1

Introduction

The U.S. and other governments spend far more money subsidizing the production of clean energy technologies, such as electric vehicles, wind turbines, and solar cells, than they do on clean energy research and development (R&D).[1] Why? A major reason is that many believe that costs fall as a function of cumulative production in a so-called learning or experience curve, and thus stimulating demand is the best way to reduce costs. According to such a curve, product costs drop a certain percentage each time cumulative production doubles as automated manufacturing equipment is introduced and organized into flow lines.[2] Although such a learning curve does not explicitly exclude activities performed outside of a factory, the fact that learning curves link cost reductions with cumulative production focuses our attention on the production of a final product and implies that learning gained outside of a factory is either unimportant or is driven by that production.

But is this true? Are cumulative production and its associated activities in a factory the most important sources of cost reductions for clean energy or any other technology for that matter? Among other things, this book shows that most improvements in wind turbines, solar cells, and electric vehicles are being implemented outside of factories and that many of them are only indirectly related to production. Engineers and scientists are increasing the physical scale of wind turbines, increasing the efficiencies as well as reducing the material thicknesses of solar cells,[3] and improving the energy storage densities of batteries for electric vehicles, primarily in laboratories and not in factories. This suggests that increases in production volumes, particularly those of existing technologies, are less important than increases in spending on R&D (i.e., supply-side approaches)—an argument that Bill Gates[4] and other business

leaders regularly make. Although demand and thus demand-based subsidies do encourage R&D,[5] only a small portion of these subsidies will end up funding R&D activities.

Should this surprise us? Consider computers (and other electronic products such as mobile phones[6]). The implementation of automated equipment and its organization into flow lines in response to increases in production volumes are not the main reasons for the dramatic reduction in the cost of computers over the last 50 years. The cost of computers dropped primarily for the same reason that their performance rose: continuous improvements in integrated circuits (ICs). Furthermore, improvements in the cost and performance of ICs were only partly from the introduction of automated equipment and its organization into flow lines. A much more important cause was large reductions in the *scale* of transistors, memory cells, and other dimensional features, where these reductions required improvements in semiconductor-manufacturing equipment. This equipment was largely developed in laboratories, and these developments depended on advances in science; their rate of implementation depended more on calendar time (think of Moore's Law) than on cumulative production volumes of ICs.[7]

NEW QUESTIONS AND NEW APPROACHES

We need a better understanding of how improvements in cost and performance emerge and of why they emerge more for some technologies than for others, issues that are largely ignored by books on management (and economics). While most such books are about innovative managers and organizations, and their flexibility and open-mindedness, they don't help us understand why some technologies experience more improvements in cost and performance than do others. In fact, they dangerously imply that the potential for innovation is everywhere and thus all technologies have about the same potential for improvement.

Nothing can be further from the truth. ICs, magnetic disks, magnetic tape, optical discs, and fiber optics experienced what Ray Kurzweil calls "exponential improvements" in cost and performance in the second half of the 20th century, while mechanical components and products assembled from them did not.[8] Mobile phones, set-top boxes, digital televisions, the Internet, automated algorithmic trading (in hedge funds, for example), and online education also experienced large improvements over the last 20 years because they benefited from improvements in the previously mentioned technologies. A different set of technologies (e.g., steam engines, steel, locomotives, and automobiles) experienced large improvements in both cost and performance

in the 18th and 19th centuries. An understanding of why some technologies have more potential for improvements than do others is necessary for firms, governments, and organizations to make good decisions about clean energy and new technologies in general.

We also need a better understanding of how science and the characteristics of a technology determine the potential of new technologies. Although there is a large body of literature on how advances in science facilitate advances in technology in the so-called linear model of innovation,[9] many of these nuances are ignored once learning curves and cumulative production are considered. For example, improvements in solar cell efficiency and reductions in material thicknesses involve different sets of activities, and the potential for these improvements depends on the types of solar cells and on levels of scientific understanding for each type. Lumping together the cumulative production from different types of solar cells causes these critical nuances to be ignored and thus prevents us from implementing the best policies.

