10

COST CURVES AND SUPPLY CURVES *

By Jacob Viner ||

It is the primary purpose of this article to develop a graphical exposition of the manner in which supply curves are dependent upon the different possible types of technological and pecuniary cost situations, under the usual assumptions of atomistic competition and of rational economic behavior on the part of the producers. No attempt is made here at realistic description of the actual types of relationship between costs and supply, and the purpose is the more modest one of presenting the formal types of relationship which can be conceived to exist under certain simplifying assumptions. Analysis of this kind derives obviously from the path-breaking contribution of Alfred Marshall in his Principles of Economics. Interest in this type of problem has been largely confined to the Anglo-Saxon countries, and in these countries there has been a tendency until recent years for economists to accept and reproduce the general lines of Marshall's analysis somewhat uncritically and without much further elaboration. I have no very serious fundamental criticism to make of Marshall's analysis of the supply side of the exchange value problem. But Marshall's treatment is highly elliptical. A striking illustration of his tendency to telescope his argument is his common practice in his graphs of labelling cost curves and supply curves alike with the symbols *ss*, conventionally used for supply curves, and thus diverting the attention of his readers, and perhaps also occasionally his own attention, from the necessity of selecting from among the many possible types of cost curve that one which in the given circumstances alone has claims to being considered as also a supply curve. Marshall, moreover, although he made valuable additions to the conceptual

* *Zeitschrift für Nationalökonomie*, Vol. III (1931), pp. 23–46. Reprinted, by the courtesy of the publisher and the author, without change from the original text, with the addition of a supplementary note.

|| Princeton University.

terminology necessary for analysis of this type, nevertheless worked with vocabulary lacking sufficient terms to distinguish clearly from each other all the significant types of cost phenomena, and here also the terminological poverty tended to lead to inadequate classification not only on the part of his followers but on his own part. Marshall's analysis was excessively simple even on the basis of his own simplifying assumptions, and inadequately precise in formulation, and his followers have standardized an even simpler type of exposition of the relationship of cost to price.

In recent years a number of English economists, notably Pigou, Sraffa, Shove, Harrod and Robertson, have presented in the Economic Journal a series of criticisms, elaborations, and refinements of the Marshallian analysis which, in my opinion, go a long way both towards bringing out clearly the contribution contained in its implications as well as in its explicit formulations, and towards completing and correcting it where that is necessary. The indebtedness of the present paper to their writings is considerable and is freely acknowledged. But I have been presenting charts such as those contained in this article to my students at the University of Chicago for a long period antedating the writings referred to above, and if in the course of years these charts have undergone substantial revision and, as I am convinced, correction, chief credit is due to the penetrating criticisms of my students.

The analysis which follows is based on the usual assumptions and presuppositions of the Marshallian type of economics. As compared to the Lausanne School type of analysis, it contents itself with examination of the conditions of a partial equilibrium of a special sort, and does not inquire into the repercussions of the postulated changes in cost or demand conditions on the general equilibrium situation. Like all partial equilibrium analysis, including the allegedly "general" equilibrium theories of the Lausanne School, it rests on assumptions of the *caeteris paribus* order which posit independence where in fact there is some degree of dependence. For such logically invalid assumptions there is the pragmatic defense that they permit of more detailed analysis of certain phases of economic interdependence than would be possible in their absence, and that to the extent that they are fictions uncompensated by counterbalancing fictions, it is reasonable to believe that the errors in the results obtained will be almost

invariably quantitative rather than qualitative in character, and will generally be even quantitatively of minor importance. As compared to the Austrian School, there is, I believe, no need either for reconciliation or for apology. On the somewhat superficial level on which analysis of the present type is conducted the basic issue as between the English and the Austrian Schools does not enter explicitly into the picture, and in so far as it has any bearing on the conclusions, this bearing is again quantitative rather than qualitative in character. The Austrian School starts with the assumption, usually tacit, never emphasized, that the supplies of all the elementary factors of production are given and independent of their rates of remuneration. The English School emphasizes, perhaps overemphasizes, the dependence of the amounts of certain of the elementary factors, notably labor and waiting, on their rates of remuneration. The techniques of analysis of each school are in essentials identical, and each school, if it were to apply its techniques to the situation postulated by the other, would reach identical conclusions. The difference in the assumptions of the two schools has bearing on the quantitative but not on the qualitative behavior of the prices of the elementary factors and therefore also of the money costs of their products, as the demands for these factors and products change. The conflict between the two schools has greater significance for the theory of the value of the elementary factors of production, i.e., for the theory of distribution, than for the theory of particular commodity price determination. For the present analysis, where it is assumed either that the prices of the elementary factors remain unaltered or that they undergo changes of a kind consistent with the basic assumptions of either school, the differences between the two schools would not affect qualitatively the character of the findings. All of the propositions laid down in this paper should, I believe, be acceptable to, or else should be rejected by, both schools.

The procedure which will be followed, will be to begin in each case with the mode of adjustment of a particular concern to the given market situation when the industry as a whole is supposed to be in stable equilibrium. This particular concern is not to be regarded as having any close relationship to Marshall's "representative firm." It will not be assumed to be necessarily typical of its industry with respect to its size, its efficiency, or the rate of slope

of its various cost curves, but it will be assumed to be typical, or at least to represent the prevailing situation, with respect to the general qualitative behavior of its costs as it varies its own output or, in certain situations, as the industry of which it is part varies its output. All long-run differences in efficiency as between concerns will be assumed, however, to be compensated for by differential rates of compensation to the factors responsible for such differences, and these differential rates will be treated as parts of the ordinary long-run money costs of production of the different concerns. In the long-run, therefore, every concern will be assumed to have the same total costs per unit, except where explicit statement to the contrary is made. It will be assumed, further, that for any industry, under long-run equilibrium conditions, the same relationships must exist for every concern between its average costs, its marginal costs, and market price, as for the particular concern under special examination. But the reasoning of this paper would still hold if the realistic concession were made that in every industry there may be a few concerns which are not typical of their industry with respect to the qualitative behavior of their costs as output is varied either by themselves or by the industry as a whole, and which therefore do not wholly conform to these assumptions. It may be conceded, for instance, that in an industry in which for most producers expansion of their output means lower unit costs there should be a few producers for whom the reverse is true.

Short-Run Equilibrium for an Individual Concern

Chart I, which represents the behavior of money costs in the short-run for a single concern with a plant of a given scale, is the fundamental graph, and is incorporated in or underlies all the succeeding ones.[1] It is assumed that this concern is not of sufficient importance to bring about any change in the prices of the factors as a result of a change in its output. Since unit money costs of

[1] The charts were drawn for me by Y. K. Wong of the University of Chicago. Where in any chart one curve is derived from another or a combination of other curves presented in the same chart, it is drawn mathematically to scale. No attempt has been made, however, to maintain the same scales as between different charts. An attempt has been made to use mnemonic symbols for the various curves, *MC* for instance indicating marginal cost, *P* indicating price, and so forth. It is hoped that this will facilitate reading of the charts.

production are the sum of the products of the amounts of the factors used in the production of one unit multiplied by the prices of the factors, any change in unit money costs as output varies must in this case be due, therefore, to changes in the amounts of the factors required for the production of one unit, or to use Walras' term, to changes in the "technological coefficients of production." The "short-run" is taken to be a period which is long enough to permit of any desired change of output technologically possible without altering the scale of plant, but which is not long enough to permit of any adjustment of scale of plant. It will be arbitrarily

Short-Run Cost Curves

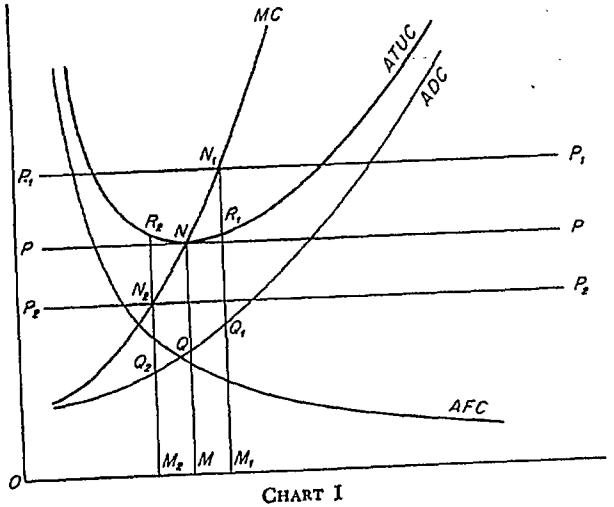

CHART I

assumed that all of the factors can for the short-run be sharply classified into two groups, those which are necessarily fixed in amount, and those which are freely variable. "Scale of plant" will be used as synonymous with the size of the group of factors which are fixed in amount in the short-run, and each scale will be quantitatively indicated by the amount of output which can be produced at the lowest average cost possible at that scale. The costs associated with the fixed factors will be referred to as the "fixed costs" and those associated with the variable factors will be called the "direct costs." It is to be noted that the "fixed costs" are fixed only in their aggregate amounts and vary with output in their amount per unit, while the "direct costs" are variable in their

aggregate amount as output varies, as well as, ordinarily at least, in their amount per unit. Amounts of output are in this as in all the succeeding charts measured along the horizontal axis from *O*, and money costs and prices along the vertical axis from *O*.

The curve *AFC* represents the trend of the average fixed costs per unit as output is increased. Since these are the costs associated with the parts of the working combination which, by hypothesis, are absolutely fixed in their aggregate amount, this curve must be a rectangular hyperbola.[2] The curve *ADC* represents the trend of average direct costs per unit as output is increased. Since the increase in output is the result of the application, to a constant amount of "fixed" factors, of increased amounts of the variable factors, the law of diminishing returns, if it is operating, should make the output per unit of the variable factor employed diminish, i.e., should make the "direct" technical coefficients of production increase, as total output increases. As the prices of the factors by assumption remain constant, the average direct costs must also increase as output increases, if the law of diminishing returns is operative. It is assumed, not, I believe, without justification, that within the useful range of observation the law of diminishing returns is operative, and the average direct cost curve is therefore drawn positively inclined throughout.[3] The curve *ATUC* represents the trend of average total (i.e., fixed plus direct) unit costs as output is increased, and is, of course, the sum of the ordinates of the *ADC* and *AFC* curves. It is necessarily *U*-shaped for all industries having any substantial fixed costs, and is in this respect a universal short-run curve qualitatively descriptive of the short-run behavior of average costs of practically all concerns and all industries which cannot quickly and completely adjust the amounts of all the factors they use to variations in their rates of output. But the relative lengths and the relative rates of inclination of the negatively inclined and the positively inclined portions of the curve will differ from concern to concern and from industry to industry, depending upon the relative importance of the fixed to the total costs and upon the degree of sharpness with which the

[2] I.e., the equation to the curve will be of the form $xy = c$.

[3] It is also drawn concave upward, to indicate the progressively sharper operation of the law of diminishing returns as the fixed factors are more intensively exploited.

law of diminishing returns is operative for the variable factors. The curve MC represents the trend of marginal costs as output is increased. Any point on it represents the increase in aggregate costs as output at that point is increased by one unit.[4]

The marginal cost curve must cut the average cost curve at the lowest point of the latter. At the point of intersection, average cost and marginal cost are of course equal. But average cost is equal to marginal cost only when average cost is constant, i.e., when the average cost curve is a horizontal line.[5] The point of intersection of the marginal cost curve with the average cost curve when the latter is concave upwards must therefore be at the lowest point of the latter, where its tangent is a horizontal line.[6]

If this particular producer is an insignificant factor in his industry, i.e., if atomistic competition prevails, he may reasonably assume that no change in his output, and especially no change consistent with the maintenance of the scale of plant at its original level, will have any appreciable effect on the price of his product. Under these conditions, the partial demand curve for his product may be taken as a horizontal line whose ordinate from the base is equal to the prevailing price.[7] It will be to his interest to carry production to the point where marginal cost equals price, i.e., his short-run MC curve will also be his rational short-run supply curve. If price is MN, this will mean an output of OM and no extra profit or loss on his operations, i.e., the quasi-rent on his fixed investment

[4] If y_a = average fixed cost per unit, y_b = average direct cost per unit, and x = output, then $ATUC = y_a + y_b$, and $MC = \frac{d[(y_a + y_b)x]}{dx}$. It is important to note that no consideration need be given to the fixed costs, if they really are absolutely fixed, in computing the marginal cost. Since $xy_a = c$, and $\frac{dc}{dx} = o, \ldots \frac{d[(y_a + y_b)x]}{dx} = \frac{d(xy_b)}{dx}$.

[5] If x = output, and y = average cost, marginal cost $= \frac{d(xy)}{dx}$. If $y = c$, then $\frac{d(xy)}{dx} = y$. If y is an increasing function of x then $\frac{d(xy)}{dx} > y$. If y is a decreasing function of x, then $\frac{d(xy)}{dx} < y$.

[6] For a mathematical proof, see Henry Schultz, "Marginal Productivity and the Pricing Process," *The Journal of Political Economy*, Vol. XXXVIII (1929), p. 537, note 33.

[7] This is equivalent to saying that the partial demand for his product has infinite elasticity.

per unit of output, NQ, would be equal to the fixed costs per unit. If price is P_1, output will be OM_1, and the quasi-rent per unit of output, N_1Q_1, will be in excess of the fixed costs per unit, R_1Q_1. If P_2 is the price, the output will be OM_2, and the quasi-rent per unit of output will be N_2Q_2, or less than the fixed costs per unit, R_2Q_2. All of these situations are consistent with short-run equilibrium, which, as far as individual producers are concerned, requires only that marginal cost equal price. The short-run supply curve for the industry as a whole is not shown in this chart, but is simply the sum of the abscissae of the individual short-run marginal cost (= individual supply) curves.[8]

Long-Run Equilibrium

The long-run is taken to be a period long enough to permit each producer to make such technologically possible changes in the scale of his plant as he desires, and thus to vary his output either by a more or less intensive utilization of existing plant, or by varying the scale of his plant, or by some combination of these methods. There will therefore be no costs which are technologically fixed in the long-run,[9] and if in fact the scale of plant is not altered as long-run output alters, it will be the result of voluntary choice and not of absolute technological compulsion. For an industry as a whole long-run variations in output can result from more or less intensive use of existing plants, or from changes in the scale of plants, or from changes in the number of plants, or from some combination of these. Under long-run equilibrium conditions changes in output, whether by an individual producer or by the industry as a whole, will be brought about by the economically optimum method from the point of view of the individual producers, so that each producer will have the optimum scale of plant for his long-run output. To simplify the analysis, it will be assumed that in each industry the optimum type of adjustment to a long-run variation in output for that industry as a whole will not only be alike for all producers but will involve only one of the three possible methods of adjustment listed above; namely, change in intensity of use of existing plants, change in scale of plants, and change

[8] It is shown in Chart II.

[9] This is, of course, not inconsistent with the proposition that at any moment within the long-run there will be costs which from the short-run point of view are fixed.

in number of plants. The theoretical static long-run, it should be noted, is a sort of "timeless" long-run throughout which nothing new happens except the full mutual adjustment to each other of the primary factors existing at the beginning of the long-run period. It is more correct, therefore, to speak of long-run equilibrium in terms of the conditions which will prevail after a long-run, rather than during a long-run. Long-run equilibrium, once established, will continue only for an instant of time if some change in the primary conditions should occur immediately after equilibrium in terms of the pre-existing conditions had been reached. The only significance of the equilibrium concept for realistic price theory is that it offers a basis for prediction of the direction of change when equilibrium is not established. Long before a static equilibrium has actually been established, some dynamic change in the fundamental factors will ordinarily occur which will make quantitative changes in the conditions of equilibrium. The ordinary economic situation is one of disequilibrium moving in the direction of equilibrium rather than of realized equilibrium.

For long-run equilibrium not only must marginal cost of output from existent plant equal price for each individual producer, but it must also equal average cost. If this were not the case, there would be either abnormal profits or losses, which would operate either to attract capital into the industry or to induce withdrawal of capital from the industry, and in either case would tend to bring about a change in output. For long-run equilibrium it is further necessary not only that each producer shall be producing his portion of the total output by what is for him, under existing conditions, the optimum method, but that no other producer, whether already in the industry or not, shall be in a position to provide an equivalent amount of output, in addition to what he may already be contributing, at a lower cost. The relations of costs to supply in the long-run will depend on the technological conditions under which output can be most economically varied, and the succeeding discussion will consist in large part of a classification and analysis of these conceivable types of technological conditions.

"Ricardian" Increasing Costs

Chart II illustrates a special case corresponding to the Ricardian rent theory in its strictest form. Let us suppose that a given

industry is already utilizing all of the supply available at any price of a necessary factor of production, so that the output of the industry as a whole can be increased only by the more intensive utilization of the absolutely limited factor. Suppose also that no appreciable economies are to be derived, whatever the output of the industry as a whole, by a combination into larger productive units, or a subdivision into smaller productive units, of the existing concerns. In order further to simplify the analysis, it is assumed that the identical portions of the working-combination which in this case remain technologically fixed in amount whatever may be

"Ricardian" Increasing Costs

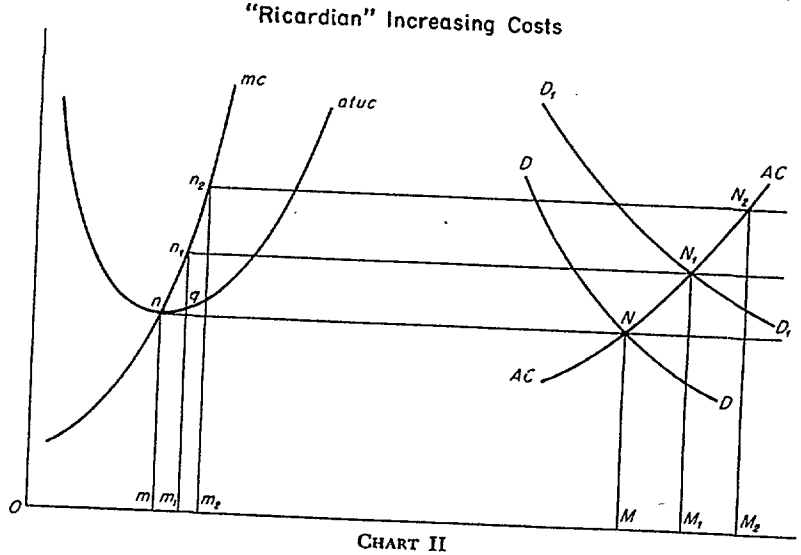

Chart II

the short-run variations in output also remain economically fixed in amount whatever long-run variations in output may occur. If the particular concern whose costs are indicated in the left-hand portion of Chart II and the particular concern with which Chart I is concerned were identical, and if the two charts were drawn to the same scale, the *MC* curve in Chart I and the *mc* curve in Chart II would be identical, although the former represents the short-run trend and the latter represents the long-run trend of marginal costs as output is varied, i.e., for these assumptions, the short-run and the long-run marginal cost curves would be identical. The *atuc* curve in Chart II, continuing these assumptions, would simply represent the short-run variations in average cost for this particular concern

as output was varied, when long-run price was mn or MN,[10] and would be in all respects identical with the $ATUC$ curve of Chart I. When long-run price was MN, this concern would be in both short-run and long-run equilibrium when its output was Om, and its average cost, its marginal cost, and price were all equal.

