Positive Feedbacks in the Economy [§]

by

W. Brian Arthur [*]

26 November 1989

[§]Published in *Scientific American*, 262, 92-99, Feb. 1990.

[*] Morrison Professor of Population Studies and Economics, Stanford University, Food Research Institute, 311 W.Encina Hall, Stanford, Calif. 94305

Positive Feedbacks in the Economy

Conventional economic theory is built on the assumption of diminishing returns. Economic actions eventually engender a negative feedback that leads to a predictable equilibrium for prices and market shares. Negative feedback tends to stabilize the economy because any major changes will be offset by the very reactions they generate. The high oil prices of the 1970's encouraged energy conservation and increased oil exploration, precipitating a predictable drop in prices by 198x. According to conventional theory the equilibrium marks the "best" outcome possible under the circumstances: the most efficient use and allocation of resources.

Such an agreeable picture often does violence to reality. In many parts of the economy stabilizing forces appear not to operate. Instead, positive feedback magnifies the effect of small economic shifts; the economic models that describe such effects differ vastly from the conventional ones. Diminishing returns imply a single equilibrium point for the economy, but positive feedback—increasing returns—make for multiple equilibrium points. There is no guarantee that the particular economic outcome selected from among the many alternatives will be the "best" one. Furthermore, once chance economic forces select a particular path, it may become locked in regardless of the advantages of other paths. If one product or nation in a competitive marketplace gets ahead by "chance" it tends to stay ahead and even increase its lead. Predictable, shared markets are no longer guaranteed.

In the last few years I and other economic theorists at Stanford, the Santa Fe Institute, and elsewhere have been developing a view of the economy based on positive feedbacks. Increasing-returns economics has roots in economic thinking that go back for seventy or more years, but its application to the economy as a whole is largely new. The theory has strong parallels with modern non-linear physics (instead of the pre-20th century physical models that underlie conventional economics), it requires new and challenging mathematical techniques, and it appears to be the appropriate theory for understanding modern high-technology economies.

The history of video-cassette recorders furnishes a simple example of positive feedback. The VCR market started out with two competing formats selling at about the same price: VHS and Beta. Each format is subject to increasing returns with increasing market share: large numbers of VHS recorders encourage video outlets to stock more prerecorded tapes in VHS format, thus increasing the value of owning a VHS recorder and leading more people to buy one. (The same would, of course, be true for Beta-format players.) As the two systems compete a small lead in market share may enhance the competitive position of one of the systems and help it further increase its lead.

Such a market is initially unstable. Both systems were introduced about the same time and thus started with roughly equal market shares, but those shares fluctuated early on due to external circumstance, "luck" and other actions by companies maneuvering for position. Increasing returns on early gains eventually tilted the competition toward VHS: it accumulated enough of an advantage to take essentially the entire VCR market. However, it would have been impossible at the outset of the competition to say in advance which system would win. Furthermore, if the claim the Beta was technically superior is true, then the market's choice does not represent the best economic outcome.

Conventional economic theory would predict a different result for competition between two technologies or products performing the same function. An example is the competition between water power and coal to drive electrical generators. As hydroelectric plants take more of the market, engineers must exploit more costly dam sites, thus increasing the chance that a coal-fired plant will be cheaper. As coal plants take more of the market they bid up the price of coal (or trigger the imposition of costly pollution controls), thus tipping the balance toward hydro. The two end up sharing the market in a predictable proportion that best exploits the potentials of each, in contrast to what happens for video recorders.

The evolution of the VCR market would not have surprised the great Victorian economist Alfred Marshall, one of the founders of today's conventional economics. In his 1890 Principles of Economics he noted that if firms' production costs fall as their market shares increase, a firm that by good fortune gained a high proportion of the market early on would be able to best its rivals; "whichever firm first gets off to a good start" would corner the market. Marshall did not follow up this observation, however, and theoretical economics in this century has until recently largely ignored it.

Marshall did not believe that increasing returns applied everywhere in the economy; agriculture and mining—the mainstays of economies of his time—were subject to diminishing returns caused by limited amounts of fertile land or high-quality ore deposits. Manufacturing, on other hand, enjoyed increasing returns because large plants allowed improved organization. Modern economists do not see economies of scale as a reliable source of increasing returns. Sometimes large plants have proved more economical; often they have not.