Part of the problem is that we don't understand what causes a time lag (often a long one) between advances in science, improvements in technology that are based on these advances, and the commercialization of technology. And without such an understanding, how can firms and governments make good decisions about clean energy? More fundamentally, how can they understand the potential for Schumpeter's so-called creative destruction and new industry formation? A new industry is defined as a set of products or services based on a new concept and/or architecture where these products or services are supplied by a new collection of firms and where their sales are significant (e.g., greater than $5 billion). According to Schumpeter, waves of new technologies (which are often based on new science) have created new industries, along with opportunities and wealth for new firms, as they have destroyed existing technologies and their incumbent suppliers.

This is a book about why specific industries emerge at certain moments in time and how improvements in technologies largely determine this timing. For example, why did the mainframe computer industry emerge in the 1950s, the personal computer (PC) industry in the 1970s, the mobile phone and automated algorithmic trading industries in the 1980s, the World Wide Web in the 1990s, and online universities in the 2000s? On the other hand, why haven't the personal flight, underwater, and space transportation industries emerged, in spite of large expectations for them in the 1960s?[10] Similarly, why haven't large electric vehicle, wind, and solar industries yet emerged, or when will such industries emerge that can exist without subsidies?

Parts of these questions concern policies and strategies. When did governments introduce the right polices and when did firms introduce the right

strategies? But parts also involve science and technology, and, as mentioned previously, they have been largely ignored by management books on technology and innovation,[11] even as the rates of scientific and technological change have accelerated and the barriers to change have fallen.[12] When was our understanding of scientific phenomena or the levels of performance and cost for the relevant technologies sufficient for industry formation to occur? We need better answers to these kinds of questions in order to complement research on government policies and firms' R&D strategies. For example, understanding the factors that impact on the timing of scientific, technical, and economic feasibility can help firms create better product and technology road maps, business models, and product introduction strategies. They can help entrepreneurs understand when they should quit existing firms and start new ones.[13] And they can help universities better teach students how to look for new business opportunities and address global problems; such problems include global warming, other environmental emissions, the world's dependency on oil and minerals from unstable regions, and the lack of clean water and affordable housing in many countries.

Some of the problems that arise when firms misjudge the timing of economic feasibility can be found in the mobile phone industry. In the early 1980s, studies concluded that mobile phones would never be widely used, while in the late 1990s studies concluded that the mobile Internet was right around the corner. Some would argue that we underestimated the importance of mobile communication, but I would argue that these studies misjudged the rate at which improvements in performance and cost would occur. The 1980s studies should have been asking what consumers would do when Moore's Law made handsets free and talk times less than 10 cents a minute. The 1990s studies should have been addressing the levels of performance and cost needed in displays, microprocessor and memory ICs, and networks before various types of mobile Internet content and applications could become technically and economically feasible.[14]

Chapters 2 and 3 (Part I) address the potential of new technologies using the concept of *technology paradigm* primarily advanced by Giovanni Dosi.[15] Few scholars or practitioners have attempted to use the technology paradigm to assess the potential of new technologies or to compare different ones.[16] One key aspect of this paradigm is geometrical scaling, which is a little-known idea initially noticed in the chemical industries (and in living organisms).[17] Part I shows how a technology paradigm can help us better understand the potential for new technologies where technologies with a potential for large improvements in cost and performance often lead to the rise of new industries. Part I and the rest of this book also show how implementing a technology and

exploiting the full potential of its technology paradigm require advances in science and improvements in components.

One reason for using the term "component" is to distinguish between components and systems in what can be called a "nested hierarchy of subsystems."[18] Systems are composed of subsystems, subsystems are composed of components, and components may be composed of various inputs including equipment and raw materials. This book will just use the terms *systems* and *components* to simplify the discussion. For example, a system for producing integrated circuits is composed of components such as raw materials and semiconductor-manufacturing equipment.

TECHNOLOGICAL DISCONTINUITIES AND A TECHNOLOGY PARADIGM

A technology paradigm can be defined at any level in a nested hierarchy of subsystems, where we are primarily interested in large changes in technologies, or what many call technological discontinuities. These are products based on a different set of concepts and/or architectures from that of existing products, and they are often defined as the start of new industries.[19] For example, the first mainframe computers, magnetic tape–based playback equipment, and transistors (like new services such as automated algorithmic trading and online universities) were based on a different set of concepts than were their predecessors: punch card equipment, phonograph records, and vacuum tubes, respectively. On the other hand, minicomputers, PCs, and various forms of portable computers only involved changes in architectures.