Suppose now, that owing to a long-run increase of market demand from DD to D_1D_1, long-run price rises to M_1N_1. It will pay our producer to increase his output to Om_1, at which point the new marginal cost, m_1n_1, will be equal to the new price. If the prices of all the factors remain the same, the new price will be higher than the new average cost m_1q. But it is impossible, for a case such as this, to adhere to the assumption that the prices of all the factors remain the same. Given an absolutely limited amount of one of the factors, no change in the prices of the other factors, and a rise in the long-run demand for and in the long-run price of its product, and the long-run price of this absolutely scarce factor must rise. Let us suppose that the fixed factor is land. Its price or rent will rise until there ceases to be any excess of marginal over average cost. The *atuc* curve in Chart II therefore has only short-run significance. A long-run increase in the price of the product will cause an increase in the price of land-use, and therefore a rise in the entire *atuc* curve. The increase in land-rent, however, will have no effect on marginal costs, and therefore on the long-run *mc* curve, for it will be due to the increase in price of the product and not to the increase in output of this particular concern. Even if this producer maintained his output at Om, after long-run price had risen to M_1N_1, the *atuc* curve would rise in the same manner and degree. It would always shift upward in such a way, however, that the *mc* curve would intersect it at its lowest point,[11] i.e., rent

[10] The qualifying phrase in italics is important. Its significance is explained in the next paragraph of the text.

[11] Each successive short-run *atuc* curve of a particular producer, as the long-run price of his product rises, consists of the ordinates of his former *atuc* curve plus a new rent charge fixed in total amount regardless of his output, and therefore of the form $xy = c$. As was pointed out in note 4, page 204, the vertical addition of a rectangular hyperbola to an average cost curve does not affect the marginal cost curve derivable from it. The same *mc* curve can, therefore, continue to be the short-run marginal cost curve, even when the short-run average cost curve is undergoing long-run changes consistently with the conditions assumed in this case.

for land would rise just sufficiently to make the new lowest average cost equal the new equilibrium marginal cost. When the long-run price was M_1N_1, therefore, average cost, marginal cost, and price would be equal for each producer under long-run equilibrium.

The AC curve in the right-hand portion of Chart II represents the long-run supply curve for the industry as a whole, and is simply the sum of the abscissas of the individual *mc* curves. It is also a long-run average cost curve for the industry as a whole inclusive of rent, and a long-run marginal cost curve for the industry as a whole exclusive of rent. For the individual producer, the changes in rent payments required as demand changes are due primarily to the changes in demand, secondarily to the changes in output of the industry as a whole, and only to an insignificant degree to his own changes in output. The individual producer will therefore not take the effect on his rent payments of increased output on his own part into account, and the supply curve for the industry as a whole will therefore be the marginal cost curve for the industry as a whole exclusive of rent.[12]

This appears to be the case usually designated in the textbooks as the case of "increasing costs." I have labelled it as "Ricardian increasing costs" to indicate its close relationship to the Ricardian rent theory. It is to be noted that as output increases the long-run average costs rise even if the increase of rents is disregarded and that there are increasing unit technological costs, therefore, whether the technical coefficients are weighted by the original or by the new prices of the factors. There are increasing marginal costs in every possible sense of the term costs.

If *mc* were the short-run marginal cost curve for a scale adapted to a long-run equilibrium output of Om, and if not all the factors which were technologically fixed in the short-run remained economically fixed in the long-run as output was increased, then, since

[12] For the industry as a whole, however, the increase of output as demand increases will affect rent, on the one hand by influencing price and gross receipts, and on the other hand by influencing gross expenses. Depending upon the shift in position and the elasticity of the demand curve and upon the rate of slope of the industry marginal cost curve exclusive of rent, an increase of output when demand increases may make rent either greater or less than if output were kept constant. But under atomistic competition the possible results of keeping output constant when demand rises will play no part in the determination of output, of price, or of rent.

there would be less scope for the operation of the law of diminishing returns, the long-run marginal cost curve for the particular concern would be different from and less steeply inclined than the *mc* curve, and the new short-run $atuc_1$ curve for a long-run equilibrium scale of output of, for example, Om_1 would have no simple relationship to the *atuc* curve in Chart II. Similarly, the long-run supply curve for the industry as a whole, since it is the sum of the abscissas of the individual long-run marginal cost curves, would then also be less steeply inclined than the *AC* curve in Chart II, which would then be only a short-run supply curve for the industry as a whole, when the long-run equilibrium output of the industry was *OM*.

Constant Costs

In the short-run, for industries which have any fixed costs whatsoever, constant marginal costs as output is varied are wholly inconceivable if the law of diminishing returns is operative, and constant average costs are inconceivable if there are increasing marginal costs as required by the law of diminishing returns.[13]

In the long-run, however, constant costs are theoretically conceivable under two kinds of circumstances. The first case is when each producer can vary his scale of production without affecting his long-run average costs. The situation in this case for any individual concern will be as represented in Chart III. The curves $atuc_1$ and mc_1 represent, respectively, the short-run trends of average and marginal costs as output is varied from a plant of scale *OA*. The curves $atuc_2$ and mc_2, similarly represent, respectively, the short-run trends of average and marginal costs as output is varied from a plant of scale *OB*; and similarly, for scales *OC* and *OD*. In the long-run any output would be produced from the optimum scale for this output. The long-run average cost curve would therefore be the horizontal line *AC*, which passes through the lowest points of all the short-run *atuc* curves. Where average

[13] Let x = output, y_a = average fixed costs per unit, y_b = average direct costs per unit, and c and k be two different constants. Suppose that short-run average costs are constant, i.e., that $y_a + y_b = k$. But $xy_a = c$. Then $xy_b = kx - c$, and marginal cost, or $\frac{d(xy_b)}{dx} = \frac{d(kx - c)}{dx} = k$, which is inconsistent with the law of diminishing returns.

costs are constant as output varies, average cost and marginal cost are always identical.[14] This horizontal line would therefore also be the individual producer's long-run supply curve.

This case presents certain difficulties when perfect competition prevails which make it impossible to indicate graphically the relationship between the long-run supply curves of the individual concern and the industry as a whole. Read as an ordinary supply curve, the *AC* line indicates that in the long-run this concern would be unwilling to operate at any price under *AN*, would be

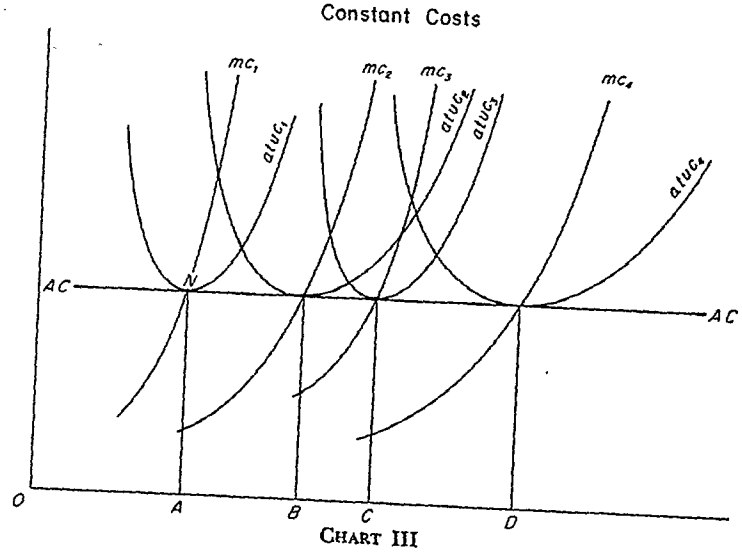

CHART III

willing to produce any amount at a price *AN*, and would be anxious to produce unlimited quantities at any price over *AN*. If the costs of different producers in the industry are not uniform, then the lowest cost concern would tend to monopolize the industry. If the costs of different producers are uniform, the supply curve for the industry would be indefinite, and in the long-run there would be a constant tendency toward overproduction, with consequent losses and a reaction toward underproduction. Actual long-run price and output would be unstable, but would oscillate above and below stable points of equilibrium price and equilibrium output.

The second conceivable case of long-run constant costs, not

[14] See note 5, page 204.

illustrated graphically here, would be presented by a situation in which all of the concerns within the industry and an indefinite number of potential members of the industry can operate at long-run minimum average costs uniform as between the different concerns, but with average costs increasing for each as its output increases. The long-run output of the industry would then consist of the sum of the outputs of all the member concerns, each operating at that scale at which its costs are at the minimum common to all, and variations of output for the industry as a whole would result wholly from variations in the number of producers, each of whom would maintain a constant output while he remained in the industry. For the industry as a whole, therefore, long-run production would take place under conditions of constant long-run average and marginal cost, uniform for all producers and equal to each other, although each concern would be operating subject to short-run increasing average and marginal costs. Here also actual long-run price and output for the industry as a whole would tend to be unstable, but would oscillate above and below stable points of equilibrium price output.

The situation would in these two cases be somewhat analogous to that of a thermostatic control which aims at maintaining a uniform temperature, which is stimulated into operation only when there is a significant degree of variation from the desired temperature, and which succeeds only in keeping the ever-present variations from the desired temperature from exceeding narrow limits in either direction. Completely stable equilibrium under constant cost conditions is only conceivable on the assumption of some departure from perfect competition, in consequence of which variations in output by individual producers, or entrance into the industry by new producers or withdrawal of old, are subject to some difficulty even in the long-run after the equilibrium price and output have once been momentarily established.

Net Internal Economies of Large-Scale Production

We owe to Marshall the important distinction between the "internal" and the "external" economies resulting from increased output. For present purposes we will use the term "net internal economies of large-scale production" to mean net reductions in costs to a particular concern resulting from a long-run expansion

in its output when each output is produced from a plant of the optimum scale for that output. The word "net" is introduced to make it clear that increase in output may result at the same time in economies and in diseconomies and that it is only the excess of the former over the latter to which reference is made here. Internal economies of large-scale production are primarily a long-run phenomenon, dependent upon appropriate adjustment of scale of plant to each successive output. They should not be confused with the economies resulting from "spreading of overhead," which are a short-run phenomenon, represented by the negative inclination of the average fixed cost curve in Chart I. Internal economies of large-scale production need not be relatively greater for those particular costs which in the short-run are the fixed costs than for those particular costs which in the short-run are the direct costs. In the long-run, in any case, there are no technologically fixed or overhead costs, if the definitions here followed of "long-run" and of "fixed costs" are adhered to. Internal economies of large-scale production are independent of the size of output of the industry as a whole, and may be accruing to a particular concern whose output is increasing at the same time that the output of the industry as a whole is undergoing a decline. It is for this reason that Marshall gave them the name of internal, to distinguish them from the external economies which are dependent on something outside the particular concerns themselves, namely, the size of output of the industry as a whole.

Internal economies may be either technological or pecuniary, that is, they may consist either in reductions of the technological coefficients of production or in reductions in the prices paid for the factors as the result of increases in the amounts thereof purchased. Illustrations of technological internal economies would be savings in the labor, materials, or equipment requirements per unit of output resulting from improved organization or methods of production made possible by a larger scale of operations. Pecuniary internal economies, on the other hand, would consist of advantages in buying, such as "quantity discounts" or the ability to hire labor at lower rates, resulting from an increase in the scale of purchases.[15]

[15] Pecuniary internal economies are, theoretically, as likely to result from expansion of output from a given plant as from expansion of output brought

Chart IV illustrates the behavior of the cost curves for a particular concern which enjoys net internal economies of large-scale production. As in Chart III the *ac* curves and the *mc* curves represent the short-run variations in average and marginal costs respectively, as output is varied from plants of each indicated scale. The *AC* curve represents the long-run trend of average costs, that is, the trend of average costs when each output is produced from a plant of the optimum scale for that output, and is drawn so as to connect the points of lowest average cost for each scale of plant.[16] The *MC* curve is the long-run marginal curve for this particular concern when the *AC* curve is interpreted as a continuous curve. It represents the increment in aggregate costs resulting from a unit increase in output, when each output is produced from a plant of the optimum scale for that output. It is to be noted that while the short-run marginal cost curves are positively inclined, the long-run marginal cost curve is negatively inclined.[17]

about by increase of scale of plant. But it is only the latter form of expansion of output which is likely to be great enough to result in significant pecuniary internal economies.

[16] The *AC* curve would represent a continuous trend only if it is assumed that scale of plant can be modified by small increments. If the curve is interpreted as a discontinuous one, then only the points N, N_1, N_2, . . . on it are significant, and the significant long-run costs for the intervals between are the lowest short-run average costs available for the indicated outputs. It may be noticed that at certain points the short-run *ac* curves are drawn so as to sink below the long-run *AC* curve. If the *AC* curve is interpreted as having significance only at the N points, this is of no consequence. But if the *AC* curve is interpreted as a continuous curve, this is an error. My instructions to the draftsman were to draw the *AC* curve so as never to be above any portion of any *ac* curve. He is a mathematician, however, not an economist, and he saw some mathematical objection to this procedure which I could not succeed in understanding. I could not persuade him to disregard his scruples as a craftsman and to follow my instructions, absurd though they might be.

[17] If y, y_1, y_2, are the short-run average costs for scales of plant, OM, OM_1, and OM_2, respectively, as indicated by the *ac* curves; Y = long-run average cost, as indicated by the *AC* curve; x = output; mc, mc_1, and mc_2 indicate the short-run marginal costs as represented by the *mc* curves; and MC indicates the long-run marginal cost, as represented by the *MC* curve, then:

$$mc = \frac{d(xy)}{dx};\ mc_1 = \frac{d(xy_1)}{dx};\ mc_2 = \frac{d(xy_2)}{dx};\ \text{and } MC = \frac{d(xY)}{dx};$$

$$\text{and } \frac{d^2(xy)}{dx^2} > 0;\ \text{and } \frac{d^2(xY)}{dx^2} < 0.$$

The familiar proposition that net internal economies of large-scale production and long-run stable equilibrium are inconsistent under competitive conditions is clearly illustrated by this chart. When price is MN, this concern, if operating with the scale of plant represented by the short-run curves *ac* and *mc*, is in short-run equilibrium when its output is OM, for its short-run marginal cost is then equal to price. It will not be in long-run equilibrium, however, for its long-run marginal cost will then be only MQ, or less than price. Provided that no change in its output will affect market price, it will pay this concern to enlarge its plant whatever the price may

Net Internal Economies of Large-Scale Production

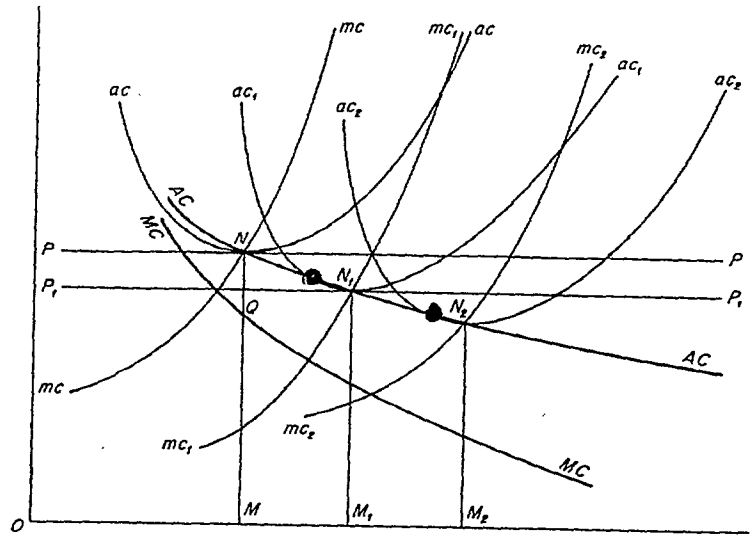

CHART IV

be, and whatever its existing scale of plant may be. If thereby it grows so large that its operations exert a significant influence on price, we pass out of the realm of atomistic competition and approach that of partial monopoly. Even then, however, it would still be profitable for this concern to enlarge its plant and increase its output as long as long-run marginal cost was lower than long-run marginal revenue, or the increment in aggregate receipts resulting from a unit increment in output, after allowance for any reduction in price.[18]

[18] If Y_p = long-run price, X = long-run output, and Y_c = long-run average cost, long-run marginal cost would be $\frac{d(XY_c)}{dX}$, long-run marginal revenue

For any particular concern operating under these conditions, and *a fortiori* for an industry as a whole consisting of such concerns, there is no definite long-run supply curve. At any price MN higher than the asymptote of the AC curve, this producer will be willing to produce any quantity not less than OM.

To negatively-inclined long-run cost curves such as the AC and MC curves in Chart IV, Marshall has denied the characteristic of "reversibility," i.e., of equal validity whether output is increasing or decreasing, on the ground that some of the economies accruing when the output of a concern, or of an industry as a whole, is increased will be retained if the output of the concern or of the industry returns to its original dimensions.[19] This reasoning appears to involve a confusion between static and dynamic cost curves. The reductions in costs as output is increased indicated by curves such as the AC and MC curves in Chart IV are purely functions of size of output when scale is adjusted to output and not of lapse of actual time during which improved processes may happen to be discovered. The economies associated with output OM are economies which are not available for any output less than OM. The only basis on which the irreversibility of these curves, as static curves, could logically be posited would be the existence of possible economies of a type adapted to any scale of output but discoverable only when output is great, where invention, but not its exploitation, was a function of scale of output.

would be $\frac{d(XY_p)}{dX}$, and it would pay to carry production to the point where long-run marginal cost equalled long-run marginal revenue, or $\frac{d(XY_c)}{dX} = \frac{d(XY_p)}{dX}$. Under atomistic competition, $\frac{d(XY_p)}{dX} = Y_p$, which is independent of this particular concern's output. Whatever the price, therefore, this concern would always have an incentive to increase its long-run output as long as long-run marginal cost remained less than that price. If partial monopoly resulted, however, marginal revenue, or $\frac{d(XY_p)}{dX}$, would become a function of market demand and of competitor's supply and would be smaller than Y_p, and a point of stable long-run equilibrium might exist, depending on how the other producers reacted to variations in output by this one. If complete monopoly resulted, there would probably be a definite point of stable equilibrium. These questions, however, are beyond the range of this paper.

[19] *Principles of Economics*, 8th ed., 1922, p. 808.

Net Internal Diseconomies of Large-Scale Production

Cases are clearly conceivable where increase of scale of plant would involve less efficient operation and consequently higher unit costs. The prevailing opinion in the United States that for most types of agriculture the one-family farm is still the optimum mode of agricultural organization would indicate that in this country at least agriculture was subject to net internal diseconomies of large-scale production after an early stage in the size of the farm-unit had been reached. But when increase of output by means of the increase of scale of existing plants involves a substantial increase in unit costs, it will always be possible for the industry as a whole to avoid the net internal diseconomies of large-scale production by increasing its output through increase in number of plants without increase in their scale.[20] This case has no practical importance, therefore, except as it represents an economic barrier against increase in scale of plants, and it is not worth while to illustrate it graphically.