I would update Marshall's insight by observing that the parts of the economy that are resource-based (agriculture, bulk-goods production, mining) are still for the most part subject to diminishing returns. Here conventional economics rightly holds sway. The parts of the economy that are knowledge-based, on the other hand, are largely subject to increasing returns. Products such as computers, pharmaceuticals, missiles, aircraft, automobiles, software, telecommunications equipment or fiber optics are complicated to design and to manufacture. They require large initial investments in research, development and tooling, but once sales begin incremental production is relatively cheap. A new airframe or aircraft engine, for example, typically costs between $2 and 3 billion to design, develop, certify, and put into production. Each copy thereafter costs perhaps $50 to 100 million. Unit costs fall and profits increase as more units are built.

Furthermore, increased production brings additional benefits: producing more units means gaining more experience in the manufacturing process and understanding how to produce additional units even more cheaply. Moreover, experience gained with one product or technology can make it easier for a company to produce new products incorporating similar or related technologies. Japan, for example, leveraged an initial investment in building precision instruments into a facility for building consumer electronics products and then the integrated circuits that went into them.

Not only do the costs of producing high-technology products fall as a company makes more of them, the benefits of using them increase. Many items such as computers or telecommunications equipment work in networks that require compatibility; when one brand gains significant market share (like VHS or the IBM PC) people have a strong incentive to buy more of the same product so as to be able to exchange information with those using it already. Finally, hi-tech products, unlike bulk goods, require specialized marketing and good relationships with customers. Increasing market share requires

building a network of such ties; the more extensive this network the easier further increases become.

If increasing-returns mechanisms are important, why have they been largely ignored until recently? Some would say that complicated products—high technology—for which increasing returns are so prevalent, are themselves a recent phenomenon. This is true, but only part of the answer. After all, in the 1940's and 1950's economists like Gunnar Myrdal and Nicholas Kaldor identified "cumulative causation" or positive feedback mechanisms that did not involve technology. Orthodox economists avoided increasing returns for deeper reasons.

Some economists found the existence of more than one solution to the same problem distasteful—unscientific. "Multiple equilibria," wrote Josef Schumpeter in 1954, "are not necessarily useless, but from the standpoint of any exact science the existence of a uniquely determined equilibrium is, of course, of the utmost importance, even if proof has to be purchased at the price of very restrictive assumptions; without any possibility of proving the existence of uniquely determined equilibria—or at all events, of a small number of possible equilibria—at however high a level of abstraction, a field of phenomena is really a chaos that is not under analytical control."

Other economists could see that increasing returns would destroy their familiar world of unique, predictable equilibria and along with this the notion that the market's choice was always best. Moreover, if one or a few firms came to dominate a market, the assumption of perfect competition, that no firm is large enough to affect market prices on its own (which makes economic problems easy to analyze), would also be a casualty. When John Hicks surveyed these possibilities in 1939 he drew back in alarm. "The threatened wreckage," he wrote, "is that of the greater part of economic theory." Economists restricted themselves to diminishing returns, which presented no anomalies and could be analyzed completely.

Still others were perplexed by the question of how a market would select one among several possible solutions. In Marshall's example, the initially largest firm has the lowest production costs and must inevitably win in the market. In that case, why would smaller firms compete at all? On the other hand, if by some chance a market started with several identical firms, their market shares would remained poised in an unstable equilibrium forever.

Studying such problems in 1979, I believed I could see a way out of many of these difficulties. In the real world, if several similar-sized firms entered a market together, small fortuitous events—unexpected orders, chance meetings with buyers, managerial whims—would help determine which ones achieved early sales and, over time, which firm came to dominate. Economic activity is quantized by individual transactions that are too small to foresee, and these small "random" events could cumulate and become magnified by positive feedbacks over time to determine which solution was reached. This suggested that situations dominated by increasing returns should be modeled not as static, deterministic problems, but rather as dynamic processes with random events, and with natural positive feedbacks or non-linearities. With this strategy an increasing-returns market could be recreated theoretically and watched as its corresponding process unfolded again and again. Sometimes one solution would emerge, sometimes (under identical conditions) another. It would be impossible to know in advance which of the multiple solutions would emerge in any given run, but it would be possible to record the particular set of random events leading to each solution and to study the probability that a particular solution will emerge under a certain set of initial conditions. The idea was simple and it may well have occurred to economists in the past. But making it work called for non-linear random-process theory that did not exist in their day.