Building from Giovanni Dosi's characterization and using an analysis of many technologies (See the Appendix for the research methodology), Chapter 2 and the rest of this book define a technology paradigm in terms of (1) a technology's basic concepts or principles and the trade-offs that are defined by them; (2) the directions of advance within these trade-offs, where advance is defined by a technological trajectory (or more than one);[20] (3) the potential limits to trajectories and their paradigms; and (4) the roles of components and scientific knowledge in these limits.[21] Partly because this book is concerned with understanding when a new technology might offer a superior value proposition, Chapter 2 focuses on the second and third items and shows how there are four broad methods of achieving advances in performance and cost along technological trajectories: (1) improving the efficiency by which basic concepts and their underlying physical phenomena are exploited; (2) radical new processes; (3) geometrical scaling; and (4) improvements in "key" components.

In doing so, Chapter 2 shows how improvements in performance and/or price occur in a rather smooth and incremental manner over multiple generations of discontinuities. While some argue that these improvements can be represented by a series of S-curves where each discontinuity initially leads to *dramatic* improvements in performance-to-price ratios,[22] this and succeeding chapters show that such dramatic changes in the rates of improvement are relatively rare. Instead, this book's analyses suggest that there are smooth rates of improvement that can be characterized as incremental over multiple generations of technologies, and that these incremental improvements in a technological trajectory enable one to roughly understand near-term trends in performance and/or price/cost for new technologies.

GEOMETRICAL SCALING

Chapter 3 focuses on geometrical scaling as a method of achieving improvements in the performance and cost of a technology. Geometrical scaling refers to the relationship between the geometry of a technology, its scale, and the physical laws that govern it. As others describe it, the "scale effects are permanently embedded in the geometry and the physical nature of the world in which we live."[23]

As a result of geometrical scaling, some technologies benefit from either large increases (e.g., engines or wind turbines) or large reductions (e.g., ICs) in physical scale. For example, consider the pipes and reaction vessels that make up chemical plants, which benefit from increases in scale While economies of scale generally refer to amortizing a fixed cost over a large volume, at least until the capacity of a plant is reached, geometrical scaling refers to the fact that the output from pipes varies as a function of one dimension (radius) squared whereas the costs of pipes vary as a function of this dimension (radius) to the first power. Similarly, the output from a reaction vessel varies as a function of one dimension (radius) cubed whereas the costs of the reaction vessels vary as a function of one dimension (radius) squared. This is why empirical analyses have found that the costs of chemical plants rise only about two-thirds for each doubling of output and thus increases in the scale of chemical plants have led to dramatic reductions in the cost of many chemicals.[24]

Other technologies benefit from *reductions* in scale. The most well-known examples of this type of geometrical scaling can be found in ICs, magnetic disks and tape, and optical discs, where reducing the scale of transistors and storage regions has led to enormous improvements in the cost and performance of these technologies.[25] This is because reductions in scale lead to improvements in both performance and costs. For example, placing more

transistors or magnetic or optical storage regions in a certain area increases speed and functionality and reduces both the power consumption and size of the final product, which are typically considered improvements in performance for most electronic products; they also lead to lower material, equipment, and transportation costs. The combination of increased performance and reduced costs as size is reduced has led to exponential improvements in the performance-to-cost ratio of many electronic components.

Like Chapter 2, Chapter 3 and other chapters show how geometrical scaling is related to a nested hierarchy of subsystems. Chapter 3 demonstrates that benefiting from geometrical scaling in a higher-level "system" depends on improvements in lower-level supporting "components,"[26] and that large benefits from geometrical scaling in a lower-level "key component" can drive long-term improvements in the performance and cost of a higher-level "system." In the second instance, these long-term improvements may lead to the emergence of technological discontinuities in systems, particularly when the systems do not benefit from increases in scale. Part II shows how exponential improvements in ICs and magnetic storage densities led to discontinuities in computers and magnetic recording and playback equipment, as well as in semiconductors. Chapter 9 explores this for other systems.

In fact, most of the disruptive innovations covered by Clayton Christensen, who many consider to be the guru of innovation,[27] benefit from geometrical scaling (and experience exponential improvements) in either the "system" or a key "component" in the system. This suggests that there is a "supply-side" aspect to Christensen's theory of disruptive innovation that is very different from his focus on the demand side of technological change. While his theory suggests to some that large improvements in performance and costs along a technological trajectory *automatically* emerge once a product finds a low-end niche, and so finding the low-end niche is the central challenge of creating disruptive innovations,[28] Chapters 3 and 4 show how geometrical scaling explains why some low-end technological discontinuities became disruptive innovations and why these low-end technological discontinuities initially emerged. A search for potentially disruptive technologies, then, should consider the extent to which a system or a key component in it can benefit from rapid rates of improvement through, for example, geometrical scaling.

