in silence. The threat of exit will typically be made by the loyalist—that is, by the member who cares—who leaves no stone unturned before he resigns himself to the painful decision to withdraw or switch.

The relationship between voice and exit has now become more complex. So far it has been shown how easy availability of the exit option makes the recourse to voice less likely. Now it appears that the *effectiveness* of the voice mechanism is strengthened by the possibility of exit. The willingness to develop and use the voice mechanism is reduced by exit, but the ability to use it with effect is increased by it. Fortunately, the contradiction is not insoluble. Together, the two propositions merely spell out the conditions under which voice (a) will be resorted to and (b) bids fair to be effective: there should be the possibility of exit, but exit should not be too easy or too attractive as soon as deterioration of one's own organization sets in.

The correctness of this proposition can be illustrated by the extent to which parties are responsive to the voice of the membership. The parties of totalitarian one-party systems have been notoriously unresponsive—as have been the parties of multi-party systems. In the former case, the absence of the possibility of either voice or exit spelled absolute control of the party machinery by whatever leadership dominated the party. But in the second case, with both exit and voice freely available, internal democracy does not get much of a chance to develop either because, with many parties in the field, members will usually find it tempting to go over to some other party in case of disagreement. Thus they will not fight for "change from

within." In this connection it may be significant that Michels's "Iron Law of Oligarchy" according to which all parties (and other large-scale organizations) are invariably ruled by self-serving oligarchies was based on first-hand acquaintance primarily with the multi-party systems of Continental Western Europe. The best possible arrangement for the development of party responsiveness to the feelings of members may then be a system of just a very few parties, whose distance from each other is wide, but not unbridgeable. In this situation, exit remains possible, but the decision to exit will not be taken lightheartedly. Hence voice will be a frequent reaction to discontent with the way things are going and members will fight to make their voice effective. This prediction of our theory is confirmed by the lively internal struggles characteristic of parties in existing two-party systems, however far they may be from being truly democratic. Even in parties in nontotalitarian almost-one-party systems, as for example the Congress party of India and the PRI (Partido Revolucionario Institucional) of Mexico, voice has been more in evidence than in many of the often highly authoritarian or oligarchic parties of multi-party systems.*

In two-party systems, exit can happen not only as a result of a member or group of members of one party going over to the other, but because it is always possible to launch a third party. Hence, if voice is to be given a fair try by the members, such launching must not be too easy —a condition that is usually fulfilled by the very existence and tradition of the two-party system, as well as by the institutional obstacles ordinarily placed in the way of third parties. On the other hand, if voice is to be at its most effective, the threat of exit must be credible, particularly when it most counts. In American presidential politics this set of conditions for maximizing the effectiveness of voice means that a group of party members should be able to stay within the party up to the nominating convention and still be able to form a third party between the end of the convention and election time. If exit is made too difficult by requiring the group to qualify as a party at a date *prior* to the convention, the dissenting group must either exit before the convention or go to the convention without being able to make an effective threat of exit. More stringent conditions for exit fail here to strengthen voice; rather they make for either premature exit or for less effective voice. The point is well put by Alexander Bickel:

> The characteristic American third party . . . consists of a group of people who have tried to exert influence within one of the major parties, have failed, and later decide to work on the outside. States in which there is an early qualifying date tend to force such groups to forego major-party primary and other prenomination activity and organize separately, early in an election year. For if they do not they lose all opportunity for action as a third party later.[4]

The author adds that this is counterproductive from the point of view of the two-party system; the same judgment can be made from the point of view of achieving

* A related point of considerable importance is suggested to me by the recent article of Michael Walzer, "Corporate Authority and Civil Disobedience," *Dissent* (September–October 1969), pp. 396–406. The strict democratic controls to which supreme political authority is subjected in Western democracies are contrasted in the article with the frequently total absence of such controls in corporate bodies functioning within these same states. As the author shows, this absence or feebleness of voice in most commercial, industrial, professional, educational, and religious organizations is often justified by the argument that "if [their members] don't like it where they are, they can leave" (p. 397), something they cannot do in relation to the state itself. Walzer argues strongly that this argument is a poor excuse which should not be allowed to stand in the way of democratization; but as a matter of positive political science, it is useful to note that the greater the opportunities for exit, the easier it appears to be for organizations to resist, evade, and postpone the introduction of internal democracy even though they function in a democratic environment.

4. Alexander M. Bickel, "Is Electoral Reform the Answer?" *Commentary* (December 1968), p. 51.

party responsiveness to its members through the most effective mix of voice and exit.