Each increasing returns problem, it seemed, would need to be studied as its own random process. But many of them turned out to fit a general, non-linear probability schema. It can be pictured as follows: There is a gigantic table to which balls are added one at a time; they can be of several possible colors—white, red, green, or blue. The color of the ball to be added next is unknown, but the probability of a given color depends on the current proportions of colors on the table. If an increasing proportion of balls of a given color increases the probability of adding another ball of that color, the system will demonstrate positive feedback. The question is: given the function that assigns the probabilities given current proportions, what will be the long run proportions of colors on the table? This is like tossing a strange coin whose probability of Heads varies with the proportion of Heads tossed previously, and asking what will be the the long-run proportion of Heads.

In 1931 the mathematician George Polya had solved a very particular version of this problem where the probability of adding a color always equaled its current proportion.

Three US probability theorists, Bruce Hill, David Lane, and William Sudderth in 1980 solved a more general, non-linear version. In 1983 two Soviet probability theorists, Yuri Ermoliev and Yuri Kaniovski and I found the solution to a very general version. As balls continue to be added, we showed, the proportions of each color must settle down to a value that is a "fixed point" of the probability function—a set of values where the probabilities of adding each color equal their proportions. With increasing returns there can be several such sets. This meant that given an increasing-returns problem, we could determine the possible patterns or solutions that could emerge by solving the much easier problem of finding its sets of fixed points. With such tools economists can now define increasing returns problems with precision, identify their possible solutions and study the process by which a solution is reached. Increasing returns are no longer "a chaos that is not under analytical control."

In the real-world, the balls might represent new companies and their colors the region where they decide to settle. Suppose that firms enter an industry one by one and choose their locations so as to maximize profit. Also suppose that firms' profits increase if they are near other firms (their suppliers or customers). The geographical preferences of each firm (the intrinsic benefits it gains from being in a particular region) vary; unknown chance events determine the preferences of the next firm to enter the market.

The first firm to enter the industry picks a location based purely on geographical preference. The second firm decides based on preference modified by the benefits from locating near the first firm. The third firm is influenced by the positions of the first two firms, and on. If some location by good fortune attracts more firms than the others in the early stages of this evolution, the probability that it will attract more firms increases. Industrial concentration becomes self-reinforcing. The random historical sequence of firms entering the industry determines which pattern of regional settlement results. But the theory shows that not all patterns are possible. If the attractiveness exerted by the presence of other firms continues to rise without levelling off as more firms are added, the only possible solutions are where one region dominates and shuts out all others. If the attractiveness levels off, other solutions become possible where regions share the industry. Our new tools tell us which types of solutions can occur under which economic conditions.

Do some regions in fact amass a large proportion of an industry because of historical chance rather than geographical superiority? Santa Clara County in California (Silicon Valley) is a likely example. In the 1940's and early 1950's certain key people in the US electronics industry—William Hewlett and David Packard, the Varian brothers, William Shockley—set up shop near Stanford University; the local availability of engineers, supplies and components that these early firms helped create made location in Santa Clara extremely advantageous for the 900 or so firms that followed. If these early entrepreneurs had preferred other places the densest concentration of electronics in the country might well have been somewhere else. Not every location might have been suitable, but certainly many other university towns might have been candidates.