Some readers may find the emphasis on supply-side factors in Chapters 2 and 3 (Part I) to be excessive and thus may classify the author as a believer in so-called technological determinism. Nothing could be further from the truth. I recognize that there is an interaction between market needs and product designs, that increases in demand encourage investment in R&D, and that the technologies covered in this book were "socially constructed."[29] The

relevance of this social construction is partly reflected in the role of new users in many of the technological discontinuities covered in Part II, where these new users and changes in user needs can lead to the rise of new industries.[30] For example, the emergence of industries represented by microbreweries and artisanal cheeses is more the result of changes in consumer taste than of changes in technology. Some of these changes come from rising incomes that have led to the emergence of many industries serving the rich or even the super rich. When the upper 1 percent of Americans receives 25 percent of total income, many industries that cater to specialized consumer tastes will naturally appear.[31]

This book focuses on supply-side factors because industries that have the potential to significantly enhance most lives or improve overall productivity require dramatic improvements in performance and cost. As Paul Nightingale says about Giovanni Dosi's theory of technology paradigms, drawing on the research of Nathan Rosenberg and David Mowery,[32] "'Market pull' theories are misleading, not because they assume innovation processes respond to market forces but because they assume that the response is *unmediated*. As a consequence, they cannot explain why so many innovations are not forthcoming despite huge demand, nor why innovations occur at particular moments in time and in particular forms."[33] For example, the world *needs* inexpensive solar, wind, and other sources of clean energy, and large subsidies are increasing demand and R&D spending for them. But even with these large subsidies, significant improvements in cost and performance will not be forthcoming if such technologies do not have the potential for dramatic reductions in cost. And if they don't have such a potential, the world needs to look for other solutions.

A second reason for focusing on supply-side factors is that unless we understand the technological trajectories and the factors that have a direct impact on them, such as scaling, how can we accelerate the rates of improvement in cost and performance? Since much of the management literature on learning primarily focuses on the organizational processes involved with learning, it implies that organizational issues have a bigger impact on the potential for improving costs and performance than does the nature of the technology.[34] Thus, while this literature implies that solving energy and environmental problems is primarily an organizational issue, geometrical scaling and the other three methods of achieving advances in performance and cost remind us that the *potential* for improving cost and performance depends on a technology's characteristics.[35] Without a *potential* for improvements, it would be difficult for organizational learning to have a large impact on the costs and performance of a technology, no matter how innovative an organization is.

THE TIMING OF TECHNOLOGICAL DISCONTINUITIES

Chapters 4 through 6 (Part II) analyze technological discontinuities in part because these discontinuities often form the basis for new industries. For example, the first mainframe computers, minicomputers, personal computers, personal digital assistants, audiocassette players, videocassette recorders, camcorders, memory ICs, and microprocessors, as well as automated algorithmic trading and online education, are typically defined in this way. Like other discontinuities, they were based on a different set of concepts and/or architectures than were existing products. The characterization of a system's architecture is also considered important because the ability to characterize a system's concept partly depends on one's ability to characterize its potential architectures.

But what determines the timing of these discontinuities? Since the characterization of a concept or architecture and an understanding of the relevant scientific phenomenon usually precede the commercialization of a technology, we can look at the timing of technological discontinuities in relation to them. How long before the emergence of technological discontinuities were the necessary concepts and/or architectures characterized? And why is there a time lag and in many cases a long one, between a characterization of these concepts and architectures and both the commercialization and diffusion of the technology?[36]

These questions are largely ignored by academic researchers. While there is wide agreement on the descriptions and timing of specific technological discontinuities, most research focuses on the existence and reasons for incumbent failure and in doing so mostly treats these discontinuities as "bolts of lightning." For example, the product life cycle cyclical, and disruptive models of technological change do not address the sources of technological discontinuities; instead, their emphasis on incumbent failure implies that any time lag is due to management (e.g., cognitive) failure.[37]