Net External Economies of Large Production

External economies are those which accrue to particular concerns as the result of the expansion of output by their industry as a whole, and which are independent of their own individual outputs. If an industry which enjoys net external economies of large production increases its output—presumably through increase in number of plants—the average costs of the member concerns of that industry will fall even though each concern maintains a constant scale of plant and a constant output. Like internal economies, external economies may be either technological or pecuniary. Illustrations of technological external economies are difficult to find, but a better organization of the labor and raw materials markets with respect to the availability of laborers and materials when needed by any particular plant, and improvement

[20] Increase of scale should be distinguished from increase in output from the same scale of plant. In the former, all the factors are increased in about the same proportions; in the latter some factors remain fixed in amount. Whenever it is generally possible to increase all the factors in about the same proportion, i.e. to increase scale of plant, it is also possible, alternatively at least, to increase the number of plants.

in productive technique resulting from "cross-fertilization," or the exchange of ideas among the different producers, appear to be possible sources of technological external economies resulting from the increase in size of the industry as a whole. Illustrations of pecuniary external economies would be reductions in the prices of services and materials resulting from the increase in the amounts of such services and materials purchased by the industry as a whole. Pecuniary external economies to industry *A* are likely to be internal or external economies to some other industry *B*. If

Net External Economies of Large Production

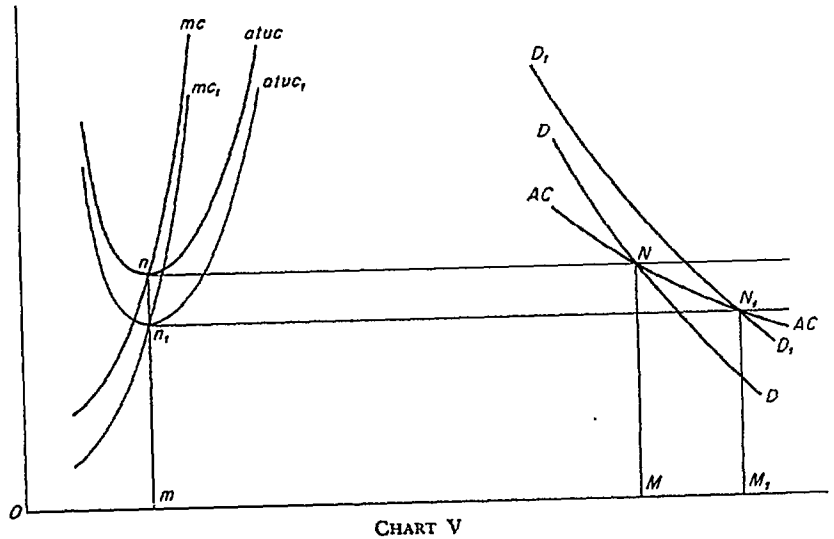

CHART V

industry *A* purchases materials in greater quantity, their price may fall because industry *B* can then produce them at lower unit cost. But cases are theoretically conceivable where pecuniary external economies to industry *A* may not be economies to any other industry, as, for instance, if laborers should have a preference, rational or irrational, for working in an important rather than in a minor industry, and should therefore be willing to accept lower wages as the industry expands.

Chart V illustrates the case of net external economies of large production, irrespective of whether these economies are technological, or pecuniary, or both. As always, each concern will in the long-run tend to produce its output from the optimum scale

for that output, and given that scale, to carry production to the point where its average and marginal costs are both equal to price. If *Om* represents the optimum scale of plant for the particular producer, i.e., the scale at which he can produce at the lowest average cost, if the long-run price is *mn* or *MN*, and if the long-run output for the industry as a whole is *OM*, this producer will be in long-run equilibrium when his output is *om*, and his average and his marginal cost are both *mn*. Suppose now that long-run demand rises from *DD* to D_1D_1, and that long-run output of the industry as a whole increases, as the result of increase in the number of producers, from *OM* to OM_1. Since, by assumption, this industry is subject to net external economies of large production, the short-run average and marginal cost curves of each particular concern will fall in the manner indicated in the left-hand portion of Chart V. This particular concern will be in long-run equilibrium with the new situation when its output is *om*, as before, but its long-run average and marginal costs will have fallen from *mn* to mn_1. The *AC* curve represents the trend of the individual average (and also marginal) costs as output of the industry as a whole changes by the amounts indicated on the horizontal axis. Any point on this curve represents the long-run average cost for every individual producer, and therefore for the industry as a whole, when the output of the industry as a whole is as indicated. It is theoretically the same as the supply curve for the industry as a whole. The long-run marginal cost curve for the industry as a whole is not shown on the chart. It would fall below the *AC* curve.[21] Its only relationship to the short-run marginal cost curves of the individual concerns would be that it was a function of the downward shifting of the lowest points on the individual short-run *atuc* and *mc* curves as the output of the entire industry increased. Under atomistic competition this marginal cost curve would have no influence on supply, since individual producers would not take it into account in deciding either upon their continuance in

[21] If X = output of the industry as a whole, and Y_a = long-run average cost for the industry as a whole as represented by the *AC* curve, the *MC* curve for the industry as a whole would be $\frac{d(XY_a)}{dX}$, $< Y_a$. If average cost for a particular producer $= y_a$, then $y_a = f(X)$, and at long-run equilibrium, $y_a = Y_a$.

or their entrance into the industry or upon their scale of output when in the industry.[22]

Net External Diseconomies of Large Production

Although it has not ordinarily been given consideration, the case of net external diseconomies of large production is of indisputable practical importance. Pecuniary diseconomies of this kind will always tend to result from the expansion of output of an industry because the increased purchases of primary factors and materials which this entails must tend to raise their unit prices. In order that pecuniary diseconomies shall not result from the expansion of an industry's output, it is necessary, for both primary factors of production and materials, that the increase in demand by this industry shall be accompanied by a corresponding and simultaneous decrease in demand by other industries or increase in supply of the factors and materials themselves, or, failing this, that the materials, because of net external or internal economies in the industries producing them, should have negatively inclined supply curves.[23] These pecuniary external diseconomies, however,

[22] Employing terminology resembling that used by Pigou in his *The Economics of Welfare*, the marginal private net cost would exceed the marginal industry net cost. If the output of an additional producer be represented by ΔX, and the average cost of his output and of the outputs of the other producers by $y_a = f(X)$, then the marginal private net cost would be y_a, and the marginal industry net cost would be $\frac{\Delta(XY_a)}{\Delta X}$, $< y_a$.

[23] It is worth pointing out that negative supply curves for the primary factors of production will not prevent an increased demand for them from a particular industry from resulting in an increase in their unit prices and therefore are not a barrier to pecuniary external diseconomies for that industry in so far as their primary factor costs are concerned. The negatively inclined supply curves of primary factors have a different meaning from the negatively inclined supply curves for commodities. If labor has a negatively inclined supply curve that means not that willingness to hire labor in greater quantities will result in a fall in the wage-rate, but, what is very different, that fewer units of labor will be offered for hire when a high rate of wages is offered than when a lower rate is offered. In the case of commodities, any point on a negatively inclined supply curve must be interpreted to mean that at the indicated price, the indicated quantity or more of the commodity can be purchased. In the case of labor, any point on a negatively inclined supply curve must be interpreted to mean that when the indicated wage-rate is obtainable, the indicated quantity of labor, but no more, will be available for

may be more than counterbalanced by technological external economies, and need not necessarily result therefore in net external diseconomies. External technological diseconomies, or increasing technical coefficients of production as output of the industry as a whole is increased, can be theoretically conceived, but it is hard to find convincing illustrations. One possible in-

Net External Diseconomies of Large Production

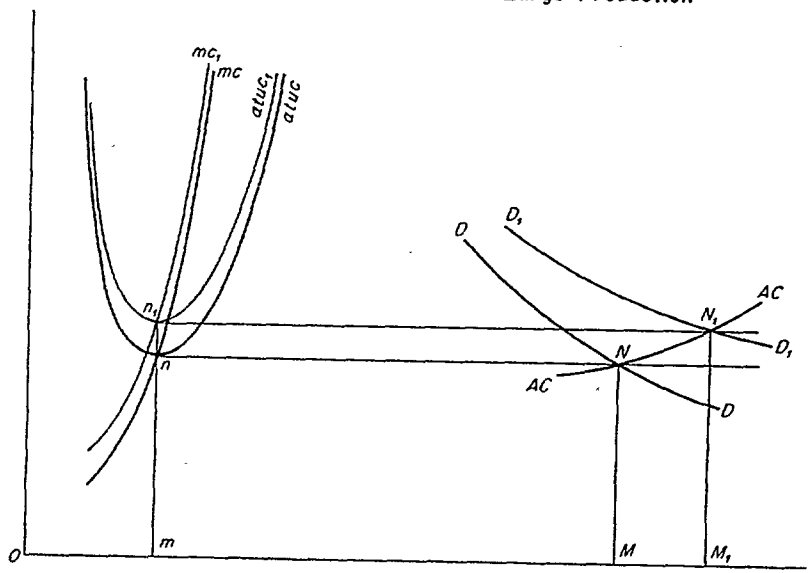

Chart VI

stance might be higher unit highway transportation costs when an industry which provides its own transportation for materials and products expands its output and thereby brings about traffic congestion on the roads.

Chart VI illustrates the case of net external diseconomies of large production, whether technological or pecuniary. When the long-run equilibrium outputs of the industry as a whole are OM and OM_1, respectively, the *atuc* and $atuc_1$ curves represent the respective trends of short-run average costs, the *mc* and mc_1 curves

hire. If the negatively inclined supply curve for labor has an elasticity of less than unity, as seems probable, it must be assumed that labor will prefer a high wage rate and partial employment to a low wage rate and fuller employment, and therefore will resist any movement toward the lower points on its supply curve.

represent the trends of short-run marginal costs and mn and mn_1 represent the long-run equilibrium average and marginal costs, for one individual producer. The reverse of the conditions when net external economies of large production are present, in this case the long-run equilibrium average and marginal costs of the individual concern rise as the output of the industry as a whole increases. The AC curve represents the trend of the individual average (and also marginal) long-run costs and therefore also of the industry long-run average cost as the industry as a whole varies its output. This is also the long-run supply curve for the industry as a whole. The long-run marginal cost curve for the industry as a whole is not shown on the chart. It would rise above the AC curve.[24] Since the individual producers will not concern themselves with the effect on the costs of other producers of their own withdrawal from or entrance into the industry, and since in this case it is assumed that variation in output takes place only through variation in number of producers, the marginal cost curve for the industry as a whole will, under competitive conditions, have no influence on output.[25]

Particular Expenses Curves

In the foregoing analysis of the relation of cost to supply, it has been throughout maintained, explicitly or implicitly, that under long-run static competitive equilibrium marginal costs and average costs must be uniform for all producers. If there are particular units of the factors which retain permanently advantages in value productivity over other units of similar factors, these units, if hired, will have to be paid for in the long-run at differential rates proportional to their value productivity, and if em-

[24] As for Chart V, if X = output of industry as a whole and Y_a = long-run average cost for industry as a whole, as represented by the AC curve, the marginal cost curve for the industry as a whole would be $\frac{d(XY_a)}{dX}$. If for the individual concern, y_a = average cost, then $y_a = f(X)$, and at long-run equilibrium $y_a = Y_a$.

[25] In Pigou's terminology, the marginal industry net cost would exceed the marginal private net cost. If the output of an additional concern be represented by ΔX, and his average cost by $y_a = f(X)$, then the marginal private net cost would be y_a, and the marginal industry net cost would be $\frac{\Delta(XY_a)}{\Delta X} > y_a$.

ployed by their owner should be charged for costing purposes with the rates which could be obtained for them in the open market and should be capitalized accordingly. In the short-run, the situation is different. There may be transitory fluctuations in the efficiency of particular entrepreneurs or of particular units of the factors, and it would neither be practicable nor sensible to recapitalize every unit of invested resources with every fluctuation in their rate of yield. Even in the short-run, there must be equality as between the marginal costs of different producers under equilibrium conditions,[26] but there may be substantial variations as between the average costs, and therefore as between the net rates of return on original investment, of different producers.

Statistical investigations of individual costs in the United States, based in the main on unrevised cost accounting records, have shown that the variations in average costs as between different producers in the same industry at the same time are very substantial, and that ordinarily a significant proportion of the total output of an industry appears to be produced at an average cost in excess of the prevailing price. To some extent these variations in cost can be explained away as due (1) to different and, from the point of view of economic theory unsatisfactory, methods of measuring costs, and especially the costs associated with the relatively fixed factors of production, (2) to regional differences in f.o.b. factory costs and in prices which, in an area as large as the United States, can be very substantial for bulky commodities without implying the absence of keen competition and (3), to the absence of atomistic competition. But even aside from such considerations, it should be obvious that such findings are in no way inconsistent with the propositions of equilibrium price theory as outlined above. Under short-run equilibrium the average costs,

[26] Since a time-interval is always present between the sale contract and at least some of the stages of hiring of factors and of actual production, there is opportunity under short-run equilibrium for some divergence between price and marginal cost, and therefore, between the marginal costs of different producers. It would be a more precise way of formulating the short-run theory to say that since all producers, if acting rationally, carry production to the point where anticipated marginal cost will equal anticipated price, and since price, in a perfect market, is uniform for all, marginal cost tends to be uniform for all producers, and variations as between different producers result only from errors in anticipation.

including the fixed costs, of any particular producer need bear no necessary relationship to price, except that the average direct costs must not exceed price. These statistical costs, moreover, are not the equilibrium costs of the theoretical short-run, but are the costs as they exist at an actual moment of time when short-run equilibrium with the fundamental conditions as they exist at the moment may not have been attained, and when these fundamental conditions are themselves liable to change at any moment.

It may be worth while, however, to show the relationship of the distribution of particular average costs within an industry at particular actual moments of time to the general supply conditions of the industry under assumptions of long-run equilibrium. To a curve representing the array of actual average costs of the different producers in an industry when the total output of the industry was a given amount, these individual costs being arranged in increasing order of size from left to right, Marshall gave the name of "particular expenses curve,"[27] and American economists have called such curves "bulk-line cost curves,"[28] "accountants' cost curves," and "statistical cost curves." In Chart VII, the curves AN, BN_1, and CN_2, are supposed to be the appropriate particular expenses curves for an industry subject to net external economies of large production, when the output of the industry as a whole is OM, OM_1, and OM_2, respectively. Because the industry is subject to net external economies, the entire particular expenses curves are made to shift downward as the output of the industry expands. (If the industry were subject to net external diseconomies of production, the particular expenses curves would shift upwards as the output of the industry expands. Corresponding modifications

[27] See *Principles*, 8th ed., Appendix *H*, p. 911. It will be noticed that his particular expenses curve, *SS* is drawn so as to project somewhat beyond the point of total output for the industry as a whole *A*. This is an error, and no significance can be given to the part of the curve projecting beyond the point of total output of the industry as a whole. If the output of the industry were to increase up to the terminal point of this curve, the entire curve would acquire a different locus.

[28] "Bulk-line cost curves" because if a perpendicular is dropped to the horizontal axis from the point of intersection of the price-line and the curve, the greater part or the "bulk" of the output would be to the left of this "bulk-line." See F. W. Taussig, "Price-Fixing as seen by a Price-Fixer," *The Quarterly Journal of Economics*, Vol. XXXIII.

in the chart would have to be made as other assumptions with respect to the conditions under which the industry can expand its output were introduced.) It is to be understood also that no dynamic changes in prices of the factors or in average technological cost conditions for the industry as a whole are occurring except such as are associated with variations in output of the industry as a whole.

The *HC* curve is a curve connecting the points of highest-cost for each successive output. These highest-costs, though often so designated, are not marginal costs in the strict sense of the term,

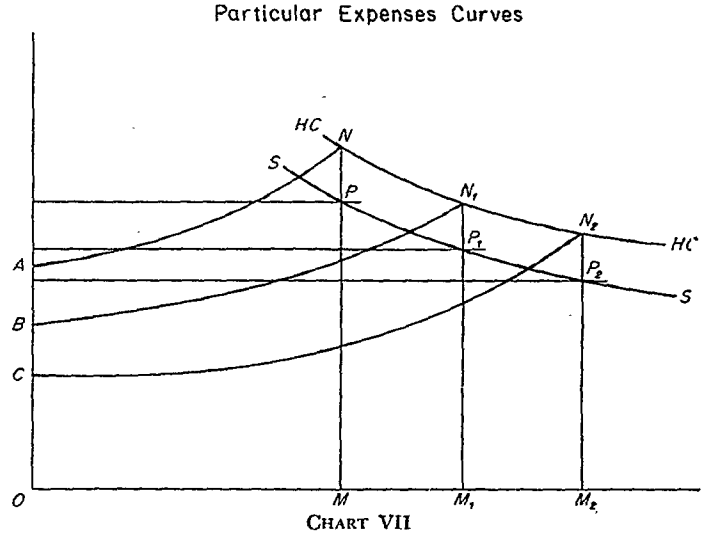

CHART VII

but are in each case simply the average costs of that produc[er] whose average costs are the highest in the industry. If the statistical indications and also certain *a priori* considerations are to be followed, these highest average costs are likely to be, except in "bad years," distinctly higher than the true marginal costs,[29] and

[29] If the AN, BN_1, and CN_2 curves were the actual particular expenses curves when the actual outputs of the industry as a whole were OM, OM_1, and OM_2 respectively, the actual marginal cost curve for the industry as a whole would be a curve representing the differences per unit increase of output between the aggregate costs represented by the successive areas, $AOMN$, BOM_1N_1, COM_2N_2, . . . as output was increased from OM to OM_1, to OM_2, to . . . It would be negatively inclined, and would be much below the *HC* curve.

so drawn in this graph. The P, P_1, P_2 lines represent price, and are drawn to intersect the particular expenses curves below their highest points, in conformity with the statistical findings. The curve *SS*, drawn through the P, P_1, P_2 points representing actual prices prevailing when the outputs are OM, OM_1 and OM_2, respectively, is a sort of actual semi-dynamic[30] supply curve.

What is the ordinary relationship between the *HC* curve and the *SS* curve under fully dynamic conditions cannot be postulated on *a priori* grounds, and only statistical investigation can throw much light on it. American investigators of particular expenses curves believe that they have already demonstrated stable and predictable relations between them and price, but a reasonable degree of scepticism still seems to be justified. One point, however, is clear on *a priori* even more than on inductive grounds. If the *SS* curve in Chart VII were not ordinarily below, and substantially below, the *HC* curve, the familiar and continuously present phenomenon of bankruptcy would be inexplicable.

It is possible, moreover, to devise a theory of even long-run static equilibrium which still leaves room for an excess of the *HC* over the *SS* curves, and therefore for bankruptcy as a phenomenon consistent with long-run equilibrium. For such a theory, however, long-run equilibrium would apply only to the industry as a whole, and would be a sort of statistical equilibrium between rate of output and rate of consumption. None of the individual producers under this theory need be in long-run equilibrium at any time. At any moment, some producers would be enjoying exceptional profits, and others incurring heavy losses. The particular expenses curve could remain positive in its inclination and fixed in its locus, but there would be necessarily a constant process of shifting of their position on that curve on the part of the individual producers, and an equality in rate of withdrawal of producers from the industry through bankruptcy or otherwise, on the one hand, and of entrance of new producers into the industry, on the other hand. A theory of this sort would leave room for pure profits even in a static state.

[30] "Semi-dynamic" because certain types of dynamic changes have been assumed not to occur.