Two conclusions stand out from this discussion: (1) the detail of institutional design can be of considerable importance for the balance of exit and voice; (2) this balance, in turn, can help account for the varying extent of internal democracy in organizations.

Boycott

Boycott is another phenomenon on the border line between voice and exit, just like the threat of exit. Through boycott, exit is actually consummated rather than just threatened; but it is undertaken for the specific and explicit purpose of achieving a change of policy on the part of the boycotted organization and is therefore a true hybrid of the two mechanisms. The threat of exit as an instrument of voice is here replaced by its mirror image, the promise of re-entry: for it is understood that the member-customer will return to the fold in case certain conditions which have led to the boycott are remedied.

Boycott is often a weapon of customers who do not have, at least at the time of the boycott, an alternative source of supply for the goods or services they are ordinarily buying from the boycotted firm or organization, but who can do temporarily without them. It is thus a temporary exit without corresponding entry elsewhere and is costly to both sides, much like a strike. In this respect also it combines characteristics of exit, which causes losses to the firm or organization, with those of voice, which is costly in time and money for the member-customers.

Elements for a model of loyalist behavior

It may be helpful to set up a more formal model of what happens when choice between two competing goods or organizations is affected by loyalty. For the purpose of

this inquiry, it will be assumed once again that the normally bought product or the organization to which one belongs begins to deteriorate. The focus will now be on organizations and their policies, rather than on firms and their products. Quality deterioration must therefore be redefined in subjective terms: from the member's viewpoint, it is equivalent to increasing disagreement with the organization's policies.

In figure 1 the horizontal axis measures quality of an

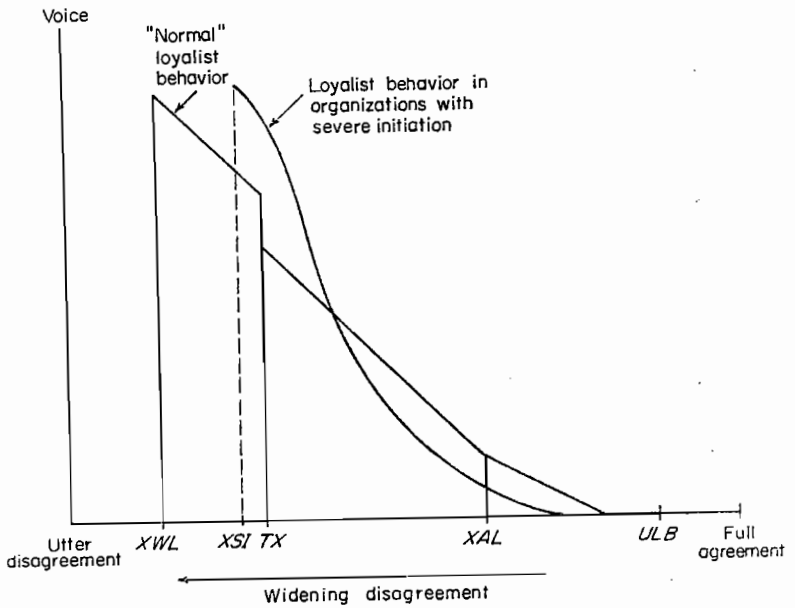


Figure 1. Loyalist behavior in the face of increasing disagreement with an organization

organization which is moving from the point where the member finds himself in complete agreement with its policies to the point of total disagreement. The vertical axis measures the amount of effective voice that is forthcoming in response to various degrees of disagreement.

At some point in the process of the organization's heading in the "wrong" direction, members will begin attempts to use their influence to correct and reverse the process, and these attempts will become stronger as disagreement widens. There comes a point in this process at which exit would take place in the absence of loyalty (*XAL*—point of e*X*it in the *A*bsence of *L*oyalty). Loyalty now acts as a brake on the decision to exit. The *loyal* member does *not* exit, *but something happens to him:* he begins to be acutely unhappy about continuing as a member, contracts qualms or *Bauchschmerzen* (bellyaches) as the phrase went among German Communist party members dissatisfied with the party line. Normally he will make stronger attempts than hitherto to change the line and will intensify the use of voice in its various forms for this purpose; hence we show a kink in the voice function at this point, and a steeper slope after it. Then, as disagreement widens further, the member will have thoughts of exit and threaten it (*TX*—point of *T*hreat of e*X*it) if that action can be at all expected to enhance the effectiveness of voice. The threat of exit means a discontinuous increase in the amount of voice that is forthcoming; this explains the vertical slope of the voice function at this point. Finally, loyalty reaches its breaking point and exit ensues (at point *XWL*—point of e*X*it *W*ith *L*oyalty). The strength of the grip which loyalty has on the customer or member can be measured either by the distance between *XAL* and *TX* or by that between *XAL* and *XWL*. These two distances define two different varieties of loyalty. The former represents loyalty with no thought of exit—in many basic organizations, exit is normally entirely outside the horizon of the member, even though he may be quite unhappy about his condition as member. The distance between *XAL* and *XWL* represents a more inclusive concept of loyalist behavior. The distance *TX-XWL* represents the portion of the process of deterioration during which the member thinks