But do we really believe that management failure for either cognitive or organizational reasons is why it took more than 100 years to implement Charles Babbage's computing machine in spite of early government funding?[38] Although Babbage defined the basic concept of the computer in the 1820s and subsequently built a prototype, general-purpose computers did not emerge until the 1940s or diffuse widely in developed countries until the 1980s. Is this time lag merely due to narrow-minded managers and policy makers, or is something else going on? More important, in combination with a theory that technological discontinuities initially experience dramatic improvements in performance and price, an emphasis on incumbent failure as the main reason

for a long time lag suggests that there are many technological discontinuities with a potential for *dramatic* improvements in performance and costs just waiting to be found. According to this logic, if only managers and policy makers could overcome their cognitive limitations, firms and governments could find new technologies that would *quickly* replace existing ones and thus solve big problems such as global warming.

This book disagrees with such an assessment and shows how the timing of many discontinuities can be analyzed. Building from research done by Nathan Rosenberg and his colleagues on the role of complementary technologies in the implementation of new ones,[39] Part II shows that insufficient components were the reason for the time lag between the identification and characterization of concepts and architectures that formed the basis of technological discontinuities and the commercialization (and diffusion) of these discontinuities.[40] Chapters 4, 5, and 6 present a detailed analysis of the discontinuities in computers, magnetic recording and playback equipment, and semiconductors, respectively. One reason for choosing these "systems" is that few argue that there were market failures for discontinuities in them, unlike the discontinuities of more "complex network" systems such as broadcasting or mobile phones, which are addressed in Part III.[41] A second reason is that there have been many discontinuities in these and related systems, and thus there are many "data points" to analyze.[42] Finally, the time lag for each discontinuity in these systems was primarily due to one or two insufficient components, which is very different from the mechanical sector, where novel combinations of components have probably played a more important role than have improvements in one or two components individually.[43] Partly because it is possible to design many of these systems in a modular way,[44] their performance was primarily driven by improvements in "key" components (which is the fourth broad method of achieving advances in a system's performance and costs), and these improvements also drove the emergence of system discontinuities.

For example, the implementation of minicomputers, personal computers, and most forms of portable computers primarily depended on improvements in one type of component, the IC, as the discontinuities were all based on concepts and architectures that had been characterized by the late 1940s.[45] Similarly, the implementation of various discontinuities in magnetic-based audio and video recording equipment primarily depended on improvements in one type of component, the magnetic recording density of tape, as these discontinuities were all based on concepts and architectures that had been characterized by the late 1950s. In other words, in spite of the increasing variety of components that can be combined in many different ways, improvements in a single type of component had a larger impact on the emergence

of these discontinuities (and on the performance of these systems) than did so-called novel combinations of multiple components (or technologies). This conclusion enables us to go beyond the role of complementary technologies in the time lag to analyze the specific levels of performance that were needed in single types of components before new systems—that is, discontinuities—could be implemented.

SYSTEMS, COMPONENTS, AND DISCONTINUITIES

Chapters 4, 5, and 6 analyze the impact of improvements in a single type of component on the emergence of discontinuities in systems in two different ways, where both of these ways are facilitated by the smooth and incremental manner in which improvements in performance and/or price have been occurring.

First, building from the role of trade-offs in technology paradigms and marketing theory,[46] these chapters show how improvements in components have changed the trade-offs that suppliers (i.e., designers) and users make when they consider systems and how this change leads to the emergence of discontinuities. Technology paradigms define a set of trade-offs between price and various dimensions of performance, which suppliers consider when they design or compare systems, while users make trade-offs between price and various dimensions of performance. In both cases, improvements in components can change the way these trade-offs are made by both suppliers and users.

Second, economists use the term "minimum threshold of performance" to refer to the performance that is necessary before users will consider purchasing a system.[47] For example, users would not purchase a PC until PCs could perform a certain number of instructions per second in order to process specific software applications. When a single type of component such as a microprocessor has a large impact on the performance of a system such as a PC, a similar threshold exists for the components in these systems. Thus, PCs could not perform a certain number of instructions per second until a microprocessor could meet certain levels of performance.