SUPPLEMENTARY NOTE (1950) *

I do not take advantage of the opportunity to revise my 1931 article. Even the error in Chart IV (page 215) is left uncorrected, so that future teachers and students may share the pleasure of many of their predecessors of pointing out that if I had known what an "envelope" was I would not have given my excellent draftsman the technically impossible and economically inappropriate assignment of drawing an *AC* curve which would pass through the lowest cost points of all the *ac* curves and yet not rise above any *ac* curve at any point. It is left also to the reader to modify the general contour of the ATUC curve of Chart I in conformity with the evidence which inductive studies seem to provide that the trough of this curve has a negative inclination throughout almost all of its possible course.

I feel it incumbent upon me, however, so as to avoid propagating serious error, to carry the analysis of costs a stage further in one respect by departing here from the traditional Marshallian pattern of assumptions to which the article adheres. The partial-equilibrium nature of the Marshallian assumptions leaves a wider range of possibilities to the long-run tendencies of costs for an expanding industry than is consistent with general-equilibrium analysis. I first saw this in 1938, and thereafter pointed it out to my students at the University of Chicago. But the first, and, to my knowledge, still the only, analysis in print similar to what I have in mind[31] is in Joan Robinson's excellent article, "Rising Supply Price," *Economica,* VIII, February, 1941 (see page 241 of the present volume), which has not attracted the attention which in my opinion it eminently deserves. What follows is, I think, in substantial harmony with her argument, but is so presented as to provide a link with the analysis in my 1931 article.

The most significant long-run behavior of costs for many appli-

* Reprinted from *Readings in Economic Analysis,* ed. by R. V. Clemence, Addison-Wesley Press, Cambridge, 1950, Vol. II, pp. 31–35, by the courtesy of the publisher and the author, with an additional note.

[31] I have since found the same doctrine expounded in an earlier article by R. F. Harrod, "Notes on Supply," *The Economic Journal,* Vol. XL (1930), pp. 232–241, especially pp. 240–241. [Note added in 1951.]

cations of value theory to concrete economic issues is the trend of unit costs, average or marginal, for a particular commodity (or group of commodities) as the total output of such commodity expands while the economy as a whole remains stable (or relatively stable), so that the expansion of output of this commodity is of necessity simultaneous with a corresponding contraction of output of all other commodities considered in the aggregate. Let us assume that in an otherwise stable economy a shift of wants occurs from other commodities to cloth, with a consequent expansion in the output of cloth. Except by coincidence, and even that conceivable only at a "point" rather than over a substantial range, the cloth industry will be using the various "factors" (or "ingredients," or "resources," or "input items") in proportions somewhat different from those in which the economy as a whole, and the contracting section of it, uses them. As the cloth industry expands, therefore, and bids for more factors, the contracting industries will not, at prevailing prices for the factors, be releasing factors in the same proportions in which the cloth industry is trying to acquire them; at prevailing prices for the factors, those which the woolen industry uses relatively heavily will be in short supply while those which it uses relatively lightly will be in excess supply. There will consequently occur a realignment of the prices of the factors, with those used relatively heavily by the cloth industry rising in price and those used relatively lightly by it falling in price.

Thus all industries must tend to be subject to "external net pecuniary diseconomies of large production" when they expand relative to the economy of which they are a part. The entrepreneurs in an expanding industry, to lessen the impact of these pecuniary diseconomies, will endeavor to reduce the ratio of their use of the factors which have risen in price to their use of the factors which have fallen in price. But the extent to which such change in the proportions in which the factors are combined is technically feasible and economically profitable is limited by the operation of the law of diminishing returns, i.e., increase in the relative use of the cheaper factors results in decreased ratios of output to input of these factors. It is not possible therefore completely to escape the pecuniary diseconomies resulting from the relative changes in the prices of the factors by altering the proportions in which the factors are used, if it is assumed, as I do here,

that the law of diminishing returns is operating in the long run.

There is presented on page 231 an arithmetical illustration of the conclusions derivable from this reasoning with respect to money costs per unit of product and allocation of resources as between different industries when in an economy of stable size a shift of wants of given extent in favor of cloth results in an expansion of the cloth industry. Case I represents what are for present purposes the essential characteristics of the assumed original equilibrium of the economy as a whole, and Case II represents a possible new equilibrium, consistent with all the assumptions made, after there has been full adjustment to the shift in wants. It is assumed in Case II that all the factors are fixed in amount, and also that the total national income remains at $320. There are in Case II as compared to Case I: an increase in the output of cloth; a rise in the price of factor B used relatively heavily by the cloth industry and decreases in the prices of other factors; a relative decrease in the use of factor B as compared to other factors for each industry (although not for all industry in the aggregate); a rise in the average cost and in the price of cloth (there would also be a rise in the marginal cost of cloth but this is not brought out explicitly in the illustration) and falls in the average costs and prices of all other commodities taken as a whole. The degrees of change from Case I indicated in the illustration are in every instance arbitrary, though consistent with equilibrium for the economy as a whole, but the directions of change are in every instance necessary ones.

The assumption that the factors are fixed in amount, i.e., that the amounts offered for hire are independent of their rates of remuneration, is an unnecessary one; though for fortuitous historical reasons it is a popular assumption in economic theory and even sometimes presented as dogma which it is not respectable to depart from, it is in fact wholly arbitrary and unrealistic. It is easy to modify the arithmetical illustration, however, to adapt it to other types of assumptions as to the character of the supply functions of the factors. I will not take the space required to do this here, but will confine myself to a summary account of the character of the necessary changes in the results which follow changes in the assumptions as to the supply functions of the factors, all other assumptions remaining as before.

Case III. Assume that each of the factors has a supply function such that the quantity offered for hire is an increasing function of the rate of remuneration. All price and cost changes as compared to Case I will be the same in direction as in Case II, but less in degree. The total quantity of factor B used by the economy as a whole will be greater and of factors B and C will be less than in Case I or Case II.

Case IV. Assume that factor B has a fixed supply, while the amounts offered of factors A and C are increasing functions of their rates of remuneration. The cost and price of cloth will rise more than in Cases II or III and the cost and prices of other commodities will fall less than in Cases II or III.

Case V. Assume that the quantity of factor B offered for hire is an increasing function of its rate of remuneration, while the supplies of factors A and C are fixed. The cost and price of cloth will rise, but less than in Cases II, III, or IV, while the prices of other commodities will fall more than in Cases II, III, or IV.

Case VI. Assume that the supply functions of all of the factors of production can be represented graphically by "rising-backward" curves, i.e., that as higher rates of remuneration are offered for them smaller quantities are supplied.[32] The rise in the cost and the price of cloth would be greater and the fall in the costs and prices of other commodities would also be greater than in any of the preceding Cases II to V.

In all these cases an increase in the long-run output of cloth can occur only at higher cost, and these or allied cases cover all the assumptions as to the supply functions of the factors which seem to me to be conceivable as realities if the possibility of migration of factors to or from the economy in question is excluded. If the reasoning here presented is valid, there is therefore a universal long-run "law" of increasing money costs as output changes in response to shifts in wants in an economy of constant national

[32] "Rising-backward" supply curves need to be interpreted differently from ordinary negatively-inclined Marshallian supply curves, even when they are geometrically identical. In the former case each point on the curve represents a maximum quantity; in the latter case each point on the curve represents a minimum quantity. Negatively-inclined supply curves of the second type for basic factors of production seem to me so improbable as to make analysis of their consequences pedantic.

Factor	Rate of Remuneration Per Unit	Total Quantity of Factors Used: By Cloth Industry	By Other Industries	By All Industries	Total Payments to Factors	Industry	Output: Units	Input per Unit of Output ("Technical Coefficients"): A	B	C	Average Cost = Price
Case I											
A	$4.00	1	29	30	$120	Cloth	4	1/4	5/4	2/4	$8.00
B	4.00	5	15	20	80	Other	36	29/36	15/36	28/36	8.00
C	4.00	2	28	30	120	All	40				8.00
				80	$320						
Case II											
A	$3.40	6	24	30	$102	Cloth	8	6/8	8/8	5/8	$10.30
B	5.50	8	12	20	110	Other	30	24/20	12/30	25/30	7.92.
C	3.60	5	25	30	108	All	38				8.42+
				80	$320						

money income. The "law" will operate unambiguously, however, only after the expanding industry has reached the stage beyond which there are no net technological or efficiency advantages of increasing the scale of plants in order to increase output of the particular commodities concerned as compared to increasing the number of plants, i.e., where there are no "net technological economies of large-scale production." When this stage is ordinarily reached is a question of fact, but I know of no convincing evidence that the optimum-efficiency size, as measured by the ratio of optimum-plant-capacity to total output of the industry, is not quite moderate for any industry of appreciable size outside the fields of transportation and communication, where "plant" is difficult to define.

THE THEORY OF ECONOMIC DEVELOPMENT

An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle

JOSEPH A. SCHUMPETER

TRANSLATED BY

REDVERS OPIE

OXFORD UNIVERSITY PRESS
LONDON OXFORD NEW YORK

CHAPTER II

THE FUNDAMENTAL PHENOMENON OF ECONOMIC DEVELOPMENT

I

THE social process, which rationalises [1] our life and thought, has led us away from the metaphysical treatment of social development and taught us to see the possibility of an empirical treatment; but it has done its work so imperfectly that we must be careful in dealing with the phenomenon itself, still more with the concept in which we comprehend it, and most of all with the word by which we designate the concept and whose associations might lead us astray in all manner of undesirable directions. Closely connected with the metaphysical preconception — more precisely with the ideas which grow out of metaphysical roots and become preconceptions if, neglecting unbridgeable gulfs, we make them do the work of empirical science — even if not itself such a metaphysical preconception, is every search for a "meaning" of history. The same is true of the postulate that a nation, a civilisation, or even the whole of mankind, must show some kind of uniform unilinear development, as even such a matter-of-fact mind as Roscher assumed and as the innumerable philosophers and theorists of history in the long brilliant line from Vico to Lamprecht took and still take for granted. Here, too, belong all kinds of evolutionary thought that centre in Darwin — at least if this means no more than reasoning by analogy — and also the psychological prejudice which consists in seeing more in motives and acts of volition than a reflex of the social process. But the evolutionary idea is now discredited in our field, especially with his-

[1] This is used here in Max Weber's sense. As the reader will see, "rational" and "empirical" here mean, if not identical, yet cognate, things. They are equally different from, and opposed to, "metaphysical," which implies going beyond the reach of both "reason" and "facts," beyond the realm, that is, of science. With some it has become a habit to use the word "rational" in much the same sense as we do "metaphysical." Hence some warning against misunderstanding may not be out of place.

torians and ethnologists, for still another reason. To the reproach of unscientific and extra-scientific mysticism that now surrounds the "evolutionary" ideas, is added that of dilettantism. With all the hasty generalisations in which the word "evolution" plays a part, many of us have lost patience.

We must get away from such things. Then two facts still remain: first the fact of historical change, whereby social conditions become historical "individuals" in historical time. These changes constitute neither a circular process nor pendulum movements about a centre. The concept of social development is defined by these two circumstances, together with the other fact: that whenever we do not succeed in adequately explaining a given historical state of things from the preceding one, we do indeed recognise the existence of an unsolved but not insoluble problem. This holds good first of all for the individual case. For example, we understand Germany's internal political history in 1919 as one of the effects of the preceding war. It also holds good, however, for more general problems.

Economic development is so far simply the object of economic history, which in turn is merely a part of universal history, only separated from the rest for purposes of exposition. Because of this fundamental dependence of the economic aspect of things on everything else, it is not possible to explain *economic* change by previous *economic* conditions alone. For the economic state of a people does not emerge simply from the preceding economic conditions, but only from the preceding total situation. The expository and analytical difficulties which arise from this are very much diminished, practically if not in principle, by the facts which form the basis of economic interpretation of history; without being compelled to take a stand for or against this view, we can state that the economic world is relatively autonomous because it takes up such a great part of a nation's life, and forms or conditions a great part of the remainder; wherefore writing economic history by itself is obviously a different thing from writing, say, military history. To this must be added still another fact which facilitates the separate description of any of the divisions of the social process. Every sector of social life is, as it were, in-

habited by a distinct set of people. The heteronomous elements generally do not affect the social process in any such sector directly as the bursting of a bomb "affects" all things which happen to be in the room in which it explodes, but only through its data and the conduct of its inhabitants; and even if an event occurs like the one suggested by our metaphor of a bursting bomb, the effects only occur in the particular garb with which those primarily concerned dress them. Therefore, just as describing the effects of the Counter Reformation upon Italian and Spanish painting always remains history of art, so describing the economic process remains economic history even where the true causation is largely non-economic.

The economic sector, again, is open to an endless variety of points of view and treatments, which one can array, for example, according to the breadth of their scope — or we might just as well say according to the degree of generalisation which they imply. From an exposition of the nature of the economic life of the Niederaltaich monastery in the thirteenth century to Sombart's exposition of the development of economic life in western Europe, there runs a continuous, logically uniform thread. Such an exposition as Sombart's is theory, and indeed theory of economic development in the sense in which we intend it for the moment. But it is not economic theory in the sense in which the contents of the first chapter of this book are economic theory, which is what has been understood by "economic theory" since Ricardo's day. Economic theory in the latter sense, it is true, plays a part in a theory like Sombart's, but a wholly subordinate one: namely, where the connection of historical facts is complicated enough to necessitate methods of interpretation which go beyond the analytic powers of the man in the street, the line of thought takes the form offered by that analytical apparatus. However, where it is simply a question of making development or the historical outcome of it intelligible, of working out the elements which characterise a situation or determine an issue, economic theory in the traditional sense contributes next to nothing.[1]

[1] If economists, nevertheless, have always had something to say on this theme, this is only because they did not restrict themselves to economic theory, but — and

We are not concerned here with a theory of development in this sense. No historical evolutionary factors will be indicated — whether individual events like the appearance of American gold

indeed quite superficially as a rule — studied historical sociology or made assumptions about the economic future. Division of labor, the origin of private property in land, increasing control over nature, economic freedom, and legal security — these are the most important elements constituting the "economic sociology" of Adam Smith. They clearly relate to the social framework of the economic course of events, not to any immanent spontaneity of the latter. One can also consider this as Ricardo's theory of development (say in Bücher's sense), which, moreover, exhibits the line of thought which earned the characterisation of "pessimist" for him: namely the "hypothetical prognosis" that in consequence of the progressive increase of population together with the progressive exhaustion of the powers of the soil (which can according to him only temporarily be interrupted by improvements in production) a position of rest would eventually appear — to be distinguished *toto coelo* from the ideal momentary position of rest of the equilibrium of modern theory — in which the economic situation would be characterised by an hypertrophy of rent, which is something totally different from what is understood above by a theory of development, and still more different from what we shall understand by it in this book. Mill worked out the same line of thought more carefully, and also distributed color and tone differently. In essence, however, his Book IV, "Influence of the Progress of Society on Production and Distribution," is just the same thing. Even this title expresses how much "progress" is considered as something non-economic, as something rooted in the data that only "exercises an influence" upon production and distribution. In particular his treatment of improvements in the "arts of production" is strictly "static." Improvement, according to this traditional view, is something which just happens and the effects of which we have to investigate, while we have nothing to say about its occurrence *per se*. What is thereby passed over is the subject matter of this book, or rather the foundation stone of its construction. J. B. Clark (Essentials of Economic Theory), whose merit is in having consciously separated "statics" and "dynamics," saw in the dynamic elements a disturbance of the static equilibrium. This is likewise our view, and also from our standpoint an essential task is to investigate the effect of this disturbance and the new equilibrium which then emerges. But while he confines himself to this and just like Mill sees therein the meaning of dynamics, we shall first of all give a theory of these causes of disturbances in so far as they are more than mere disturbances for us and in so far as it seems to us that essential economic phenomena depend upon their appearance. In particular: two of the causes of disturbance enumerated by him (increase of capital and population) are for us, as for him, merely causes of disturbance, however important as "factors of change" they may be for another kind of problem just indicated in the text. The same is true of a third (changes in the direction of consumers' tastes) which will later be substantiated in the text. But the other two (changes in technique and in productive organisation) require special analysis and evoke something different again from disturbances in the theoretical sense. The non-recognition of this is the most important single reason for what appears unsatisfactory to us in economic theory. From this insignificant-looking source flows, as we shall see, a new conception of the economic process, which overcomes a series of fundamental difficulties and thus justifies the new statement of the problem in the text. This statement of the problem is more nearly parallel to that of Marx. For according to him there is an *internal* economic development and no mere adaptation of economic life to changing data. But my structure covers only a small part of his ground.

production in Europe in the sixteenth century, or "more general" circumstances like changes in the mentality of economic men, in the area of the civilised world, in social organisation, in political constellations, in productive technique, and so forth — nor will their effects be described for individual cases or for groups of cases.[1] On the contrary, the economic theory the nature of which was sufficiently expounded to the reader in the first chapter will simply be improved for its own purposes, by building onto it. If this were also to enable this theory to perform better than hitherto its service to the other kind of theory of development, the fact would still remain that the two methods lie in different planes.

Our problem is as follows. The theory of the first chapter describes economic life from the standpoint of a "circular flow," running on in channels essentially the same year after year — similar to the circulation of the blood in an animal organism. Now this circular flow and its channels do alter in time, and here we abandon the analogy with the circulation of the blood. For although the latter also changes in the course of the growth and decline of the organism, yet it only does so continuously, that is by steps which one can choose smaller than any assignable quantity, however small, and always within the same framework. Economic life experiences such changes too, but it also experiences others which do not appear continuously and which change the framework, the traditional course itself. They cannot be understood by means of any analysis of the circular flow, although they are purely economic and although their explanation is obviously among the tasks of pure theory. Now such changes

[1] Therefore one of the most annoying misunderstandings that arose out of the first edition of this book was that this theory of development neglects all historical factors of change except one, namely the individuality of entrepreneurs. If my representation were intended to be as this objection assumes, it would obviously be nonsense. But it is not at all concerned with the concrete factors of change, but with the method by which these work, with the *mechanism of change*. The "entrepreneur" is merely the bearer of the mechanism of change. And I have taken account not of one factor of historical change, but of none. We have still less to do here with the factors which in particular explain the changes in the economic organisation, economic custom, and so on. This is still another problem, and although there are points at which all these methods of treatment collide, it means spoiling the fruit of all if they are not kept apart and if each is not allowed the right to grow by itself.

and the phenomena which appear in their train are the object of our investigation. But we do not ask: what changes of this sort have actually made the modern economic system what it is? nor: what are the conditions of such changes? We only ask, and indeed in the same sense as theory always asks: how do such changes take place, and to what economic phenomena do they give rise?