about exit and is liable to use the threat of exit for the purpose of changing the policies of the organization. This threat being in some situations a particularly potent weapon, the total volume of effective voice that is generated in the course of the process of deterioration may be more closely related to that distance than to the total stretch of loyalist behavior *(XAL-XWL)*.

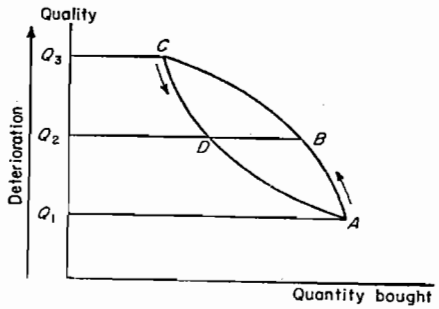
With the help of this model, speculation about the loyalist's behavior can be carried a little further. Suppose he has exited (exit from a product means ordinarily "entry" into a competing product, whereas exit from an organization can mean simply passage from the set of members to the set of nonmembers) and the product or organization he has left achieves recovery: At what point of the organization's "road back" will he re-enter? It seems quite unlikely that he will do so as soon as recovery reaches point *XWL* at which he exited. Just because he suffered between *XAL* and *XWL* he will now wait *at least* until the product or the organization has returned to point *XAL* at which previously he began to have qualms. He may very well require higher quality as an extra margin of insurance that renewed slippage will not immediately saddle him with *Bauchschmerzen* once again; in many cases, of course, the whole process may have left behind such scars that re-entry is altogether inconceivable. Thus the points of exit and re-entry will be far from identical; the distance between them, if it could be measured, would yield another way of measuring the strength of loyalty for different products and organizations.

If progressive deterioration and then improvement of quality in the above model is replaced by successive declines and then increases in the prices of assets, loyalist behavior is seen to be akin to that of the naïve, small, odd-lot investor who typically sells stocks cheap to stop his losses and buys back dear after stock values have risen considerably beyond those at which they were sold. Unlike

such investors, however, the loyalist is not necessarily a "sucker"; his sticking with the deteriorating product or organization should have as counterpart an increase in the chances of their recovery. It is only if such recovery fails to occur that he looks like, and turns out to be, a sucker. But in that case he has lost the bet on recovery that is implicit in loyalist behavior.

An observation of interest to the economist: loyalist behavior as sketched out here leads to a breakup of the traditional demand curve which establishes a one-to-one relationship between price (or quality) and quantity bought into two distinct curves. When a loyalty-commanding product first deteriorates and then improves, there will be one demand schedule for the downward movement in quality, with low demand elasticities at the beginning and high ones eventually as intolerable deterioration finally does lead to exit of the loyalists, and quite another one as quality recovers. During the improvement phase, elasticities will be low in the low-quality ranges and will only eventually become higher as improvement is confirmed.[5]

5. This proposition is easily diagrammed. The figure below shows quantity bought on the horizontal axis and quality (deterioration) on the vertical axis. Suppose quality first stands at Q_1, then deteriorates gradually to Q_3 and thereafter recovers slowly back to Q_1. Curve *ABC* then shows the demand schedule for the deterioration phase while curve *CDA* portrays demand for the recovery phase. Depending on the phase of the decline-recovery cycle, demand for quality Q_2 is either Q_2B or Q_2D.



Demand is of course always likely to be a function not only of current, but to some extent also of previous, quality because of inertia and lags in perception. Loyalty strongly reinforces this influence of past performance of the firm or organization on present behavior of the customers or members.