On a grander scale, if small events in history had been different would the pattern of cities themselves have been different? I believe the answer is yes. To the degree that certain locations were natural harbors or junction points on rivers or lakes, the pattern of cities today reflects not chance but geography. To the degree that industry and people are attracted to places where people and industry have already gathered, small chance concentrations early on may have become magnified into today's configuration of urban centers. "Chance and necessity," to use Jacques Monod's phrase, interact. Both have played crucial roles in the development of urban centers in the US and elsewhere. My Stanford colleague Paul David has argued that Chicago, in 1830 an unpromising hamlet of 60-some inhabitants on the mud and sand around Fort Dearborn, owes its current dominance in the Great Lakes region to fortuitous circumstances in the mid-1800's that interacted with positive feedbacks to industry concentration.

Different kinds of self-reinforcing mechanisms than these regional ones work in international high-tech manufacturing and trade. Countries that gain high volume and experience in a high-technology industry can reap advantages of lower cost and higher quality that may make it possible for them to shut other countries out. For example, in the early 1970's Japanese auto makers began to sell significant number of small cars in the US. As Japan gained market volume without much opposition from Detroit, its engineers and production workers gained experience, its costs fell and its products improved. These factors, together with improved sales networks, allowed the Japanese to increase its share of the US market; workers gained more experience, costs fell further, and quality improved again. Before Detroit responded in a serious way this positive-feedback loop had helped Japanese companies to make serious inroads into the US market for small cars.

Similar sequences of events took place in the markets for television sets, integrated circuits and other products. Between 1970 and 1987 US-based manufacturing companies saw their proportion of the domestic market in record players fall from 90 percent to 1 percent, in color television from 90 to 10 and in telephones from 99 to 25. It may be possible for these manufacturers to fight back, but the rules of positive feedback imply that it is much harder to recoup a market than to hold on to it in the first place.

How should countries respond to a world economy where such rules apply? Conventional recommendations for trade policy based on constant or diminishing returns tend toward low-profile approaches. They rely on the open market, discourage monopolies and leave issues such as R&D spending to companies. Their underlying assumption is that there is a fixed world price at which producers load goods onto the market and that interfering with local costs and prices by means of subsidies or tariffs is unproductive. These policies are appropriate for the diminishing-returns parts of the economy, not for the technology-based parts where increasing returns dominate.

Policies that <u>are</u> appropriate to success in high-tech production and international trade would encourage industries to be aggressive in seeking out product and process improvements. They would strengthen the national research base on which high-tech advantages are built. They would encourage firms in a single industry to pool their resources in joint ventures that share upfront costs, marketing networks, technical knowledge and compatibility conventions. And they might even extend to strategic alliances among companies in several countries to enter a complex industry that none could tackle alone. Increasing returns theory also recommends paying close attention to timing when fostering research initiatives in new industries. There is little sense in entering a market that is already close to being locked in or where there is otherwise little chance of success. Such policies are slowly being advocated and adopted in the US.

Other policies such as subsidizing and protecting new industries such as bioengineering to capture foreign markets are debatable. Dubious feedback benefits have sometimes been cited to justify government-sponsored white elephants. Furthermore, as Paul Krugman at MIT and several other economists have pointed out, one country pursuing such policies leads to retaliation by other countries subsidizing their high-technology industries. Nobody gains. Industry and trade policy under increasing returns are

currently being studied intensely. The policies countries choose will determine not only the shape of the global economy in the 1990's, but also its winners and its losers.

Increasing returns mechanisms can also cause economies—even successful ones such as the US and Japan—to become locked into inferior technology-development paths. A technology that improves slowly at first but has enormous long-term potential could easily be shut out, thus locking an economy into a path that is both inferior and difficult to escape.

Technologies typically improve as more people adopt them and gain experience with them. This link is a positive feedback loop: the more people adopt a particular technology, the more it improves, and the more incentive there is for further adoption. Where two or more technologies (like two or more products) compete to fulfill the same purpose positive feedbacks make the market for them unstable. If one technology pulls ahead, perhaps by chance, it may gain enough in development to corner the market of potential adopters. Of course, a technology that improves more rapidly as more people adopt it stands a better chance of surviving—it has a "selectional advantage." Early superiority, however, is no guarantee of long-term dominance.