The same thing may be put somewhat differently. The theory of the first chapter describes economic life from the standpoint of the economic system's tendency towards an equilibrium position, which tendency gives us the means of determining prices and quantities of goods, and may be described as an adaptation to data existing at any time. In contrast to the conditions of the circular flow this does not mean in itself that year after year "the same" things happen; for it only means that we conceive the several processes in the economic system as partial phenomena of the tendency towards an equilibrium position, but not necessarily towards the same one. The position of the ideal state of equilibrium in the economic system, never attained, continually "striven after" (of course not consciously), changes, because the data change. And theory is not weaponless in the face of these changes in data. It is constructed so as to be able to deal with the consequences of such changes; it has special instruments for the purpose (for example the instrument called quasi-rent). If the change occurs in the non-social data (natural conditions) or in non-economic social data (here belong the effects of war, changes in commercial, social, or economic policy), or in consumers' tastes, then to this extent no fundamental overhaul of the theoretical tools seems to be required. These tools only fail — and here this argument joins the preceding — where economic life itself changes its own data by fits and starts. The building of a railway may serve as an example. Continuous changes, which may in time, by continual adaptation through innumerable small steps, make a great department store out of a small retail business, come under the "static" analysis. But "static" analysis is not only unable to predict the consequences of discontinuous changes in the traditional way of doing things; it can neither explain the occurrence of such productive revolutions nor the phenomena which accompany them. It can only investigate the new equilibrium position after the changes have occurred. It is just this occurrence of the "revolutionary" change that is our problem, the problem of economic development in a very narrow and formal sense. The reason why we so state the problem and turn aside from traditional theory lies not so much in the fact that economic changes, especially, if not solely, in the capitalist epoch, have actually occurred thus and not by continuous adaptation, but more in their fruitfulness.[1]

By "development," therefore, we shall understand only such changes in economic life as are not forced upon it from without but arise by its own initiative, from within. Should it turn out that there are no such changes arising in the economic sphere itself, and that the phenomenon that we call economic development is in practice simply founded upon the fact that the data change and that the economy continuously adapts itself to them, then we should say that there is *no* economic development. By this we should mean that economic development is not a phenomenon to be explained economically, but that the economy, in itself without development, is dragged along by the changes in the surrounding world, that the causes and hence the explanation of the development must be sought outside the group of facts which are described by economic theory.

Nor will the mere growth of the economy, as shown by the growth of population and wealth, be designated here as a process of development. For it calls forth no qualitatively new phenomena, but only processes of adaptation of the same kind as the changes in the natural data. Since we wish to direct our attention to other phenomena, we shall regard such increases as changes in data.[2]

[1] The problems of capital, credit, entrepreneurial profit, interest on capital, and crises (or business cycles) are the ones in which this fruitfulness will be demonstrated here. Yet it is not thereby exhausted. For the expert theorist I point, for example, to the difficulties which surround the problem of increasing return, the question of multiple points of intersection between supply and demand curves, and the element of time, which even Marshall's analysis has not overcome.

[2] We do this because these changes are small per annum and therefore do not stand in the way of the applicability of the "static" method. Nevertheless, their appearance is frequently a condition of development in our sense. But even though they often make the latter possible, yet they do not create it out of themselves.

Every concrete process of development finally rests upon preceding development. But in order to see the essence of the thing clearly, we shall abstract from this and allow the development to arise out of a position without development. Every process of development creates the prerequisites for the following. Thereby the form of the latter is altered, and things will turn out differently from what they would have been if every concrete phase of development had been compelled first to create its own conditions. However, if we wish to get at the root of the matter, we may not include in the data of our explanation elements of what is to be explained. But if we do not do this, we shall create an apparent discrepancy between fact and theory, which may constitute an important difficulty for the reader.

If I have been more successful than in the first edition in concentrating the exposition upon essentials and in guarding against misunderstandings, then further special explanations of the words "static" and "dynamic," with their innumerable meanings, are not necessary. Development in our sense is a distinct phenomenon, entirely foreign to what may be observed in the circular flow or in the tendency towards equilibrium. It is spontaneous and discontinuous change in the channels of the flow, disturbance of equilibrium, which forever alters and displaces the equilibrium state previously existing. Our theory of development is nothing but a treatment of this phenomenon and the processes incident to it.[1]

[1] In the first edition of this book, I called it "dynamics." But it is preferable to avoid this expression here, since it so easily leads us astray because of the associations which attach themselves to its various meanings. Better, then, to say simply what we mean: economic life changes; it changes partly because of changes in the data, to which it tends to adapt itself. But this is not the only kind of economic change; there is another which is not accounted for by influence on the data from without, but which arises from within the system, and this kind of change is the cause of so many important economic phenomena that it seems worth while to build a theory for it, and, in order to do so, to isolate it from all the other factors of change. The author begs to add another more exact definition, which he is in the habit of using: what we are about to consider is that kind of change arising from within the system *which so displaces its equilibrium point that the new one cannot be reached from the old one by infinitesimal steps*. Add successively as many mail coaches as you please, you will never get a railway thereby.

II

These spontaneous and discontinuous changes in the channel of the circular flow and these disturbances of the centre of equilibrium appear in the sphere of industrial and commercial life, not in the sphere of the wants of the consumers of final products. Where spontaneous and discontinuous changes in consumers' tastes appear, it is a question of a sudden change in data with which the businessman must cope, hence possibly a question of a *motive* or an opportunity for other than gradual adaptations of his conduct, but not of such other conduct itself. Therefore this case does not offer any other problems than a change in natural data or require any new method of treatment; wherefore we shall neglect any spontaneity of consumers' needs that may actually exist, and assume tastes as "given." This is made easy for us by the fact that the spontaneity of wants is in general small. To be sure, we must always start from the satisfaction of wants, since they are the end of all production, and the given economic situation at any time must be understood from this aspect. Yet innovations in the economic system do not as a rule take place in such a way that first new wants arise spontaneously in consumers and then the productive apparatus swings round through their pressure. We do not deny the presence of this nexus. It is, however, the producer who as a rule initiates economic change, and consumers are educated by him if necessary; they are, as it were, taught to want new things, or things which differ in some respect or other from those which they have been in the habit of using. Therefore, while it is permissible and even necessary to consider consumers' wants as an independent and indeed the fundamental force in a theory of the circular flow, we must take a different attitude as soon as we analyse *change*.

To produce means to combine materials and forces within our reach (cf. *supra*, Chapter I). To produce other things, or the same things by a different method, means to combine these materials and forces differently. In so far as the "new combination" may in time grow out of the old by continuous adjustment in small steps, there is certainly change, possibly growth, but neither a

new phenomenon nor development in our sense. In so far as this is not the case, and the new combinations appear discontinuously, then the phenomenon characterising development emerges. For reasons of expository convenience, henceforth, we shall only mean the latter case when we speak of new combinations of productive means. Development in our sense is then defined by the carrying out of new combinations.

This concept covers the following five cases: (1) The introduction of a new good — that is one with which consumers are not yet familiar — or of a new quality of a good. (2) The introduction of a new method of production, that is one not yet tested by experience in the branch of manufacture concerned, which need by no means be founded upon a discovery scientifically new, and can also exist in a new way of handling a commodity commercially. (3) The opening of a new market, that is a market into which the particular branch of manufacture of the country in question has not previously entered, whether or not this market has existed before. (4) The conquest of a new source of supply of raw materials or half-manufactured goods, again irrespective of whether this source already exists or whether it has first to be created. (5) The carrying out of the new organisation of any industry, like the creation of a monopoly position (for example through trustification) or the breaking up of a monopoly position.

Now two things are essential for the phenomena incident to the carrying out of such new combinations, and for the understanding of the problems involved. In the first place it is not essential to the matter — though it may happen — that the new combinations should be carried out by the same people who control the productive or commercial process which is to be displaced by the new. On the contrary, new combinations are, as a rule, embodied, as it were, in new firms which generally do not arise out of the old ones but start producing beside them; to keep to the example already chosen, in general it is not the owner of stage-coaches who builds railways. This fact not only puts the discontinuity which characterises the process we want to describe in a special light, and creates so to speak still another kind of discontinuity in addition to the one mentioned above, but it also explains important features of the course of events. Especially in a competitive economy, in which new combinations mean the competitive elimination of the old, it explains on the one hand the process by which individuals and families rise and fall economically and socially and which is peculiar to this form of organisation, as well as a whole series of other phenomena of the business cycle, of the mechanism of the formation of private fortunes, and so on. In a non-exchange economy, for example a socialist one, the new combinations would also frequently appear side by side with the old. But the economic consequences of this fact would be absent to some extent, and the social consequences would be wholly absent. And if the competitive economy is broken up by the growth of great combines, as is increasingly the case to-day in all countries, then this must become more and more true of real life, and the carrying out of new combinations must become in ever greater measure the internal concern of one and the same economic body. The difference so made is great enough to serve as the water-shed between two epochs in the social history of capitalism.

We must notice secondly, only partly in connection with this element, that whenever we are concerned with fundamental principles, we must never assume that the carrying out of new combinations takes place by employing means of production which happen to be unused. In practical life, this is very often the case. There are always unemployed workmen, unsold raw materials, unused productive capacity, and so forth. This certainly is a contributory circumstance, a favorable condition and even an incentive to the emergence of new combinations; but great unemployment is only the consequence of non-economic events — as for example the World War — or precisely of the development which we are investigating. In neither of the two cases can its existence play a fundamental rôle in the explanation, and it cannot occur in a well balanced circular flow from which we start. Nor would the normal yearly increment meet the case, as it would be small in the first place, and also because it would normally be absorbed by a corresponding expansion of production within the circular flow, which, if we admit such increments, we must think of as adjusted

to this rate of growth.[1] As a rule the new combinations must draw the necessary means of production from some old combinations — and for reasons already mentioned we shall assume that they *always* do so, in order to put in bold relief what we hold to be the essential contour line. The carrying out of new combinations means, therefore, simply the different employment of the economic system's existing supplies of productive means — which might provide a second definition of development in our sense. That rudiment of a pure economic theory of development which is implied in the traditional doctrine of the formation of capital always refers merely to saving and to the investment of the small yearly increase attributable to it. In this it asserts nothing false, but it entirely overlooks much more essential things. The slow and continuous increase in time of the national supply of productive means and of savings is obviously an important factor in explaining the course of economic history through the centuries, but it is completely overshadowed by the fact that development consists primarily in employing existing resources in a different way, in doing new things with them, irrespective of whether those resources increase or not. In the treatment of shorter epochs, moreover, this is even true in a more tangible sense. Different methods of employment, and not saving and increases in the available quantity of labor, have changed the face of the economic world in the last fifty years. The increase of population especially, but also of the sources from which savings can be made, was first made possible in large measure through the different employment of the then existing means.

The next step in our argument is also self-evident: command over means of production is necessary to the carrying out of new combinations. Procuring the means of production is one distinct problem for the established firms which work within the circular flow. For they *have* them already procured or else can procure them currently with the proceeds of previous production as was explained in the first chapter. There is no fundamental gap here

[1] On the whole it is much more correct to say that population grows slowly up to the possibilities of any economic environment than that it has any tendency to outgrow it and to become thereby an independent cause of change.

between receipts and disbursements, which, on the contrary, necessarily correspond to one another just as both correspond to the means of production offered and to the products demanded. Once set in motion, this mechanism works automatically. Furthermore, the problem does not exist in a non-exchange economy even if new combinations are carried out in it; for the directing organ, for example a socialist economic ministry, is in a position to direct the productive resources of the society to new uses exactly as it can direct them to their previous employments. The new employment may, under certain circumstances, impose temporary sacrifices, privations, or increased efforts upon the members of the community; it may presuppose the solution of difficult problems, for example the question from which of the old combinations the necessary productive means should be withdrawn; but there is no question of procuring means of production not already at the disposal of the economic ministry. Finally, the problem also does not exist in a competitive economy in the case of the carrying out of new combinations, if those who carry them out have the necessary productive means or can get them in exchange for others which they have or for any other property which they may possess. This is not the privilege of the possession of property *per se*, but only the privilege of the possession of disposable property, that is such as is employable either immediately for carrying out the new combination or in exchange for the necessary goods and services.[1] In the contrary case — and this is the rule as it is the fundamentally interesting case — the possessor of wealth, even if it is the greatest combine, must resort to credit if he wishes to carry out a new combination, which cannot like an established business be financed by returns from previous production. To provide this credit is clearly the function of that category of individuals which we call "capitalists." It is obvious that this is the characteristic method of the capitalist type of society — and important enough to serve as its *differentia specifica* — for forcing the economic system into new channels, for putting its means at

[1] A privilege which the individual can also achieve through saving. In an economy of the handicraft type this element would have to be emphasised more. Manufacturers' "reserve funds" assume an existing development.

the service of new ends, in contrast to the method of a non-exchange economy of the kind which simply consists in exercising the directing organ's power to command.

It does not appear to me possible to dispute in any way the foregoing statement. Emphasis upon the significance of credit is to be found in every textbook. That the structure of modern industry could not have been erected without it, that it makes the individual to a certain extent independent of inherited possessions, that talent in economic life "rides to success on its debts," even the most conservative orthodoxy of the theorists cannot well deny. Nor is the connection established here between credit and the carrying out of innovations, a connection which will be worked out later, anything to take offence at. For it is as clear *a priori* as it is established historically that credit is primarily necessary to new combinations and that it is from these that it forces its way into the circular flow, on the one hand because it was originally necessary to the founding of what are now the old firms, on the other hand because its mechanism, once in existence, also seizes old combinations for obvious reasons.[1] First, *a priori*: we saw in the first chapter that borrowing is not a necessary element of production in the normal circular flow within accustomed channels, is not an element without which we could not understand the essential phenomena of the latter. On the other hand, in carrying out new combinations, "financing" as a special act is fundamentally necessary, in practice as in theory. Second, historically: those who lend and borrow for industrial purposes do not appear early in history. The pre-capitalistic lender provided money for other than business purposes. And we all remember the type of industrialist who felt he was losing caste by borrowing and who therefore shunned banks and bills of exchange. The capitalistic credit system has grown out of and thrived on the financing of new combinations in all countries, even though in a different way in each (the origin of German joint stock banking is especially characteristic). Finally there can be no stumblingblock in our speak-

[1] The most important of which is the appearance of productive interest, as we shall see in Chapter V. As soon as interest emerges somewhere in the system, it expands over the whole of it.

ing of receiving credit in "money or money substitutes." We certainly do not assert that one can produce with coins, notes, or bank balances, and do not deny that services of labor, raw materials, and tools are the things wanted. We are only speaking of a method of procuring them.

Nevertheless there is a point here in which, as has already been hinted, our theory diverges from the traditional view. The accepted theory sees a problem in the existence of the productive means, which are needed for new, or indeed any, productive processes, and this accumulation therefore becomes a distinct function or service. We do not recognise this problem at all; it appears to us to be created by faulty analysis. It does not exist in the circular flow, because the running of the latter presupposes given quantities of means of production. But neither does it exist for the carrying out of new combinations,[1] because the productive means required in the latter are drawn from the circular flow whether they already exist there in the shape wanted or have first to be produced by other means of production existing there. Instead of this problem another exists for us: the problem of detaching productive means (already employed somewhere) from the circular flow and allotting them to new combinations. This is done by credit, by means of which one who wishes to carry out new combinations outbids the producers in the circular flow in the market for the required means of production. And although the meaning and object of this process lies in a movement of goods from their old towards new employments, it cannot be described entirely in terms of goods without overlooking something essential, which happens in the sphere of money and credit and upon which depends the explanation of important phenomena in the capitalist form of economic organisation, in contrast to other types.

Finally one more step in this direction: whence come the sums

[1] Of course the productive means do not fall from heaven. In so far as they are not given by nature or non-economically, they were and are created at some time by the individual waves of development in our sense, and henceforth incorporated in the circular flow. But every individual wave of development and every individual new combination itself proceeds again from the supply of productive means of the existing circular flow — a case of the hen and the egg.

needed to purchase the means of production necessary for the new combinations if the individual concerned does not happen to have them? The conventional answer is simple: out of the annual growth of social savings plus that part of resources which may annually become free. Now the first quantity was indeed important enough before the war — it may perhaps be estimated as one-fifth of total private incomes in Europe and North America — so that together with the latter sum, which it is difficult to obtain statistically, it does not immediately give the lie quantitatively to this answer. At the same time a figure representing the range of all the business operations involved in carrying out new combinations is also not available at present. But we may not even start from total "savings." For its magnitude is explicable only by the results of previous development. By far the greater part of it does not come from thrift in the strict sense, that is from abstaining from the consumption of part of one's regular income, but it consists of funds which are themselves the result of successful innovation and in which we shall later recognise entrepreneurial profit. In the circular flow there would be on the one hand no such rich source, out of which to save, and on the other hand essentially less incentive to save. The only big incomes known to it would be monopoly revenues and the rents of large landowners; while provision for misfortunes and old age, perhaps also irrational motives, would be the only incentives. The most important incentive, the chance of participating in the gains of development, would be absent. Hence, in such an economic system there could be no great reservoirs of free purchasing power to which one who wished to form new combinations could turn — and his own savings would only suffice in exceptional cases. All money would circulate, would be fixed in definite established channels.

Even though the conventional answer to our question is not obviously absurd, yet there is another method of obtaining money for this purpose, which claims our attention, because it, unlike the one referred to, does not presuppose the existence of accumulated results of previous development, and hence may be considered as the only one which is available in strict logic. This method of obtaining money is the creation of purchasing power by banks. The form it takes is immaterial. The issue of bank-notes not fully covered by specie withdrawn from circulation is an obvious instance, but methods of deposit banking render the same service, where they increase the sum total of possible expenditure. Or we may think of bank acceptances in so far as they serve as money to make payments in wholesale trade. It is always a question, not of transforming purchasing power which already exists in someone's possession, but of the creation of new purchasing power out of nothing — out of nothing even if the credit contract by which the new purchasing power is created is supported by securities which are not themselves circulating media — which is added to the existing circulation. And this is the source from which new combinations *are* often financed, and from which they would have to be financed *always*, if results of previous development did not actually exist at any moment.

These credit means of payment, that is means of payment which are created for the purpose and by the act of giving credit, serve just as ready money in trade, partly directly, partly because they can be converted immediately into ready money for small payments or payments to the non-banking classes — in particular to wage-earners. With their help, those who carry out new combinations can gain access to the existing stocks of productive means, or, as the case may be, enable those from whom they buy productive services to gain immediate access to the market for consumption goods. There is never, in this nexus, granting of credit in the sense that someone must wait for the equivalent of his service in goods, and content himself with a claim, thereby fulfilling a special function; not even in the sense that someone has to accumulate means of maintenance for laborers or landowners, or produced means of production, all of which would only be paid for out of the final results of production. Economically, it is true, there is an essential difference between these means of payment, if they are created for new ends, and money or other means of payment of the circular flow. The latter may be conceived on the one hand as a kind of certificate for completed production and the increase in the social product effected through it,

and on the other hand as a kind of order upon, or claim to, part of this social product. The former have not the first of these two characteristics. They too are orders, for which one can immediately procure consumption goods, but not certificates for previous production. Access to the national dividend is usually to be had only on condition of some productive service previously rendered or of some product previously sold. This condition is, in this case, not yet fulfilled. It will be fulfilled only after the successful completion of the new combinations. Hence this credit will in the meantime affect the price level.

The banker, therefore, is not so much primarily a middleman in the commodity "purchasing power" as a *producer* of this commodity. However, since all reserve funds and savings to-day usually flow to him, and the total demand for free purchasing power, whether existing or to be created, concentrates on him, he has either replaced private capitalists or become their agent; he has himself become the capitalist par excellence. He stands between those who wish to form new combinations and the possessors of productive means. He is essentially a phenomenon of development, though only when no central authority directs the social process. He makes possible the carrying out of new combinations, authorises people, in the name of society as it were, to form them. He is the ephor of the exchange economy.