These remarks make it tempting to introduce the concept of *unconscious* loyalist behavior. A situation similar to the one in which the points of exit and of re-entry do not coincide has been described by psychologists. If, say, the likeness of a cat is made to change gradually into that of a dog through a succession of images shown to a subject and if later the same series is shown in reverse order, the eye behaves as though it were "loyal" to whatever figure it started with: when the sequence is shown in the cat to dog direction, a majority of images will be labeled "cat," and vice versa.[6] To this extent then, the general difficulties of recognizing change are a breeding ground for unconscious loyalist behavior in case of deterioration, as well as for prolonged reluctance toward entry or re-entry in case the organization improves.[7] Since unconscious loyalist behavior is by definition free from felt discontent, it will not lead to voice. This behavior whose onset is marked by point

6. K. R. L. Hall, "Perceiving and Naming a Series of Figures," *Quarterly Journal of Experimental Psychology*, 2:153–162 (1950). Similar results have been obtained in experiments designed to investigate how diverse bits and pieces of information are combined and integrated. When, for example, several personality trait adjectives are read to the subjects of the experiment, the over-all judgment about the person described by the adjectives depends on the order in which the adjectives have been named, with the earlier-named ones apparently receiving a higher weight. For instance, the sequence "intelligent, prudent, moody, self-centered" produces a better over-all impression than the reverse sequence. This phenomenon is known as "primacy effect." See Norman H. Anderson, "Primacy Effects in Personality Impression Formation," *Journal of Social Psychology*, 2:1–9 (June 1965), and literature there noted.

7. Robert Jervis, "Hypotheses on Misperception," *World Politics*, 20:439–453 (April 1968), and Albert O. Hirschman, "Underdevelopment, Obstacles to the Perception of Change, and Leadership," *Daedalus* (Summer 1968), pp. 925–936.

ULB (*U*nconscious *L*oyal *B*ehavior) is loyalist only from the point of view of an outside observer who feels that voice- or exit-justifying deterioration has indeed set in. The member is simply unaware of the degree of deterioration that is taking place.

The model which has been outlined will be useful in considering now certain variants of loyalist behavior.

Loyalist Behavior as Modified by Severe Initiation and High Penalties for Exit

Loyalty has so far been hailed as a force which, in the act of postponing exit, strengthens voice and may thus save firms and organizations from the dangers of excessive or premature exit. Something has already been said, however, about situations in which loyalty does not play so providential a role. The various institutions designed to foster loyalty have obviously not been established with the purpose of elaborating an improved mixture of voice and exit; when they do so, it is unwittingly, "as a result of human action, not of human design." [8]

It is always pleasant for the social scientist to discover such hidden and unintended harmonies, but the discovery carries with it an obligation to look out for situations that fall short of harmony. In the present case, the opportunities for a nonoptimal outcome are numerous. It is possible for loyalty to overshoot the mark and thus to produce an exit-voice mix in which the exit option is unduly neglected. Secondly, it must be realized that loyalty-promoting institutions and devices are not only uninterested in stimulating voice at the expense of exit: indeed they are often meant to *repress* voice alongside exit. While feedback through exit or voice is in the long-run interest of organization managers, their short-run interest is to entrench themselves and to enhance their freedom to act as they wish, unmolested as far as possible by either desertions *or* complaints of members. Hence management can be relied on to think of a variety of institutional devices aiming at anything but the combination of exit and voice which may be ideal from the point of view of society.

8. This phrase, used by F. A. Hayek as the title of an essay in *Studies in Philosophy, Politics, and Economics* (Chicago: University of Chicago Press, 1967), is traced by him to Adam Ferguson's *Essay on the History of Civil Society* (1767).

High fees for entering an organization and stiff penalties for exit are among the main devices generating or reinforcing loyalty in such a way as to repress either exit or voice or both. How do these devices affect our model of loyalist behavior? The concept of unconscious loyalist behavior can serve to open up the subject. As was just shown, this type of behavior cannot give rise to voice; and because like all loyal behavior it also postpones exit, it will be prized by organizations whose management wishes members to refrain from both exit and voice. Such organizations will be looking for devices converting, as it were, conscious into unconscious loyalist behavior.

Actually there often is no clear dividing line between these two types of behavior, because the customer or member of the organization may have a considerable stake in *self-deception*, that is, in fighting the realization that the organization he belongs to or the product he has bought are deteriorating or defective. He will particularly tend to repress this sort of awareness if he has invested a great deal in his purchase or membership. In organizations entry into which is expensive or requires severe initiation, recognition by members of any deterioration will therefore be delayed and so will be the onset of voice. By the same token, however, it may be expected that once deterioration is adverted to, members of an organization that requires severe initiation will fight hard to prove that they were right after all in paying that high entrance fee. Thus while the onset of voice will be delayed by severe initiation, resort to it is likely to be *more active* than is ordinarily the case during a subsequent phase of loyalist

behavior. The high cost of entry will change the time-pattern of voice, but may well not reduce its aggregate volume.[9]