Part II draws a number of conclusions from these analyses. First, the new concepts or architectures that form the basis of discontinuities in systems were known long before the discontinuities were implemented. In other words, the characterization of concepts or architectures was usually not the bottleneck either for the discontinuities or for the creation of the industries that many of these discontinuities represent. Instead, the bottleneck was one or two types of components that were needed to implement the discontinuities. Thus, improvements in components can gradually make new types of systems (i.e., discontinuities) possible, and the thresholds of performance (and price)

that are needed in specific components before a new system is economically feasible can be analyzed.

Second, finding new customers and applications, which partly reflect heterogeneity in customer needs,[48] can reduce the minimum thresholds of performance for the components needed to implement discontinuities. Chapters 4, 5, and 6 provide many examples of how new customers and applications (and methods of value capture) enabled discontinuities to be successfully introduced before the discontinuities provided the levels of performance and/or price that the previous technology provided. In other words, these new customers, applications, and method of value capture reduced the minimum thresholds of performance for these systems and their key components. However, although this was important from the standpoint of competition between firms, the impacts on these thresholds of new customers, applications, and methods of value capture (and the heterogeneity in customer needs they reflected) were fairly small when compared to the many orders of magnitude in system performance that came from improvements in component performance.

Third, one reason that discontinuities emerged in computers and in magnetic recording and playback equipment is that they did not benefit from geometrical scaling to the extent that their components did. ICs and magnetic recording density experienced exponential improvements in cost and performance because they benefited from dramatic reductions in scale (i.e., geometrical scaling). However, since computers and magnetic recording systems do not benefit from increases in scale (as do, for example, engines), it was natural that smaller versions emerged and replaced larger ones.

Fourth, the demand for many of these improvements in components was initially driven by other systems and/or industriesThis enabled the new systems to receive a "free ride" from existing industries as improvements in components "spilled over" and made the new industries possible. This provides additional evidence that the notion of cumulative production driving cost reductions is misleading and impractical, a point that others such as William Nordhaus have made using different forms of analysis.[49] Not only is it by definition impossible for learning curves to help us understand when a potential discontinuity (one not yet produced) might provide a superior value proposition; Part II shows how improvements in components (e.g., ICs) gradually made new discontinuities economically feasible where the demand for these components was coming from other industries.

CHALLENGES FOR FIRMS AND GOVERNMENTS

Chapters 7 and 8 of Part III address a different set of questions: those that concern the challenges for firms and governments with respect to new industries.

While Parts I and II focus on when a discontinuity might become economically feasible, and thus imply that firms easily introduce and users easily adopt new technologies, Chapters 7 and 8 summarize the complexities of new industry formation and thus the challenges for firms and governments. These complexities may cause the diffusion of new technologies to be delayed, or they may enable new entrants or even new countries to dominate an industry whose previous version was dominated by other countries.

Chapter 7 focuses on competition between firms. Incumbents often fail when technological discontinuities emerge and diffuse, particularly when these discontinuities destroy an incumbent's capabilities.[50] New technologies can destroy a firm's capabilities in many areas, including R&D, manufacturing, marketing, and sales, where this destruction may be associated with the emergence of new customers. For example, Christensen argues that incumbents often fail when a low-end innovation displaces the dominant technology (thus becoming a disruptive innovation) largely because it initially involves new customers and serving them requires new capabilities.[51] Helping firms analyze the timing of technological discontinuities, which is the subject of Part II, can help them identify and prepare for these discontinuities through, for example, identifying the appropriate customers and creating the relevant new capabilities to serve them.

Other research has found that the total number of firms declines quickly following the emergence of a technological discontinuity in some industries more than in other industries where the number of firms is a surrogate for the number of opportunities.[52] This decline occurs through mergers, acquisitions, and exits, in what many call a "shakeout." The occurrence of a shakeout depends on whether large firms have advantages over smaller firms through economies of scale in operations, sales, and/or R&D. For example, economies of scale in R&D (or other activities) favor firms with large sales because they can spend more on total R&D than can firms with fewer sales. Initially, greater spending on R&D leads to more products and more products lead to more sales; this positive feedback leads to larger firms dominating an industry, where smaller firms are acquired or exit the industry.[53]

Chapter 7 focuses on these issues in more detail and on how two factors, the number of submarkets and the emergence of vertical disintegration, affect the importance of economies of scale and thus the number of opportunities for new entrants. The existence of submarkets can reduce the extent of economies of scale in R&D when each submarket requires different types of R&D; thus the existence of submarkets can prevent a shakeout. This enables a larger number of firms, including entrepreneurial start-ups, to exist in an industry or sector with many submarkets than in one with few submarkets.