III

We now come to the third of the elements with which our analysis works, namely the "new combination of means of production," and credit. Although all three elements form a whole, the third may be described as the fundamental phenomenon of economic development. The carrying out of new combinations we call "enterprise"; the individuals whose function it is to carry them out we call "entrepreneurs." These concepts are at once broader and narrower than the usual. Broader, because in the first place we call entrepreneurs not only those "independent" businessmen in an exchange economy who are usually so designated, but all who actually fulfil the function by which we define

the concept, even if they are, as is becoming the rule, "dependent" employees of a company, like managers, members of boards of directors, and so forth, or even if their actual power to perform the entrepreneurial function has any other foundations, such as the control of a majority of shares. As it is the carrying out of new combinations that constitutes the entrepreneur, it is not necessary that he should be permanently connected with an individual firm; many "financiers," "promotors," and so forth are not, and still they may be entrepreneurs in our sense. On the other hand, our concept is narrower than the traditional one in that it does not include all heads of firms or managers or industrialists who merely may operate an established business, but only those who actually perform that function. Nevertheless I maintain that the above definition does no more than formulate with greater precision what the traditional doctrine really means to convey. In the first place our definition agrees with the usual one on the fundamental point of distinguishing between "entrepreneurs" and "capitalists" — irrespective of whether the latter are regarded as owners of money, claims to money, or material goods. This distinction is common property to-day and has been so for a considerable time. It also settles the question whether the ordinary shareholder as such is an entrepreneur, and disposes of the conception of the entrepreneur as risk bearer.[1] Furthermore, the ordinary characterisation of the entrepreneur type by such expressions as "initiative," "authority," or "foresight" points entirely in our direction. For there is little scope for such qualities within the routine of the circular flow, and if this had been sharply separated

[1] Risk obviously always falls on the owner of the means of production or of the money-capital which was paid for them, hence never on the entrepreneur *as such* (see Chapter IV). A shareholder *may* be an entrepreneur. He may even owe to his holding a controlling interest the power to act as an entrepreneur. Shareholders *per se*, however, are never entrepreneurs, but merely capitalists, who in consideration of their submitting to certain risks participate in profits. That this is no reason to look upon them as anything but capitalists is shown by the facts, first, that the average shareholder has normally no power to influence the management of his company, and secondly, that participation in profits is frequent in cases in which everyone recognises the presence of a loan contract. Compare, for example, the Graeco-Roman *foenus nauticum*. Surely this interpretation is more true to life than the other one, which, following the lead of a faulty legal construction — which can only be explained historically — attributes functions to the average shareholder which he hardly ever thinks of discharging.

from the occurrence of changes in this routine itself, the emphasis in the definition of the function of entrepreneurs would have been shifted automatically to the latter. Finally there are definitions which we could simply accept. There is in particular the well known one that goes back to J. B. Say: the entrepreneur's function is to combine the productive factors, to bring them together. Since this is a performance of a special kind only when the factors are combined for the first time — while it is merely routine work if done in the course of running a business — this definition coincides with ours. When Mataja (in Unternehmergewinn) defines the entrepreneur as one who receives profit, we have only to add the conclusion of the first chapter, that there is no profit in the circular flow, in order to trace this formulation too back to ours.[1] And this view is not foreign to traditional theory, as is shown by the construction of the *entrepreneur faisant ni bénéfice ni perte*, which has been worked out rigorously by Walras, but is the property of many other authors. The tendency is for the entrepreneur to make neither profit nor loss in the circular flow — that is he has no function of a special kind there, he simply does not exist; but in his stead, there are heads of firms or business managers of a different type which we had better not designate by the same term.

It is a prejudice to believe that the knowledge of the historical origin of an institution or of a type immediately shows us its sociological or economic nature. Such knowledge often leads us to understand it, but it does not directly yield a theory of it. Still more false is the belief that "primitive" forms of a type are also *ipso facto* the "simpler" or the "more original" in the sense that they show their nature more purely and with fewer complications than later ones. Very frequently the opposite is the case, amongst other reasons because increasing specialisation may allow functions and qualities to stand out sharply, which are more difficult to recognise in more primitive conditions when mixed with others.

[1] The definition of the entrepreneur in terms of entrepreneurial profit instead of in terms of the function the performance of which creates the entrepreneurial profit is obviously not brilliant. But we have still another objection to it: we shall see that entrepreneurial profit does not fall to the entrepreneur by "necessity" in the same sense as the marginal product of labor does to the worker.

So it is in our case. In the general position of the chief of a primitive horde it is difficult to separate the entrepreneurial element from the others. For the same reason most economists up to the time of the younger Mill failed to keep capitalist and entrepreneur distinct because the manufacturer of a hundred years ago was both; and certainly the course of events since then has facilitated the making of this distinction, as the system of land tenure in England has facilitated the distinction between farmer and landowner, while on the Continent this distinction is still occasionally neglected, especially in the case of the peasant who tills his own soil.[1] But in our case there are still more of such difficulties. The entrepreneur of earlier times was not only as a rule the capitalist too, he was also often — as he still is to-day in the case of small concerns — his own technical expert, in so far as a professional specialist was not called in for special cases. Likewise he was (and is) often his own buying and selling agent, the head of his office, his own personnel manager, and sometimes, even though as a rule he of course employed solicitors, his own legal adviser in current affairs. And it was performing some or all of these functions that regularly filled his days. The carrying out of new combinations can no more be a *vocation* than the making and execution of strategical decisions, although it is this function and not his routine work that characterises the military leader. Therefore the entrepreneur's essential function must always appear mixed up with other kinds of activity, which as a rule must be much more conspicuous than the essential one. Hence the Marshallian definition of the enterpreneur, which simply treats the entrepreneurial function as "management" in the widest meaning, will naturally appeal to most of us. We do not accept it, simply because it does not bring out what we consider to be the salient point and the only one which specifically distinguishes entrepreneurial from other activities.

[1] Only this neglect explains the attitude of many socialistic theorists towards peasant property. For smallness of the individual possession makes a difference only for the petit-bourgeois, not for the socialist. The criterion of the employment of labor other than that of the owner and his family is economically relevant only from the standpoint of a kind of exploitation theory which is hardly tenable any longer.

Nevertheless there are types — the course of events has evolved them by degrees — which exhibit the entrepreneurial function with particular purity. The "promoter," to be sure, belongs to them only with qualifications. For, neglecting the associations relative to social and moral status which are attached to this type, the promoter is frequently only an agent intervening on commission, who does the work of financial technique in floating the new enterprise. In this case he is not its creator nor the driving power in the process. However, he *may* be the latter also, and then he is something like an "entrepreneur by profession." But the modern type of "captain of industry" [1] corresponds more closely to what is meant here, especially if one recognises his identity on the one hand with, say, the commercial entrepreneur of twelfth-century Venice — or, among later types, with John Law — and on the other hand with the village potentate who combines with his agriculture and his cattle trade, say, a rural brewery, an hotel, and a store. But whatever the type, everyone is an entrepreneur only when he actually "carries out new combinations," and loses that character as soon as he has built up his business, when he settles down to running it as other people run their businesses. This is the rule, of course, and hence it is just as rare for anyone always to remain an entrepreneur throughout the decades of his active life as it is for a businessman never to have a moment in which he is an entrepreneur, to however modest a degree.

Because being an entrepreneur is not a profession and as a rule not a lasting condition, entrepreneurs do not form a social class in the technical sense, as, for example, landowners or capitalists or workmen do. Of course the entrepreneurial function will *lead* to certain class positions for the successful entrepreneur and his family. It can also put its stamp on an epoch of social history, can form a style of life, or systems of moral and aesthetic values; but in itself it signifies a class position no more than it presupposes one. And the class position which may be attained is not as such an entrepreneurial position, but is characterised as landowning or

[1] Cf. for example the good description in Wiedenfeld, Das Persönliche im modernen Unternehmertum. Although it appeared in Schmoller's Jahrbuch in 1910 this work was not known to me when the first edition of this book was published.

capitalist,[1] according to how the proceeds of the enterprise are used. Inheritance of the pecuniary result and of personal qualities may then both keep up this position for more than one generation and make further enterprise easier for descendants, but the function of the entrepreneur itself cannot be inherited, as is shown well enough by the history of manufacturing families.[1]

But now the decisive question arises: why then is the carrying out of new combinations a special process and the object of a special kind of "function"? Every individual carries on his economic affairs as well as he can. To be sure, his own intentions are never realised with ideal perfection, but ultimately his behavior is moulded by the influence on him of the results of his conduct, so as to fit circumstances which do not as a rule change suddenly. If a business can never be absolutely perfect in any sense, yet it in time approaches a relative perfection having regard to the surrounding world, the social conditions, the knowledge of the time, and the horizon of each individual or each group. New possibilities are continuously being offered by the surrounding world, in particular new discoveries are continuously being added to the existing store of knowledge. Why should not the individual make just as much use of the new possibilities as of the old, and, according to the market position as he understands it, keep pigs instead of cows, or even choose a new crop rotation, if this can be seen to be more advantageous? And what kind of special new phenomena or problems, not to be found in the established circular flow, can arise there?

While in the accustomed circular flow every individual can act promptly and rationally because he is sure of his ground and is supported by the conduct, as adjusted to this circular flow, of all other individuals, who in turn expect the accustomed activity from him, he cannot simply do this when he is confronted by a new task. While in the accustomed channels his own ability and experience suffice for the normal individual, when confronted with innovations he needs guidance. While he swims with the stream in the circular flow which is familiar to him, he swims against the

[1] On the nature of the entrepreneurial function also compare my statement in the article "Unternehmer" in the Handwörterbuch der Staatswissenschaften.

stream if he wishes to change its channel. What was formerly a help becomes a hindrance. What was a familiar datum becomes an unknown. Where the boundaries of routine stop, many people can go no further, and the rest can only do so in a highly variable manner. The assumption that conduct is prompt and rational is in all cases a fiction. But it proves to be sufficiently near to reality, if things have time to hammer logic into men. Where this has happened, and within the limits in which it has happened, one may rest content with this fiction and build theories upon it. It is then not true that habit or custom or non-economic ways of thinking cause a hopeless difference between the individuals of different classes, times, or cultures, and that, for example, the "economics of the stock exchange" would be inapplicable say to the peasants of to-day or to the craftsmen of the Middle Ages. On the contrary the same theoretical picture [1] in its broadest contour lines fits the individuals of quite different cultures, whatever their degree of intelligence and of economic rationality, and we can depend upon it that the peasant sells his calf just as cunningly and egotistically as the stock exchange member his portfolio of shares. But this holds good only where precedents without number have formed conduct through decades and, in fundamentals, through hundreds and thousands of years, and have eliminated unadapted behavior. Outside of these limits our fiction loses its closeness to reality.[2] To cling to it there also, as the traditional theory does, is to hide an essential thing and to ignore a fact which, in contrast with other deviations of our assumptions from reality, is theoretically important and the source of the explanation of phenomena which would not exist without it.

[1] The same *theoretical* picture, obviously not the same sociological, cultural, and so forth.

[2] How much this is the case is best seen to-day in the economic life of those nations, and within our civilisation in the economics of those individuals, whom the development of the last century has not yet completely drawn into its stream, for example, in the economy of the Central European peasant. This peasant "calculates"; there is no deficiency of the "economic way of thinking" (Wirtschaftsgesinnung) in him. Yet he cannot take a step out of the beaten path; his economy has not changed at all for centuries, except perhaps through the exercise of external force and influence. Why? Because the choice of new methods is not simply an element in the concept of rational economic action, nor a matter of course, but a distinct process which stands in need of special explanation.

Therefore, in describing the circular flow one must treat combinations of means of production (the production-functions) as data, like natural possibilities, and admit only small [1] variations at the margins, such as every individual can accomplish by adapting himself to changes in his economic environment, without materially deviating from familiar lines. Therefore, too, the carrying out of new combinations is a special function, and the privilege of a type of people who are much less numerous than all those who have the "objective" possibility of doing it. Therefore, finally, entrepreneurs are a special type,[2] and their behavior a special

[1] Small disturbances which may indeed, as mentioned earlier, in time add up to great amounts. The decisive point is that the businessman, if he makes them, never alters his routine. The usual case is one of small, the exception one of great (*uno actu* great), disturbances. Only in this sense is emphasis put upon "smallness" here. The objection that there can be no difference in principle between small and large disturbances is not effective. For it is false in itself, in so far as it is based upon the disregard of the principle of the infinitesimal method, the essence of which lies in the fact that one can assert of "small quantities" under certain circumstances what one cannot assert of "large quantities." But the reader who takes umbrage at the large-small contrast may, if he wishes, substitute for it the contrast adapting-spontaneous. Personally I am not willing to do this because the latter method of expression is much easier to misunderstand than the former and really would demand still longer explanations.

[2] In the first place it is a question of a type of *conduct* and of a type of *person* in so far as this conduct is accessible in very unequal measure and to relatively few people, so that it constitutes their outstanding characteristic. Because the exposition of the first edition was reproached with exaggerating and mistaking the peculiarity of this conduct, and with overlooking the fact that it is more or less open to every businessman, and because the exposition in a later paper ("Wellenbewegung des Wirtschaftslebens," Archiv für Sozialwissenschaft) was charged with introducing an intermediate type ("half-static" businessmen), the following may be submitted. The conduct in question is peculiar in two ways. First, because it is directed towards something different and signifies doing something different from other conduct. One may indeed in this connection include it with the latter in a higher unity, but this does not alter the fact that a theoretically relevant difference exists between the two, and that only one of them is adequately described by traditional theory. Secondly, the type of conduct in question not only differs from the other in its object, "innovation" being peculiar to it, but also in that it presupposes aptitudes differing *in kind* and not only in degree from those of mere rational economic behavior.

Now these aptitudes are presumably distributed in an ethically homogeneous population just like others, that is the curve of their distribution has a maximum ordinate, deviations on either side of which become rarer the greater they are. Similarly we can assume that every healthy man can sing if he will. Perhaps half the individuals in an ethically homogeneous group have the capacity for it to an average degree, a quarter in progressively diminishing measure, and, let us say, a quarter in a measure above the average; and within this quarter, through a series of continually increasing singing ability and continually diminishing number of people

problem, the motive power of a great number of significant phenomena. Hence, our position may be characterised by three corresponding pairs of opposites. First, by the opposition of two real processes: the circular flow or the tendency towards equilibrium on the one hand, a change in the channels of economic routine or a spontaneous change in the economic data arising from within the system on the other. Secondly, by the opposition of two theoretical *apparatuses*: statics and dynamics.[1] Thirdly, by the opposi-

who possess it, we come finally to the Carusos. Only in this quarter are we struck in general by the singing ability, and only in the supreme instances can it become the characterising mark of the person. Although practically all men can sing, singing ability does not cease to be a distinguishable characteristic and attribute of a minority, indeed not exactly of a type, because this characteristic — unlike ours — affects the total personality relatively little.

Let us apply this: Again, a quarter of the population may be so poor in those qualities, let us say here provisionally, of economic initiative that the deficiency makes itself felt by poverty of their moral personality, and they play a wretched part in the smallest affairs of private and professional life in which this element is called for. We recognise this type and know that many of the best clerks, distinguished by devotion to duty, expert knowledge, and exactitude, belong to it. Then comes the "half," the "normal." These prove themselves to be better in the things which even within the established channels cannot simply be "dispatched" (erledigen) but must also be "decided" (entscheiden) and "carried out" (durchsetzen). Practically all business people belong here, otherwise they would never have attained their positions; most represent a selection — individually or hereditarily tested. A textile manufacturer travels no "new" road when he goes to a wool auction. But the situations there are never the same, and the success of the business depends so much upon skill and initiative in buying wool that the fact that the textile industry has so far exhibited no trustification comparable with that in heavy manufacturing is undoubtedly partly explicable by the reluctance of the cleverer manufacturers to renounce the advantage of their own skill in buying wool. From there, rising in the scale we come finally into the highest quarter, to people who are a type characterised by super-normal qualities of intellect and will. Within this type there are not only many varieties (merchants, manufacturers, financiers, etc.) but also a continuous variety of degrees of intensity in "initiative." In our argument types of every intensity occur. Many a one can steer a safe course, where no one has yet been; others follow where first another went before; still others only in the crowd, but in this among the first. So also the great political leader of every kind and time is a type, yet not a thing unique, but only the apex of a pyramid from which there is a continuous variation down to the average and from it to the sub-normal values. And yet not only is "leading" a special function, but the leader also something special, distinguishable — wherefore there is no sense in our case in asking: "Where does that type begin then?" and then to exclaim: "This is no type at all!"

[1] It has been objected against the first edition that it sometimes defines "statics" as a theoretical construction, sometimes as the picture of an actual state of economic life. I believe that the present exposition gives no ground for this opinion. "Static" theory does not assume a stationary economy; it also treats of the effects of changes in data. In itself, therefore, there is no necessary connection between static theory and stationary reality. Only in so far as one can exhibit the fundamental form of the

tion of two types of conduct, which, following reality, we can picture as two types of individuals: mere managers and entrepreneurs. And therefore the "best method" of producing in the theoretical sense is to be conceived as "the most advantageous among the methods which have been empirically tested and become familiar." But it is not the "best" of the methods "possible" at the time. If one does not make this distinction, the concept becomes meaningless and precisely those problems remain unsolved which our interpretation is meant to provide for.

Let us now formulate precisely the characteristic feature of the conduct and type under discussion. The smallest daily action embodies a huge mental effort. Every schoolboy would have to be a mental giant, if he himself had to create all he knows and uses by his own individual activity. And every man would have to be a giant of wisdom and will, if he had in every case to create anew all the rules by which he guides his everyday conduct. This is true not only of those decisions and actions of individual and social life the principles of which are the product of tens of thou-

economic course of events with the maximum simplicity in an unchanging economy does this assumption recommend itself to theory. The stationary economy is for uncounted thousands of years, and also in historical times in many places for centuries, an incontrovertible fact, apart from the fact, moreover, which Sombart emphasised, that there is a tendency towards a stationary state in every period of depression. Hence it is readily understood how this historical fact and that theoretical construction have allied themselves in a way which led to some confusion. The words "statics" and "dynamics" the author would not now use in the meaning they carry above, where they are simply short expressions for "theory of the circular flow" and "theory of development." One more thing: theory employs two methods of interpretation, which may perhaps make difficulties. If it is to be shown how all the elements of the economic system are determined in equilibrium by one another, this equilibrium system is considered as not yet existing and is built up before our eyes *ab ovo*. This does not mean that its coming into being is genetically explained thereby. Only its existence and functioning are made logically clear by mental dissection. And the experiences and habits of individuals are assumed as existing. How just these productive combinations have come about is not thereby explained. Further, if two contiguous equilibrium positions are to be investigated, then sometimes (not always), as in Pigou's Economics of Welfare, the "best" productive combination in the first is compared with the "best" in the second. And this again need not, but may, mean that the two combinations in the sense meant here differ not only by small variations in quantity but in their whole technical and commercial structure. Here too the coming into being of the second combination and the problems connected with it are not investigated, but only the functioning and the outcome of the already existing combination. Even though justified as far as it goes, this method of treatment passes over our problem. If the assertion were implied that this is also settled by it, it would be false.

sands of years, but also of those products of shorter periods and of a more special nature which constitute the particular instrument for performing vocational tasks. But precisely the things the performance of which according to this should involve a supreme effort, in general demand no special individual effort at all; those which should be especially difficult are in reality especially easy; what should demand superhuman capacity is accessible to the least gifted, given mental health. In particular within the ordinary routine there is no need for leadership. Of course it is still necessary to set people their tasks, to keep up discipline, and so forth; but this is easy and a function any normal person can learn to fulfil. Within the lines familiar to all, even the function of directing other people, though still necessary, is mere "work" like any other, comparable to the service of tending a machine. All people get to know, and are able to do, their daily tasks in the customary way and ordinarily perform them by themselves; the "director" has his routine as they have theirs; and his directive function serves merely to correct individual aberrations.