This finding implies a modification of the theory of cognitive dissonance. The theory has normally shown how people will alter their cognitions and beliefs so as to make them more consistent with some "discrepant" act or behavior they have engaged in and which is difficult to reconcile with these beliefs. In the case just noted, the act is more or less severe initiation and the cognition, in one well-known experiment, was the boring nature of the activities of the organization of which one has become a member. The theory predicted—and the experiment confirmed—that the severer the initiation the higher will be the degree of self-deception, that is, the more fascinating will the boring activities seem to the member.[10] Assume now, that there is not only some limit to self-deception but, and this is more important, room for *making* the activities of the organization more interesting as a result of members' initiative: then the same basic experimental constellation would lead to the prediction that severe-initiation members will display *more initiative* and will be *more activist* than the rest after having at first been more complacent and passive. Hence, a situation of dissonance may produce not only alterations of beliefs, attitudes, and cognitions, but could lead to *actions* designed to change the real world when that is an alternative way (and particularly when it is the only way) of overcoming or reducing dissonance.[11]

9. As is shown by the curved line in figure 1.

10. E. Aronson and J. Mills, "The Effects of Severity of Initiation on Liking for a Group," *Journal of Abnormal and Social Psychology*, 59:177–181 (1959). See also, for further refinement of the experimental results of Aronson-Mills and rebuttal of some criticisms, H. B. Gerard and G. C. Mathewson, "The Effects of Severity of Initiation on Liking for a Group: A Replication," *Journal of Experimental Social Psychology*, 2:278–287 (July 1966). See Appendix E for a fuller statement on these papers.

11. In spite of superficial resemblance, the hypothesis here proposed is fundamentally different from the one put forward and

This hypothesis is to be tested experimentally by Professor Philip Zimbardo of Stanford University and his associates.[12] Pending the outcome of these efforts, it is perhaps permissible to appeal to scattered historical evidence for illustration. Take the well-known and well-tested maxim that "revolution, like Saturn, devours its own children." Why this should be so is now easily understood: in "making the revolution" revolutionaries have paid a high personal price in risk-taking, sacrifice, and single-minded commitment. Once the revolution *is* made, a gap between the actual and the expected state of affairs is only too likely to arise. To eliminate that gap those who have paid the highest price for bringing about the new reality will be most strongly motivated to change it *anew*. In the process, they will take on some of their fellow revolutionaries who are now in positions of authority and a large number of the revolutionaries on either the one side or the other or on both will come to grief in the ensuing fight.

Another illustration of the same principle, drawn from the American experience, will be given in Chapter 8.[13]

tested in *When Prophecy Fails* by Leon Festinger, H. W. Riecken, and Stanley Schachter (Minneapolis: University of Minnesota Press, 1956). In this classic of the literature of cognitive dissonance, the authors investigated the effects on a group of believers of an unequivocal disconfirmation of their belief. In line with the theory's predictions, the believers became more vigorously engaged in proselyting activities than before. This activity, however, must be interpreted as an attempt to eliminate dissonance by "forgetting" the disconfirmation, *by drowning out the dissonant cognition*, rather than by changing it. Both the Aronson-Mills and the *Prophecy* situations are so constructed that the dissonant cognitions (boring nature of the activities of the group, nonoccurence of predicted flood) are unchangeable, once-and-for-all events. In the real world, many situations are of course iterative and are subject to change, "the next time around."

12. See Appendix E for a detailed statement on the scope and design of the proposed research.

13. See pp. 113–114. I argued elsewhere in a similar vein that efforts to rescue development projects from difficulty will be most vigorous when those responsible for the project are fully committed to it as a result of prior expenditures. Hence the later the difficulty appears the better, provided of course that it can be successfully solved. See Hirschman, *Development Projects Observed*, pp. 18–21.

Payment of a high price of entry thus does not lead necessarily to acquiescence with that for which the price has been paid, but may result in an even more determined and outspoken use of voice. It is also possible, of course, that by the time the member is no longer able to close his eyes to what is going on, deterioration has become such that exit appears as the only possible reaction to the sudden revelation of rottenness. Hence severe initiation may eventually activate exit as well as voice.[14] "You can actively flee and you can actively stay put"—this phrase of Erik Erikson is again most pertinent. It was quoted once before, in connection with the likely behavior of the quality-conscious consumer. The coincidence is not accidental, for severe initiation no doubt makes for quality-consciousness.