Vertical integration permits the late entry of firms, sometimes long after a shakeout has occurred. Furthermore, since vertical disintegration can lead to a new division of labor in an economy in which a set of new firms provides new types of products and services, it, too, can lead to the rise of new industries. While Chapter 7 primarily focuses on the emergence of high-technology industries such as computer software, peripherals, and services, and semiconductor foundries and design houses, vertical disintegration has also led to the formation of less high-tech, albeit large, industries such as janitorial services, credit collection, and training services.[54]

Chapter 8 focuses on how the challenges for firms and governments vary by type of industry, using a typology of industry formation. Whereas industries might emerge from either vertical disintegration or technological discontinuities, most of the examples in Chapter 8 are for those that emerged from the latter. The typology focuses on system complexity and whether a critical mass of users or complementary products is needed for growth to occur. Although the formation of most new industries depends on when a new technology becomes economically feasible and thus provides a superior "value proposition" to an increasing number of users, industries represented by complex systems and/or that require a critical mass of users/complementary products for growth to occur face additional challenges,[55] which may delay industry formation. Meeting these challenges might require agreements on standards, new methods of value capture and industry organization, government support for R&D, government purchases, new or modified regulations, new licenses, or even new ways of awarding licenses.

THINKING ABOUT THE FUTURE

Chapters 9 and 10 of Part IV use the conclusions from previous chapters to analyze the present and future of selected technologies. Chapter 9 looks at a broad number of electronics-related technologies such as displays, wireline and mobile phone telecommunication systems, the Internet and online services (including financial and educational services), and human-computer interfaces. Building from the notion of a technology paradigm, this chapter shows how improvements in specific components such as ICs have enabled new system-based discontinuities to become technically and economically feasible. More important, it shows how one can use an understanding of the technological trajectories in a system, or in key components of it, to analyze the timing of new discontinuities such as three-dimensional displays, cognitive radio in mobile phone systems, cloud/utility computing for the Internet, and gesture and neural-based human-computer interfaces.

Chapter 10 looks at three types of clean energy and how the four broad methods of achieving improvements in performance and costs can help us better analyze the potential for improvements in wind turbines, solar cells, and electric vehicles, so that we can provide better guidance on appropriate policies than can the typical emphasis on cumulative production. An emphasis on cumulative production says that the costs of clean energy fall as more wind turbines, solar cells, and electric vehicles are produced, that this "learning" primarily occurs within the final product's factory setting as automated equipment is introduced and organized into flow lines, that the extent of this learning depends on organizational factors, and that demand-based incentives are the best way for this learning to be achieved. Governments have responded to this emphasis on cumulative production by implementing demand-based subsidies, which firms have responded to by focusing on the production of existing technologies such as current wind turbine designs, crystalline silicon–based solar cells, and hybrid vehicles with existing lithium-ion batteries.

However, applying the three broad methods of achieving advances in performance and cost—notably, improvements in efficiency, geometrical scaling,[56] and key components—to clean energy leads to a different set of conclusions about policies. These policies involve the development of newer technologies and those that appear to have more potential for improvements than the ones currently emphasized.

For wind turbines, the key issue is geometrical scaling. Chapter 8 describes how costs per output have fallen as the physical length of turbine blades and towers have been increased, with these increases in scale requiring stronger and lighter materials. Thus, government policies should probably focus on the development of these materials through supply-based incentives such as R&D tax credits or direct funding of research. Furthermore, some evidence suggests that the limits to scaling have been reached with existing wind turbine designs, particularly those using existing materials, and so new designs are needed. Again, supply-based incentives such as R&D tax credits or direct funding of new designs will probably encourage manufacturers to develop new designs more than will demand-based subsidies.

For solar cells, improvements come from a combination of increases in efficiency and reductions in cost per area, where the latter are primarily driven by both reductions in the thicknesses of material and increases in the scale of production equipment (both are forms of geometrical scaling). The largest opportunities for these improvements are in new solar cell designs such as thin-film ones that are already cheaper on a cost-per-peak-watt basis than are crystalline silicon ones. Unfortunately, crystalline silicon ones are manufactured far more than are thin-film ones because turnkey factories are

more available for their manufacture and thus firms can more easily obtain demand-based subsidies for them. Therefore, as with wind turbines, governments should probably focus more on supply-based incentives such as R&D tax credits or direct funding of new forms of solar cell to realize the necessary improvements in efficiency and reductions in material thicknesses that appear possible with thin film.