This is so because all knowledge and habit once acquired becomes as firmly rooted in ourselves as a railway embankment in the earth. It does not require to be continually renewed and consciously reproduced, but sinks into the strata of subconsciousness. It is normally transmitted almost without friction by inheritance, teaching, upbringing, pressure of environment. Everything we think, feel, or do often enough becomes automatic and our conscious life is unburdened of it. The enormous economy of force, in the race and the individual, here involved is not great enough, however, to make daily life a light burden and to prevent its demands from exhausting the average energy all the same. But it is great enough to make it possible to meet the ordinary claims. This holds good likewise for economic daily life. And from this it follows also for economic life that every step outside the boundary of routine has difficulties and involves a new element. It is this element that constitutes the phenomenon of leadership.

The nature of these difficulties may be focussed in the following three points. First, outside these accustomed channels the individual is without those data for his decisions and those rules of

conduct which are usually very accurately known to him within them. Of course he must still foresee and estimate on the basis of his experience. But many things must remain uncertain, still others are only ascertainable within wide limits, some can perhaps only be "guessed." In particular this is true of those data which the individual strives to alter and of those which he wants to create. Now he must really to some extent do what tradition does for him in everyday life, viz. consciously plan his conduct in every particular. There will be much more conscious rationality in this than in customary action, which as such does not need to be reflected upon at all; but this plan must necessarily be open not only to errors greater in degree, but also to other kinds of errors than those occurring in customary action. What has been done already has the sharp-edged reality of all the things which we have seen and experienced; the new is only the figment of our imagination. Carrying out a new plan and acting according to a customary one are things as different as making a road and walking along it.

How different a thing this is becomes clearer if one bears in mind the impossibility of surveying exhaustively all the effects and counter-effects of the projected enterprise. Even as many of them as could in theory be ascertained if one had unlimited time and means must practically remain in the dark. As military action must be taken in a given strategic position even if all the data potentially procurable are not available, so also in economic life action must be taken without working out all the details of what is to be done. Here the success of everything depends upon intuition, the capacity of seeing things in a way which afterwards proves to be true, even though it cannot be established at the moment, and of grasping the essential fact, discarding the unessential, even though one can give no account of the principles by which this is done. Thorough preparatory work, and special knowledge, breadth of intellectual understanding, talent for logical analysis, may under certain circumstances be sources of failure. The more accurately, however, we learn to know the natural and social world, the more perfect our control of facts becomes; and the greater the extent, with time and progressive

rationalisation, within which things can be simply calculated, and indeed quickly and reliably calculated, the more the significance of this function decreases. Therefore the importance of the entrepreneur type must diminish just as the importance of the military commander has already diminished. Nevertheless a part of the very essence of each type is bound up with this function.

As this first point lies in the task, so the second lies in the psyche of the businessman himself. It is not only objectively more difficult to do something new than what is familiar and tested by experience, but the individual feels reluctance to it and would do so even if the objective difficulties did not exist. This is so in all fields. The history of science is one great confirmation of the fact that we find it exceedingly difficult to adopt a new scientific point of view or method. Thought turns again and again into the accustomed track even if it has become unsuitable and the more suitable innovation in itself presents no particular difficulties. The very nature of fixed habits of thinking, their energy-saving function, is founded upon the fact that they have become subconscious, that they yield their results automatically and are proof against criticism and even against contradiction by individual facts. But precisely because of this they become drag-chains when they have outlived their usefulness. So it is also in the economic world. In the breast of one who wishes to do something new, the forces of habit rise up and bear witness against the embryonic project. A new and another kind of effort of will is therefore necessary in order to wrest, amidst the work and care of the daily round, scope and time for conceiving and working out the new combination and to bring oneself to look upon it as a real possibility and not merely as a day-dream. This mental freedom presupposes a great surplus force over the everyday demand and is something peculiar and by nature rare.

The third point consists in the reaction of the social environment against one who wishes to do something new. This reaction may manifest itself first of all in the existence of legal or political impediments. But neglecting this, any deviating conduct by a member of a social group is condemned, though in greatly varying degrees according as the social group is used to such conduct

or not. Even a deviation from social custom in such things as dress or manners arouses opposition, and of course all the more so in the graver cases. This opposition is stronger in primitive stages of culture than in others, but it is never absent. Even mere astonishment at the deviation, even merely noticing it, exercises a pressure on the individual. The manifestation of condemnation may at once bring noticeable consequences in its train. It may even come to social ostracism and finally to physical prevention or to direct attack. Neither the fact that progressive differentiation weakens this opposition — especially as the most important cause of the weakening is the very development which we wish to explain — nor the further fact that the social opposition operates under certain circumstances and upon many individuals as a stimulus, changes anything in principle in the significance of it. Surmounting this opposition is always a special kind of task which does not exist in the customary course of life, a task which also requires a special kind of conduct. In matters economic this resistance manifests itself first of all in the groups threatened by the innovation, then in the difficulty in finding the necessary cooperation, finally in the difficulty in winning over consumers. Even though these elements are still effective to-day, despite the fact that a period of turbulent development has accustomed us to the appearance and the carrying out of innovations, they can be best studied in the beginnings of capitalism. But they are so obvious there that it would be time lost for our purposes to dwell upon them.

There is leadership *only* for these reasons — leadership, that is, as a special kind of function and in contrast to a mere difference in rank, which would exist in every social body, in the smallest as in the largest, and in combination with which it generally appears. The facts alluded to create a boundary beyond which the majority of people do not function promptly by themselves and require help from a minority. If social life had in all respects the relative immutability of, for example, the astronomical world, or if mutable this mutability were yet incapable of being influenced by human action, or finally if capable of being so influenced this type of action were yet equally open to everyone, then there would be

no special function of leadership as distinguished from routine work.

The specific problem of leadership arises and the leader type appears only where new possibilities present themselves. That is why it is so strongly marked among the Normans at the time of their conquests and so feebly among the Slavs in the centuries of their unchanging and relatively protected life in the marshes of the Pripet. Our three points characterise the nature of the *function* as well as the *conduct* or behavior which constitutes the leader type. It is no part of his function to "find" or to "create" new possibilities. They are always present, abundantly accumulated by all sorts of people. Often they are also generally known and being discussed by scientific or literary writers. In other cases, there is nothing to discover about them, because they are quite obvious. To take an example from political life, it was not at all difficult to see how the social and political conditions of France at the time of Louis XVI could have been improved so as to avoid a breakdown of the *ancien régime*. Plenty of people as a matter of fact did see it. But nobody was in a position to *do* it. Now, it is this "doing" the thing," without which possibilities are dead, of which the leader's function consists. This holds good of all kinds of leadership, ephemeral as well as more enduring ones. The former may serve as an instance. What is to be done in a casual emergency is as a rule quite simple. Most or all people may see it, yet they want someone to speak out, to lead, and to organise. Even leadership which influences merely by example, as artistic or scientific leadership, does not consist simply in finding or creating the new thing but in so impressing the social group with it as to draw it on in its wake. It is, therefore, more by will than by intellect that the leaders fulfil their function, more by "authority," "personal weight," and so forth than by original ideas.

Economic leadership in particular must hence be distinguished from "invention." As long as they are not carried into practice, inventions are economically irrelevant. And to carry any improvement into effect is a task entirely different from the inventing of it, and a task, moreover, requiring entirely different kinds of aptitudes. Although entrepreneurs of course *may* be inventors

just as they may be capitalists, they are inventors not by nature of their function but by coincidence and vice versa. Besides, the innovations which it is the function of entrepreneurs to carry out need not necessarily be any inventions at all. It is, therefore, not advisable, and it may be downright misleading, to stress the element of invention as much as many writers do.

The entrepreneurial kind of leadership, as distinguished from other kinds of economic leadership such as we should expect to find in a primitive tribe or a communist society, is of course colored by the conditions peculiar to it. It has none of that glamour which characterises other kinds of leadership. It consists in fulfilling a very special task which only in rare cases appeals to the imagination of the public. For its success, keenness and vigor are not more essential than a certain narrowness which seizes the immediate chance and *nothing else*. "Personal weight" is, to be sure, not without importance. Yet the personality of the capitalistic entrepreneur need not, and generally does not, answer to the idea most of us have of what a "leader" looks like, so much so that there is some difficulty in realizing that he comes within the sociological category of leader at all. He "leads" the means of production into new channels. But this he does, not by convincing people of the desirability of carrying out his plan or by creating confidence in his leading in the manner of a political leader — the only man he has to convince or to impress is the banker who is to finance him — but by buying them or their services, and then using them as he sees fit. He also leads in the sense that he draws other producers in his branch after him. But as they are his competitors, who first reduce and then annihilate his profit, this is, as it were, leadership against one's own will. Finally, he renders a service, the full appreciation of which takes a specialist's knowledge of the case. It is not so easily understood by the public at large as a politician's successful speech or a general's victory in the field, not to insist on the fact that he seems to act — and often harshly — in his individual interest alone. We shall understand, therefore, that we do not observe, in this case, the emergence of all those affective values which are the glory of all other kinds of social leadership. Add to this the precariousness of the

economic position both of the individual entrepreneur and of entrepreneurs as a group, and the fact that when his economic success raises him socially he has no cultural tradition or attitude to fall back upon, but moves about in society as an upstart, whose ways are readily laughed at, and we shall understand why this type has never been popular, and why even scientific critique often makes short work of it.[1]

We shall finally try to round off our picture of the entrepreneur in the same manner in which we always, in science as well as in practical life, try to understand human behavior, viz. by analysing the characteristic motives of his conduct. Any attempt to do this must of course meet with all those objections against the economist's intrusion into "psychology" which have been made familiar by a long series of writers. We cannot here enter into the fundamental question of the relation between psychology and economics. It is enough to state that those who on principle object to *any* psychological considerations in an economic argument may leave out what we are about to say without thereby losing contact with the argument of the following chapters. For none of the results to which our analysis is intended to lead stands or falls with our "psychology of the entrepreneur," or could be vitiated by any errors in it. Nowhere is there, as the reader will easily satisfy himself, any necessity for us to overstep the frontiers of observable behavior. Those who do not object to *all* psychology but only to the *kind* of psychology which we know from the traditional textbook, will see that we do not adopt any part of the time-honored picture of the motivation of the "economic man."

In the theory of the circular flow, the importance of examining motives is very much reduced by the fact that the equations of the system of equilibrium may be so interpreted as not to imply any psychic magnitudes at all, as shown by the analysis of Pareto

[1] It may, therefore, not be superfluous to point out that our analysis of the rôle of the entrepreneur does not involve any "glorification" of the type, as some readers of the first edition of this book seemed to think. We do hold that entrepreneurs *have* an economic function as distinguished from, say, robbers. But we neither style every entrepreneur a genius or a benefactor to humanity, nor do we wish to express any opinion about the comparative merits of the social organisation in which he plays his rôle, or about the question whether what he does could not be effected more cheaply or efficiently in other ways.

and of Barone. This is the reason why even very defective psychology interferes much less with results than one would expect. There may be rational *conduct* even in the absence of rational *motive*. But as soon as we really wish to penetrate into motivation, the problem proves by no means simple. Within given social circumstances and habits, most of what people do every day will appear to them primarily from the point of view of duty carrying a social or a superhuman sanction. There is very little of conscious rationality, still less of hedonism and of *individual* egoism about it, and so much of it as may safely be said to exist is of comparatively recent growth. Nevertheless, as long as we confine ourselves to the great outlines of constantly repeated economic action, we may link it up with wants and the desire to satisfy them, on condition that we are careful to recognise that economic motive so defined varies in intensity very much in time; that it is society that shapes the particular desires we observe; that wants must be taken with reference to the group which the individual thinks of when deciding his course of action — the family or any other group, smaller or larger than the family; that action does not promptly follow upon desire but only more or less imperfectly corresponds to it; that the field of individual choice is always, though in very different ways and to very different degrees, fenced in by social habits or conventions and the like: it still remains broadly true that, within the circular flow, everyone adapts himself to his environment so as to satisfy certain *given* wants — of himself or others — as best he can. In *all* cases, the *meaning* of economic action is the satisfaction of wants in the sense that there would be no economic action if there were no wants. In the case of the circular flow, we may also think of satisfaction of wants as the normal *motive*.

The latter is not true for our type. In one sense, he may indeed be called the most rational and the most egotistical of all. For, as we have seen, conscious rationality enters much more into the carrying out of new plans, which themselves have to be worked out before they can be acted upon, than into the mere running of an established business, which is largely a matter of routine. And the typical entrepreneur is more self-centred than other types,

because he relies less than they do on tradition and connection and because his characteristic task — theoretically as well as historically — consists precisely in breaking up old, and creating new, tradition. Although this applies primarily to his economic action, it also extends to the moral, cultural, and social consequences of it. It is, of course, no mere coincidence that the period of the rise of the entrepreneur type also gave birth to Utilitarianism.

But his conduct and his motive are "rational" in no other sense. And in *no* sense is his characteristic motivation of the hedonist kind. If we define hedonist motive of action as the wish to satisfy one's wants, we may indeed make "wants" include any impulse whatsoever, just as we may define egoism so as to include all altruistic values too, on the strength of the fact that they also mean something in the way of self-gratification. But this would reduce our definition to tautology. If we wish to give it meaning, we must restrict it to such wants as are capable of being satisfied by the consumption of goods, and to that kind of satisfaction which is expected from it. Then it is no longer true that our type is acting on a wish to satisfy his wants.

For unless we assume that individuals of our type are driven along by an insatiable craving for hedonist satisfaction, the operations of Gossen's law would in the case of business leaders soon put a stop to further effort. Experience teaches, however, that typical entrepreneurs retire from the arena only when and because their strength is spent and they feel no longer equal to their task. This does not seem to verify the picture of the economic man, balancing probable results against disutility of effort and reaching in due course a point of equilibrium beyond which he is not willing to go. Effort, in our case, does not seem to weigh at all in the sense of being felt as a reason to stop. And activity of the entrepreneurial type is obviously an obstacle to hedonist enjoyment of those kinds of commodity which are usually acquired by incomes beyond a certain size, because their "consumption" presupposes leisure. Hedonistically, therefore, the conduct which we usually observe in individuals of our type would be irrational.

This would not, of course, prove the absence of hedonistic motive. Yet it points to another psychology of non-hedonist character, especially if we take into account the indifference to hedonist enjoyment which is often conspicuous in outstanding specimens of the type and which is not difficult to understand.

First of all, there is the dream and the will to found a private kingdom, usually, though not necessarily, also a dynasty. The modern world really does not know any such positions, but what may be attained by industrial or commercial success is still the nearest approach to medieval lordship possible to modern man. Its fascination is specially strong for people who have no other chance of achieving social distinction. The sensation of power and independence loses nothing by the fact that both are largely illusions. Closer analysis would lead to discovering an endless variety within this group of motives, from spiritual ambition down to mere snobbery. But this need not detain us. Let it suffice to point out that motives of this kind, although they stand nearest to consumers' satisfaction, do not coincide with it.

Then there is the will to conquer: the impulse to fight, to prove oneself superior to others, to succeed for the sake, not of the fruits of success, but of success itself. From this aspect, economic action becomes akin to sport — there are financial races, or rather boxing-matches. The financial result is a secondary consideration, or, at all events, mainly valued as an index of success and as a symptom of victory, the displaying of which very often is more important as a motive of large expenditure than the wish for the consumers' goods themselves. Again we should find countless nuances, some of which, like social ambition, shade into the first group of motives. And again we are faced with a motivation characteristically different from that of "satisfaction of wants" in the sense defined above, or from, to put the same thing into other words, "hedonistic adaptation."

Finally, there is the joy of creating, of getting things done, or simply of exercising one's energy and ingenuity. This is akin to a ubiquitous motive, but nowhere else does it stand out as an independent factor of behavior with anything like the clearness with which it obtrudes itself in our case. Our type seeks out difficulties,

changes in order to change, delights in ventures. This group of motives is the most distinctly anti-hedonist of the three.

Only with the first groups of motives is private property as the result of entrepreneurial activity an essential factor in making it operative. With the other two it is not. Pecuniary gain is indeed a very accurate expression of success, especially of *relative* success, and from the standpoint of the man who strives for it, it has the additional advantage of being an objective fact and largely independent of the opinion of others. These and other peculiarities incident to the mechanism of "acquisitive" society make it very difficult to replace it as a motor of industrial development, even if we would discard the importance it has for creating a fund ready for investment. Nevertheless it is true that the second and third groups of entrepreneurial motives may in principle be taken care of by other social arrangements not involving private gain from economic innovation. What other stimuli could be provided, and how they could be made to work as well as the "capitalistic" ones do, are questions which are beyond our theme. They are taken too lightly by social reformers, and are altogether ignored by fiscal radicalism. But they are not insoluble, and may be answered by detailed observation of the psychology of entrepreneurial activity, at least for given times and places.

A course in microeconomic theory

David M. Kreps

PRINCETON UNIVERSITY PRESS

And once we have assumption $\star$, we arrive at the conclusion: In a long-run equilibrium, firms make zero profits and produce at efficient scale.

Of course, the rest of the assumptions of the theory hang together very nicely if we don't make assumption $\star$. Think, for example, of the wheat farming industry. We might well consider wheat to be a commodity good with a single, well-known price, which all wheat farmers and wheat buyers take as given. Still, some wheat farmers are lucky enough to own land that is especially good for growing wheat; these farmer-firms have a "technology" that is superior to the technology owned by farmer-firms who farm less fertile land. So we expect those farmers to make strictly positive profits.

When we don't make assumption $\star$, so it is possible that firms can make positive profits in a long-run equilibrium, we also don't expect those firms to be producing at efficient scale. In fact, they won't do so; these firms will always produce at larger-than-efficient scale. You are asked to prove this in problem 2.

In such cases, one can (and some authors do) define things so that, even though assumption $\star$ is not made, profits still equal zero in a long-run equilibrium. This is done by saying that what a firm with a "productive advantage" earns are not profits, but rather *rents* to its advantaged production capabilities. These rents come off the top as costs to be paid for the advantaged technology, so that the firm is back to making zero profits. (If you like this sort of proof-by-definition, consider why this means that the firm now is producing at the point of minimal LRAC.)

An example

To clarify some of the points made, consider the following simple example. The numbers in this example are selected because they give relatively clean and simple answers; they are not meant to be at all realistic.

In a particular economy, a product called *pfillip*, which is a nonnarcotic stimulant, is produced by a competitive industry. Each firm in this competitive industry has the same production technology, given by the production function

$$y = k^{1/6} l^{1/3},$$

where y is the amount of pfillip produced, k is the amount of *kapitose* (a specialty chemical) used in production, and l is the amount of *legume* (a common vegetable) used in production. Firms also incur fixed costs of \$1/6.

Legumes are traded in a competitive market, and the price of a unit of legume is a fixed \$1, regardless of the amount of legume demanded for the production of pfillip. The level of legume can be freely varied in the short run. Kapitose is traded in a competitive market, at price \$1/2. The amount of kapitose used by any firm in production cannot be varied in the short run but can be adjusted in the intermediate run.