A different kind of distortion of the model of loyalist behavior occurs when an organization is able to exact a *high price for exit* (over and above the forfeit of the price for entry which occurs inevitably with exit). Such a price can range from loss of life-long associations to loss of life, with such intermediate penalties as excommunication, defamation, and deprivation of livelihood. Organizations able to exact these high penalties for exit are the most traditional human groups, such as the family, the tribe, the religious community, and the nation, as well as such more modern inventions as the gang and the totalitarian party.[15] If an organization has the ability to exact a high price for exit, it thereby acquires a powerful defense against one of the member's most potent weapons: the threat of exit. Obviously, if exit is followed by severe sanctions the very idea of exit is going to be repressed

14. The activation of exit is shown in figure 1 through the location of point *XSI* (eXit of members having received *S*evere *I*nitiation) ahead of *XWL*.

15. For an account of the terror of leaving the Communist party, see Gabriel A. Almond, *The Appeals of Communism* (Princeton, 1954), ch. 12.

and the threat will not be uttered for fear that the sanction will apply to the threat as well as to the act itself. In terms of the model, point *TX* will be moved to the left and is in fact likely to disappear altogether, that is, merge with *XWL*, the point of exit when loyalty is present. This point itself may of course also be moved to the left: to deter exit is indeed a major purpose of imposing a high price for it. But in comparison with organizations that can command strong spontaneous loyalty while being unwilling or unable to impose stiff penalties for exit, the main change in members' behavior under conditions of progressive deterioration of the organization is likely to be the omission of the threat of exit rather than the postponement of exit itself.

What happens to voice in organizations where the price of exit is high? Some tentative suggestions can be advanced by distinguishing between those high-exit-price organizations where the price of entry is zero (because, as in the case of the family or nation, one enters them as a result of one's birth) and those where this price is high as well. For the latter organizations it has just been shown that the onset of felt discontent and therefore of voice will be delayed. Since the high price of exit does away, on the other hand, with the threat of exit as an effective instrument of voice, these organizations (gangs, totalitarian parties) will often be able to repress both voice and exit. In the process, they will largely deprive themselves of both recuperation mechanisms.[16]

The situation is quite different for the traditional groups, such as family and nation, which exact a high price for exit, but not for entry. Here the fact that one fully "belongs" by birthright may nurture voice and thus

16. This is a special case of the proposition, put forward by David Apter, that any increase of coercion in a society will have a price in terms of the flow of information to the powerholders. See his *Politics of Modernization* (Chicago: University of Chicago Press, 1965), p. 40.

compensate for the virtual unavailability of the threat of exit. By itself, the high price or the "unthinkability" of exit may not only fail to repress voice but may stimulate it. It is perhaps for this reason that the traditional groups which repress exit alone have proved to be far more viable than those which impose a high price for both entry and exit.

Loyalty and the Difficult Exit from "Public Goods" (and Evils)

The reluctance to exit in spite of disagreement with the organization of which one is a member is the hallmark of loyalist behavior. When loyalty is present exit abruptly changes character: the applauded rational behavior of the alert consumer shifting to a better buy becomes disgraceful defection, desertion, and treason.

Loyalist behavior, as examined thus far, can be understood in terms of a generalized concept of penalty for exit. The penalty may be directly imposed, but in most cases it is internalized. The individual feels that leaving a certain group carries a high price with it, even though no specific sanction is imposed by the group. In both cases, the decision to remain a member and not to exit in the face of a superior alternative would thus appear to follow from a perfectly rational balancing of prospective private benefits against private costs. Loyalist behavior may, however, be motivated in a less conventional way. In deciding whether the time has come to leave an organization, members, *especially the more influential ones,* will sometimes be held back not so much by the moral and material sufferings they would themselves have to go through as a result of exit, but by the anticipation that the *organization to which they belong would go from bad to worse if they left.*

This sort of behavior is the opposite of the one discussed

in Chapter 4. It was shown there that under certain conditions the most influential members might be the first to exit. The reason for which this conclusion is reversed here is that a wholly new and somewhat strange assumption has just been introduced: the member continues to care about the activity and "output" of the organization *even after he has left it.* In most consumer-product and in many member-organization relations this is of course not the case. If I become dissatisfied with the brand of soap I usually buy, and consider switching to another, I do not expect such switching to cause a worsening of the quality of my habitual brand; even if I did I presumably would not care as long as I quit buying it.[17] With the help of this counter-example, we can spell out the two conditions that underlie the special loyalist behavior now under discussion:

In the first place, exit of a member leads to further deterioration in the quality of the organization's output; secondly, the member cares about this deterioration *whether or not he stays on as a member.*