For electric vehicles, the key component is an energy storage device (e.g., a battery), and thus appropriate policies should focus on this device and not on the vehicle itself. Chapter 10 describes how improvements in lithium-ion batteries, which currently receive the most emphasis from vehicle manufacturers, are proceeding at a very slow pace, and notes that large improvements are not expected to emerge in spite of the fact that such improvements are needed before unsubsidized electric vehicles can become economically feasible. Therefore, to encourage firms to look at new forms of batteries (or other forms of energy storage device such as capacitors, flywheels,[57] or compressed air), again, governments should probably focus on supply-based incentives such as R&D tax credits or direct funding of new forms of energy storage devices.

WHO IS THIS BOOK FOR?

This book is for anyone interested in new industries and in the process of their formation, including R&D managers, high-tech marketing and business development managers, policy makers and analysts, professors, and employees of think tanks, governments, high-tech firms, and universities. The information it presents will help firms better understand when they should fund R&D or introduce new products that can be defined as a new industry. It will also help policy makers and analysts think about whether technologies have a large potential for improvement and how governments can promote the formation of industries that are based on these technologies. Furthermore, it will help uncover those technologies that have a potential for large improvements and thus a potential to become new industries, which is much more important than devising the correct policies for a given technology.

This book is particularly relevant for technologies in which the rates of improvement in performance and cost are large and thus the frequency of discontinuities is high. For firms involved with these technologies, understanding when technological discontinuities might emerge is a key issue because these discontinuities often lead to changes in market share and sometimes lead to incumbent failure. They may even lead to changes in shares at the country level. For example, the emergence of technological discontinuities impacted

the rising (and falling) shares of U.S., Japanese, Korean, and Taiwanese firms in electronics industries in the second half of the 20th century.

Thus, the information this book offers can help firms, universities, and governments better understand when discontinuities might emerge. On the one hand, scientists such as Kaku, in *Physics of the Future* (2011), Stevenson, in *An Optimist's Tour of the Future* (2011), and Diamandis and Kotler, in *Abundance* (2012)[58] discuss the scientific and technical feasibility of different technologies. On the other hand, business professors discuss the strategic aspects of new technology in terms of, for example, a business model.[59] This book helps one understand when scientifically and technically feasible technologies might become economically feasible and thus when firms, universities, and governments should begin developing business models and appropriate policies for them.

This book is also for young people, who have more at stake in the future than anyone else, and it has been written to help them think about their future. It will encourage students to think about where opportunities may emerge and thus the technologies they should study and the industries where they should begin their careers. In terms of opportunities, while the conventional wisdom is to focus students on customer needs or on what is scientifically or technically feasible, it is also important to help them understand those technologies that are undergoing improvements and how these improvements are creating opportunities in higher-level systems. This is something that few engineering classes do, partly because they focus heavily on mathematics (and are criticized for this).[60]

For example, helping students (and firms and governments) understand how reductions in the feature sizes of ICs, including bioelectronic ICs and MEMS (microelectronic mechanical systems), can help them search for new opportunities. My students have used such information to analyze 3D holograms, 3D displays, MEMS for ink-jet printing, membranes, wireless charging, wave energy, pico-projectors, 3D printing, different types of solar cells and wind turbines, cognitive radio, and new forms of human-computer interfaces (e.g., voice, gesture, neural), as well as to analyze the opportunities that are emerging from these technologies;[61] some of these presentations are a source of data in Chapter 9. Among other things, the final chapter discusses how this book can be used in university courses to help students think about and analyze the future.

Finally, the ideas discussed here can helps students and other young people look for solutions to global problems that will not be easily found. Without an understanding of technology change, how can we expect students to propose and analyze reasonable solutions? To put it bluntly, discussions of policies,

business models, and social entrepreneurship are necessary but insufficient. New technologies and improvements in existing ones provide tools that our world can use to address global problems. Therefore, proposed solutions should consider the potential for and rate of technology improvements. For example, Chapter 10 analyzes three types of clean energy and concludes that, because the potential for improvements is mixed, more radical solutions are probably necessary. We need to ask students the right questions and give them the proper tools so that they can do this type of analysis and propose more radical solutions.