Many firms could enter this industry in the long run, and firms are free to depart. All firms, both those in the industry and potential entrants, have the technology and cost structure just described.

Demand is given by the demand function $D(p) = 400 - 100p$, where p is the price of pfillip and $D(p)$ is the amount demanded at this price.[5]

What is the long-run equilibrium in this perfectly competitive market?

As a first step, let us compute the total cost function of each firm in this industry, where we assume that the firm can vary both its level of kapitose and legume inputs.

The cost-minimizing way to produce y units of output is the solution of

$$\min_{k,l} \frac{k}{2} + l, \text{ subject to } k^{1/6} l^{1/3} \geq y.$$

It is clear that the solution will have the constraint binding. (We should put in nonnegativity constraints, although they will not bind and so we've left them out.) Since the constraint will bind, we can solve for k in terms of l to obtain $k = y^6/l^2$, and so the problem becomes

$$\min_{l} \frac{y^6}{2l^2} + l.$$

The first-order condition is

$$\frac{y^6}{l^3} = 1, \quad \text{or} \quad l = y^2 \text{ and } k = y^2,$$

[5] We will use models with linear demand throughout this book, and it should always be understood that demand is nonnegative; that is, $D(p)$ is really $\max\{400 - 100p, 0\}$ in this example. We also should be careful about demand at a price of zero; what is important is that revenue at a price of zero is zero.

so that the long-run total cost for producing y is

$$TC(y) = \frac{3}{2}y^2 + \frac{1}{6}. \quad (*)$$

From this we can easily find the long-run equilibrium price: It must be the minimum value of average cost. Average costs are given by $AC(y) = (3/2)y + 1/(6y)$, which is minimized at $3/2 = 1/(6y^2)$, or $y = 1/3$. That is, producing 1/3 of a unit is the efficient scale for firms in this industry. And when firms produce 1/3 of a unit, they have average costs of \$1. So the long-run equilibrium price is \$1.

At this price, demand is $400 - 100(1) = 300$, so total industry demand is 300. And since each firm will be producing 1/3 of a unit, there will have to be 900 active firms.

To summarize, at a long-run equilibrium 900 firms will be active, each producing 1/3 of a unit, for a total industry supply of 300 units. The equilibrium price will be \$1, and each firm will be making zero profits. Each of the 900 firms will be utilizing 1/9 unit each of legume and kapitose.

Now suppose the demand curve suddenly shifts, becoming $D(p) = 750 - 150p$. What will be the industry response in the short run, in the long run, and in the intermediate run?

We will take the short run first. In the short run, there are 900 firms, each of whom is fixed at 1/9 units of kapitose. If any of these firms wishes to have output level y, they must employ enough legume l so that

$$(1/9)^{1/6} l^{1/3} = y \text{ or } (1/9)l^2 = y^6 \text{ or } l = 3y^3.$$

This gives total variable costs of $3y^3$. Hence the short-run marginal cost function of the firm is $MC(y) = 9y^2$. And so, at price p, a single firm supplies the amount $y(p)$ that solves $9y(p)^2 = p$ or $y(p) = \sqrt{p}/3$. Since there are 900 identical firms, industry supply at price p is 900 times this, or $S(p) = 300\sqrt{p}$. The short-run equilibrium price is where short-run supply equals demand:

$$300\sqrt{p} = 750 - 150p.$$

If you solve this equation for p, you get (approximately) $p = 2.1$.

The short-run equilibrium price is \$2.10. At this price, total industry demand is $750-150(2.1) = 435$, *which is divided among the 900 firms so that each firm produces approximately .483 units. This requires each firm to employ approximately .338 units of legume. Each firm must also pay the cost of the fixed 1/9 unit of kapitose and the 1/6 fixed cost, for total costs of* $.338 + (1/9)(1/2) + 1/6 = .56$ *against which each firm makes revenue of* $(2.1)(.483) = 1.014$, *for a profit per firm of .454.*

Moving to the intermediate run, we still have 900 firms active, but now each has total cost function $(3/2)y^2 + 1/6$, or marginal costs $MC(y) = 3y$. Hence at price p, each firm supplies (in the intermediate run) $p/3$ units, and the intermediate-run industry supply curve is $300p$. Intermediate-run equilibrium is where $300p = 750 - 150p$ or $p = 5/3 = 1.667$. So we conclude:

The intermediate-run equilibrium price is \$1.667. At this price, total industry demand and supply equal 500 units, and each firm produces 5/9 of a unit. This costs the firm $(3/2)(5/9)^2+1/6 = 34/54$, *against revenues of* $(5/3)(5/9) = 25/27$, *for a net profit of* $16/54 =$ \$.296. (We leave it to you to work out kapitose and legume utilization per firm.)

And the long run is easiest of all. We already know that at the long-run equilibrium price must be \$1 and each firm must produce 1/3 of a unit. So,

The long-run equilibrium price is \$1, which means that industry demand is for 600 units. Each firm produces 1/3 of a unit, so the number of firms doubles to 1,800. Each firm makes zero profits, and each firm utilizes 1/9 unit each of legume and kapitose.

We draw these "dynamics" in figure 8.3. We show there the original demand curve and the long-run supply curve in solid lines. Note that the long-run supply curve is horizontal as it must be in this sort of industry. The short-run and intermediate-run supply curves beginning from the initial equilibrium position are drawn in as dashed lines; note that the short-run supply curve has the steepest slope. We also draw in the "new" demand curve as a dashed line, and we mark the stages of industry response to the shift in demand: (1) marks the original equilibrium, (2) the new short-run equilibrium, (3) the new intermediate-run equilibrium, and (4) the new long-run equilibrium. Note that equilibrium quantities rise at each stage, and prices jump way up and then move down. Also note (not from the picture) the pattern of kapitose and legume usage by one of the originally active 900 firms. In the short run legume usage rises, then

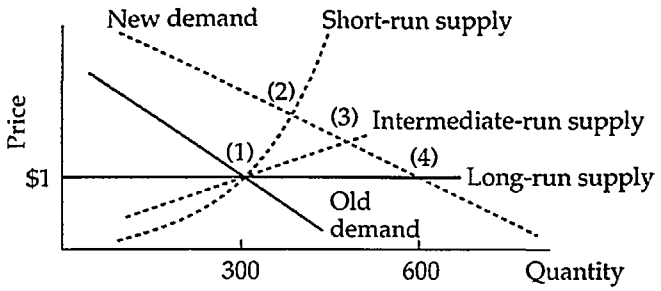

Figure 8.3. Equilibrium "dynamics" for a perfectly competitive industry.

it falls in the intermediate and long runs. Kapitose usage per firm stays constant in the short run (as it is constrained to do), then rises, and then falls.

The qualitative features of this simple example generalize substantially: In an industry where assumption ⋆ concerning potential entrants holds (and factor prices don't change with the scale of industry output; cf. the next section), long-run supply will necessarily be perfectly elastic (flat), so shifts in demand will not change the long-run equilibrium price for the good. Moreover, as asserted in section 8.2, short-run supply will generally be more inelastic than intermediate-run supply, at least for small changes in quantity, for a fixed number of firms. Hence the qualitative features of figure 8.3 can be expected to hold generally (if the story about all these runs is correct); an upwards shift in demand will be met in the short run by a small increase in quantity and a large increase in price, in the intermediate run by a further increase in quantity and a decrease in price (still above the original level), and in the long run with price retreating to the original level and output (and the number of firms) expanded to meet this increased demand.

Other changes can lead to other dynamics. For example, in problem 6 you are asked to sketch out what happens if one of the factor prices suddenly changes. Those dynamics can also be generalized beyond simple parametric examples, at least in cases where cost and demand curves are assumed to be "typical."

Such exercises illustrate the formidable power of partial equilibrium analysis in general and partial equilibrium analysis of perfect competition in particular. If cost and demand curves are "typical," which is to say demand is downward sloping and supply is nondecreasing, then the theory predicts a unique equilibrium. (Compare with general equilibrium. And, when we get to it, compare this with the predictions made using the methods of part III.) Insofar as we can empirically estimate demand functions

and cost functions, we can construct the theoretical industry supply curve and get very good tests of the theory. Moreover, this runs version of the theory makes very definite predictions about how equilibrium prices and quantities will respond dynamically to shifts in parameters of the model, again giving us a lot of empirical leverage. While our example is only an example, it indicates how powerful the theory can be.

8.3. What's wrong with partial equilibrium analysis?

At the same time, the theory is predicated on a number of assumptions that may not prove true, particularly assumptions that are required by the narrow focus of a partial equilibrium model.

For example, in a partial equilibrium analysis of the sort we have been conducting, we hold "fixed" various things left out of the model. It is typical in a partial equilibrium analysis of the market in a single product to say that one is holding fixed the prices of all commodities not in the market. For example, if we were more explicit about the demand side of our markets, we would look at consumer demand functions in the spirit of chapter 2. And it would be typical to analyze how demand for the commodity in question changes, holding fixed the price of other commodities and holding fixed the incomes the consumers have to spend. Rendered in symbols, if $x_k^i(p_1,\ldots,p_K,Y^i)$ is the demand function of consumer i for good k as a function of all prices and the consumer's income Y^i, the usual practice is to think of the industry demand curve as arising from fixing all prices except p_k, say at levels $\hat{p}_{k'}$, fixing all the Y^i, and then, for each price p_k for good k, saying that

$$D(p_k) = \sum_{i=1}^{I} x_k^i(\hat{p}_1,\ldots,\hat{p}_{k-1},p_k,\hat{p}_{k+1},\ldots,\hat{p}_K,Y^i),$$

where $i = 1,\ldots,I$ in the sum indexes the set of consumers.

In many applications, this would be the wrong thing to do, if you were really interested in industry demand. Suppose the commodity being considered is wheat. Of course, the demand for wheat depends on other factors, such as the price of corn. Should we write down the demand curve for wheat, holding fixed the price of corn? This is what is typically done and what the scheme above advises. But is it the right thing to do? As the price of wheat changes, the level of demand for corn will change. It is natural to expect that demand for corn rises as the price of wheat rises. Unless the supply of corn is perfectly elastic (is perfectly flat), the price of

corn shifts. And this then will change how much wheat will be demanded. In the short run at least, it makes little sense to suppose that the supply of corn is perfectly elastic. And so it makes little sense to suppose that the demand curve for wheat, at least in the short run, should be computed on the basis of an unchanging price of corn. *Changes in the price of wheat will, through the workings of markets, change the price of corn, which then affects the demand for wheat.* Predictions on the demand for wheat (as the price of wheat changes) that are based on a fixed price of corn will, therefore, be wrong.

If you are analyzing the market in a commodity where changes in its price (and the corresponding level of demands) won't much change the prices of other goods, then a demand curve computed under the hypothesis that other prices don't change at all won't be too far wrong. But unless this condition holds for the commodity you are interested in, you will want to think about whether there might be other prices that move substantially as the price of the good in question moves, with feedback to the demand for the good in question. If other prices can reasonably be expected to move with movements in the price of the good in question, then you must analyze more than the single market, if reasonable predictions are to emerge.

The same sort of consideration arises on the supply side of the market. Our analysis in section 8.2, for example, was predicated on an assumption that the prices of factors of production don't change. We might think to justify this assumption by appealing to the notion that many firms are in the industry, and so each is a price taker in all the factor markets. But this is no justification at all. As we change industry supply levels, we aren't looking at changes in demand for a given factor caused by a single firm; instead we are looking at changes in demand for the factor caused by changes in the activities of many firms in the industry. If this factor of production is used by many firms in the industry, if demand for this factor by firms in the industry is a large part of the demand for the factor, and if supply of the factor is not perfectly elastic, then it is wrong to build *industry* supply curves based on the supposition that the prices of the factors don't change with changes in the level of *industry* supply.

An elaborate example

Let us illustrate by continuing with the example from the previous section. We will continue to suppose that the price of legume is $1 per unit, no matter how many units of legume are bought by producers in the pfillip industry. But we suppose that the price of kapitose changes with changes in pfillip industry demand for kapitose. In particular, we suppose that if the

pfillip industry demands K units of kapitose, the price per unit that kapitose will command is $K/200$.

Why might this happen? We might suppose that kapitose is produced by an industry of perfectly competitive firms, but entry to this industry is blocked on legal grounds — say only a given set of firms are licensed to produce kapitose. If entry to the industry is restricted for this or any other reason, then even when firms in the industry are price takers, the industry supply curve may be rising; see problem 4. Or the kapitose industry might be perfectly competitive and with free entry, but potential entrants have progressively less efficient technologies for producing kapitose; see problem 5.[6]

When kapitose was assumed to trade at \$.50 per unit regardless of demand for kapitose from pfillip producers, the long-run supply curve of pfillip was perfectly elastic at price \$1. We now proceed to compute the long-run supply curve of pfillip under the newly presumed conditions. This gets a bit involved, but try to persevere.

Suppose the price of kapitose is q at some point. Each individual pfillip producer is (assumed to be) a price taker, so each computes long-run total costs of producing y units, given this factor price, as

$$\min_{k,l} qk + l + \frac{1}{6} \text{ subject to } k^{1/6}l^{1/3} \geq y.$$

If you work through the math, you will find that this leads to a total cost function

$$TC(y) = \frac{3}{2}(2q)^{1/3}y^2 + \frac{1}{6},$$

and an average cost function

$$AC(y) = \frac{3}{2}(2q)^{1/3}y + \frac{1}{6y}.$$

Efficient scale for a firm with these average costs is the solution of $(3/2)(2q)^{1/3} = 1/(6y^2)$; if you manipulate the algebra, you will find that this is

$$y^e(q) = \frac{1}{3(2q)^{1/6}},$$

where the superscript e stands for *efficient*. At this efficient scale, the level of average costs is

$$AC^e(q) = \frac{3}{2}(2q)^{1/3}\frac{1}{3(2q)^{1/6}} + \frac{1}{6/[3(2q)^{1/6}]} = (2q)^{1/6}.$$

[6] Yet another possibility is that the kapitose industry is not competitive at all. In the problems at the end of chapter 9, you will be asked to consider the case of a monopoly kapitose producer. This situation is harder to deal with because the notion of a kapitose supply curve is not well defined in a monopoly situation.

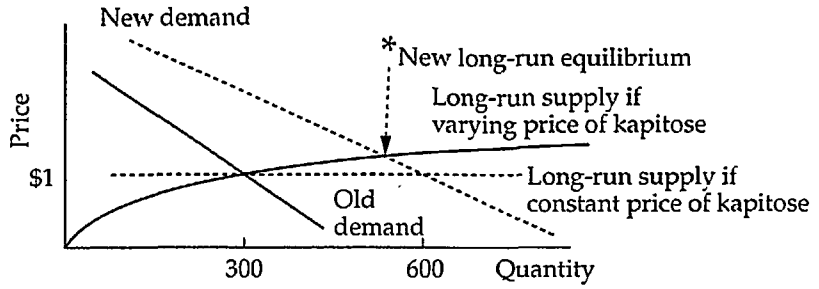

Figure 8.4. Supply and long-run equilibrium in the pfillip market when the price of kapitose rises with increasing demand for kapitose by pfillip producers.

(Don't be lazy; go through all the steps in these math derivations!) Hence if q is the long-run equilibrium price of kapitose, $(2q)^{1/6}$ is the long-run equilibrium price of pfillip. Turning this around, if p is the long-run price of pfillip, the equilibrium price of kapitose must be $q(p) = p^6/2$.

At the same time, if $q = p^6/2$ is the price of kapitose and if each firm in the pfillip industry is producing optimally at the efficient scale of $y^e = 1/[3(2q)^{1/6}] = 1/(3p)$, each firm is using $k^e = y^2/(2q)^{2/3} = 1/(18q) = 1/(9p^6)$ units of kapitose. (You need to work out the solution to the cost minimization problem above to see how we did this.)

This may seem counterintuitive to you. We have found that the level of production of the single firm is a *decreasing* function of the price of pfillip! But this is not counterintuitive as long as you remember that we are speaking here of equilibrium values. The only way the price of pfillip could be higher in a long-run equilibrium is if the price of kapitose is higher. And a higher price of kapitose lowers the efficient scale of all the firms.

Suppose long-run industry supply of pfillip is S when the price of pfillip is p. Since each firm is producing $1/(3p)$, this means we have $J = 3pS$ firms. Each of these firms uses $1/(9p^6)$ units of kapitose, so total kapitose utilization by the pfillip industry firms is

$$K = \frac{J}{9p^6} = \frac{3pS}{9p^6} = \frac{S}{3p^5}.$$

This causes the price of kapitose to be

$$\frac{p^6}{2} = q = \frac{K}{200} = \frac{S}{200(3p^5)}.$$

Solving for S in terms of p yields

$$S(p) = 300p^{11}.$$

This is the long-run equilibrium supply curve for pfillip.

In figure 8.4 we've drawn what we've discovered. The two demand curves are the two curves from the previous example: the original $D(p) = 400 - 100p$; and the shifted demand curve $D(p) = 750 - 150p$. We've dashed in the long-run supply curve of pfillip under the conditions that the price of kapitose is always \$.50; this is perfectly elastic supply at $p = 1$ since, if the price of kapitose never changes, the long-run average cost curves of firms are always the same (at any scale of industry production), and so minimum average cost is always \$1. And we've drawn in the long-run industry supply curve for the case where the price of kapitose rises with kapitose demand; this is the curve $S(p)$ derived above.

Note that the two supply curves intersect at $p = 1$, at the level $S(p) = 300$, which just happens to be the equilibrium quantity for both with the original demand curve. This is no coincidence; the numbers were picked so this would happen.[c] The point is that at higher levels of long-run equilibrium price, the long-run industry supply of pfillip is less than when kapitose costs a flat \$.50. Hence if demand shifts in the fashion of the example and the price of kapitose rises with kapitose usage, the new long-run equilibrium, marked with a $\star$ in the figure, has a price higher than \$1 and an equilibrium quantity less than the 600 we had before. (It is left to you to compute the new long-run equilibrium prices and quantities. Since this involves an eleventh degree polynomial, you will probably have to resort to numerical approximation!)

This derivation may seem quite involved to you. It is. But it has to be. We are solving for equilibria in two markets at once. We can't tell what the price of kapitose will be until we know how much is demanded by the pfillip producers. But we can't tell how much is demanded by pfillip producers until we know the price of kapitose. And we are relying on the pfillip market being in a long-run equilibrium. Putting all that together is not an easy exercise. (Now try problem 3.)

One point should be stressed about the example. The individual firm is always acting as a price taker. When we solved the individual firm's cost minimization problem, we took the price of kapitose to be an unchanging q. As we noted at the start of the chapter, this isn't quite right. As even a single firm increases its demand for kapitose, the price of kapitose rises. But this is no worse than what happens in the individual firm's profit-maximization problem, given price p. Given downward sloping demand, as a single firm raises its level of output, it depresses the market price however slightly, something else our price-taking firms are ignoring.

8.4. General equilibrium with firms

Since perfectly competitive markets may be tied together as in the previous section, we might try to put competitive firms into a general equilibrium analysis.

[c] It wasn't hard to do. Knowing that the first equilibrium has $p = 1$, $q = .5$, $Y = 300$, and industry kapitose utilization of $K = 100$, a supply curve for kapitose was selected so that at $K = 100$, q would be .5.