The first condition means that quality of a product is not invariant to the number of buyers or to the amount sold. The withdrawal of some members leads to lower quality, hence presumably still lower "demand" from the remaining members and so on—a typical case of unstable equilibrium, and of a cumulative sequence à la Myrdal. The consumer-member is here a "quality-maker" rather than, as in perfect competition, a quality-taker. Situations in which individual buyers are conscious of being price-makers rather than price-takers are, of course, familiar from the theories of monopoly and monopolistic competition. What strikes the economist as weird here is the *direction* of the relationship: In the usual price-making

17. I may, in fact, entertain the opposite "serves-them-right" reaction if I hear that a firm which has disappointed me and with which I have stopped doing business comes to grief.

situation, withdrawal of a buyer (a downward shift of the demand curve) will lead to price being lowered or, correspondingly, to quality being *improved* because the supply curve is assumed to be rising. In the present case, on the contrary, withdrawal of the quality-making "buyer" leads to a quality decline. The reason is that the "buyer" is now in reality a member and as such he is involved in both the supply and the demand sides, in both production and consumption of the organization's output. Hence, if those who have the greatest influence on quality of output are also, as is likely, more quality-conscious than the rest of the members, any slight deterioration in quality may set off their exit, which in turn will lead to further deterioration, which will lead to further exits, and so on.

In this situation, utter instability is once again avoided by the intervention of loyalist behavior and particularly by members being aware of, and recoiling from, the prospective consequences of their exit. In other words, instability may be averted if members are aware that it threatens. But there is a real question why a member should care about the consequences of his exit on the quality of the organization, to the point where the prospective decline in quality would keep him from exiting. The only rational basis for such behavior is a situation in which the output or quality of the organization *matters to one even after exit.* In other words, *full exit is impossible;* in some sense, one remains a consumer of the article in spite of the decision not to buy it any longer, and a member of the organization in spite of formal exit.

This important class of situations can again be illustrated by the competition between private and public schools. Parents who plan to shift their children from public to private school may thereby contribute to a further deterioration of public education. If they realize this prospective effect of their decision they may end up by not taking it, for reasons of general welfare or even as a

result of a private cost-benefit calculation: the lives of both parents and children will be affected by the quality of public education in their community, and if this quality deteriorates the higher educational attainments of the children to be obtained by shifting them to private school have a cost which could be so large as to counsel against the shift.

The distinction made by economists between private and public (or collective) goods is directly relevant to this discussion. *Public goods* are defined as goods which are consumed by all those who are members of a given community, country, or geographical area in such a manner that consumption or use by one member does not detract from consumption or use by another. Standard examples have been crime prevention and national defense as well as other accomplishments of public policies that are or ought to be enjoyed by everyone such as high international prestige or advanced standards of literacy and public health. The distinguishing characteristic of these goods is not only that they *can* be consumed by everyone, but that there is *no escape* from consuming them unless one were to leave the community by which they are provided. Thus he who says public goods says public evils. The latter result not only from universally sensed inadequacies in the supply of public goods, but from the fact that what is a public good for some—say, a plentiful supply of police dogs and atomic bombs—may well be judged a public evil by others in the same community. It is also quite easy to conceive of a public good turning into a public evil, for example, if a country's foreign and military policies develop in such a way that their "output" changes from international prestige into international disrepute. In view of this book's concern with deterioration and resulting exit or voice, this sort of possibility is of special interest.

The concept of public goods makes it easy to understand the notion that in some situations there can be no real exit from a good or an organization so that the

decision to exit in the partial sense in which this may be possible must take into account any further deterioration in the good that may result. What becomes difficult to grasp, in fact, once the concept of public goods is introduced is how even a partial exit from such goods is possible.

Actually, of course, a private citizen can "get out" from public education by sending his children to private school, but at the same time he *cannot* get out, in the sense that his and his children's life will be affected by the quality of public education. There are many ostensibly private goods of this sort that one can buy or refrain from buying; but they have a "public-good dimension" (often called "externalities" by economists) so that their mere production and consumption by others affects, ennobles, or degrades the lives of all members of the community. While this is perhaps not a very frequent or very important phenomenon for saleable commodities and services, it is a central feature of many organizations in relation to their members. If I disagree with an organization, say, a political party, I can resign as a member, but generally I cannot stop being a member of the society in which the objectionable party functions. If I participate in the making of a foreign policy of which I have come to disapprove, I can resign my official policy-making position, but cannot stop being unhappy as a citizen of a country which carries on what seems to me an increasingly disastrous foreign policy. In both these examples, the individual is at first both producer and consumer of such public goods as party policy and foreign policy; he can stop being producer, but cannot stop being consumer.