In 1956, for example, there were several possible ways to construct nuclear reactors: gas cooling, light water, heavy water, even liquid sodium cooling. Robin Cowan of New York University has shown that a series of trivial circumstances locked nearly 100% of the US nuclear industry in to light water. Light-water reactors were originally adapted from a highly compact unit designed to propel the first nuclear submarine, the USS Nautilus, launched in 1954. The role of the US Navy in early reactor construction contracts, efforts by the National Security Council to get a reactor—any reactor—working on land in the wake of the 1957 Sputnik launch as well as the predilections of some key officials all acted to favor the early development of light-water reactors. Construction experience led to improved light-water designs and fixed the path of the US nuclear industry by the mid-1960's. Whether other designs would, in fact, have been superior in the long run is open to question, but much of the engineering literature suggests that high-temperature gas-cooled reactors, not light water reactors, would have been better if developed.

There are many other examples of technologies locked in by "founder effects"—early events in the history of development. In 1895, the gasoline engine was held to be the

least promising option among the many motor technologies competing to power automobiles. Gasoline was noisy, dirty and explosive, and it required complicated new parts such as carburetors, ignition systems and distributors. As late as 1904 the British transportation expert William Fletcher wrote: "unless the objectionable features of the petrol carriage can be removed, it is bound to be driven from the road by its less objectionable rival, the steam-driven vehicle of the day." But a series of circumstantial events from 1890 to 1920 gave gasoline a lead that has subsequently proved unassailable.

Another positive-feedback mechanism that acts to lock in certain technological conventions or standards is the attraction of compatibility with existing products. Although a standard itself may not improve with time, increasing adoption makes it advantageous for newcomers to a field—who must exchange information or products with those already working in the field—to fall in with the standard, be it the English language, a rail gauge, a high-definition television system, a screw thread, or a typewriter keyboard. Standards that are established early (such as the 1950's-vintage computer language Fortran) can be hard to dislodge by later ones no matter how superior the would-be successors may be. Of course, locked-in technologies are eventually replaced when a new generation of advances arrives, but history has shown that where positive feedbacks are present "survival of the fittest" is not a reliable maxim.

Until recently conventional economics texts have tended to portray the economy as something akin to a large Newtonian system, with a unique equilibrium solution preordained by patterns of mineral resources, geography, population, consumer tastes and technological possibilities. In this view, perturbations or temporary shifts—like the oil shock of 1973 or the stock-market crash of 1987—are quickly negated by the opposing forces they call into action. Given knowledge of future technological possibilities, it is possible in theory to forecast the path of the economy to high accuracy as a smoothly-shifting solution to the analytical equations governing prices and quantities of goods. History is not terribly important; it merely delivers the economy to its inevitable equilibrium position.

Positive-feedback economics on the other hand finds its parallels in modern non-linear physics. Ferromagnetic materials, spin-glasses (Scientific American, July 1989), solid-state lasers and other physical systems that consist of mutually reinforcing elements show the same properties as our economic examples: they "phase lock" into one

of many possible configurations; small perturbations at critical times influence which outcome is selected; and the chosen outcome may have higher energy—be less efficient—than other possible end-states. It finds parallels in modern evolutionary thinking too. Small events, the mutations of history, are indeed often averaged away, but once in a while they become all-important in tilting parts of the economy into new structures and patterns that are then preserved and built upon in a fresh layer of development. The economy we have inherited is in part the result of historical chance.

In this view of economics, initially identical economies with significant increasing returns sectors do not necessarily select the same paths. Instead they eventually diverge. To the extent that the small events determining an overall path remain beneath the resolution of the economist's lens, accurate forecasting of the economy's future may be theoretically, not just practically, impossible.

Steering an economy with positive feedbacks so that is chooses the best of its many possible equilibrium states requires good fortune and good timing—a feel for the moments at which beneficial change from one pattern to another is most possible. Theory can help us identify these states and times. And it can guide us in applying the right amount of effort (not too little but not too much) to dislodge locked-in structures.

The English philosopher of science Jacob Bronowski once remarked that economics has long suffered from a fatally simple structure imposed on it in the 18th century. I find it exciting that this is now changing. With the acceptance of positive feedbacks, economists' theories are beginning to portray the economy not as simple but complex, not as deterministic, predictable and mechanistic, but instead as process-dependent, organic and always evolving.