It is thus possible to rationalize a wholly new type of loyalist behavior. In line with common sense (and the theory of demand), the propensity to exit has thus far been presented as a rising function of discontent with product quality, or of disagreement with the party line.

Now it can be shown that an invariant or even inverse relationship between these variables is possible. In the case of public goods, the member will compare, at any one point in the process of deterioration, the disutility, discomfort, and shame of remaining a member to the prospective damage which would be inflicted on him as a prospective nonmember and on society at large by the additional deterioration that would occur if he were to get out. The avoidance of this hypothetical damage is now the benefit of loyalist behavior, and if this benefit increases along with the cost of remaining a member, the motivation to exit need *not* become stronger as deterioration proceeds although undoubtedly our member will become increasingly unhappy. The ultimate in unhappiness and paradoxical loyalist behavior occurs when the public evil produced by the organization promises to accelerate or to reach some intolerable level as the organization deteriorates; then, in line with the reasoning just presented, the decision to exit will become ever more difficult the longer one fails to exit. The conviction that one has to stay on to prevent the worst grows stronger all the time.

Usually this sort of reasoning is an ex-post (or ex-nunc) justification of opportunism. But it must be reluctantly admitted that loyalist behavior of this type—the worse it gets the less can I afford to leave—can serve an all-important purpose when an organization is capable of dispensing public evils of truly ultimate proportions, a situation particularly characteristic of the more powerful states on the present world scene. The more wrongheaded and dangerous the course of these states the more we need *a measure of spinelessness* among the more enlightened policy makers so that some of them will still be "inside" and influential when that potentially disastrous crisis breaks out. It will be argued later that in these situations we are likely to suffer from an excess rather than from a shortage of spinelessness. It is nevertheless worth noting

that the magnitude of public evils that can today be visited upon all of us by the centers of world power has bestowed "functionality" or social usefulness on protracted spinelessness (failure to exit) provided it turns into spine (voice) at the decisive moment.

Organizations and firms producing public goods or public evils constitute the environment in which loyalist behavior (that is, postponement of exit in spite of dissatisfaction and qualms) peculiarly thrives and assumes several distinctive characteristics. For one, there is the possibility described in the last paragraphs in which we saw "right or wrong, my country" change into a seemingly perverse "the wronger the myer." Moreover, when exit does occur its nature is different from the type of exit discussed up to now. In the case of exit from organizations producing private goods, exit terminates the relationship between the customer-member and the product-organization he is leaving. True, by signaling to management that something is wrong, exit may provide a stimulus toward quality recuperation, but this effect is wholly unintended by the exiting customer-member—he "couldn't care less." In the case of public goods, on the other hand, one continues to "care" as it is impossible to get away from them entirely. In spite of exit one remains a consumer of the output or at least of its external effects from which there is no escape. Under these conditions, the customer-member will *himself* be interested in making his exit contribute to improvement of the product-organization he is leaving—an improvement which he may judge to be impossible without radical change in the way in which the organization is run. To exit will now mean to resign under protest and, in general, to denounce and fight the organization from without instead of working for change from within. In other words, the alternative is now not so much between voice and exit as between voice from within and voice from without (after exit). The exit decision then

hinges on a totally new question: At what point is one more effective (besides being more at peace with oneself) fighting mistaken policies from without than continuing the attempt to change these policies from within?

The considerable difference between "proper" exit from public goods and the kind of exit (from private goods) thus far discussed is revealed when a customer-member who exits from a public good behaves *as though* he were exiting from a private one. In a society as dominated by private goods and by styles of behavior acquired in reacting to them as the United States, such confusion may perhaps be expected. Examples from recent history come easily to mind. High officials who disagree with public policies do not blast them when they resign, but present this decision as a purely private one; one leaves because a better offer has come his way, "in fairness to my family." Similarly young men and women who find American society, its values, and the actions of its government not to their tastes are "opting out" as though they could secure for themselves a better set of values and policies without having first changed the existing set. The malaise resulting from this confusion of the two kinds of exit can be measured by the relief that *would* have been experienced if at least one of the public officials "dropping out" of the Johnson administration in disagreement over Vietnam had thereupon publicly fought official war policies; and by the relief that *was* so widely felt when the 1968 campaign of Senator Eugene McCarthy made it possible for many young Americans to do just that, instead of merely "copping out."