BBVA

FRONTIERS OF KNOWLEDGE

FRONTIERS OF KNOWLEDGE

FRONTIERS OF KNOWLEDGE

JANET ABBATE

SERGIO ALONSO

JESÚS AVILA

ABHIJIT V. BANERJEE

FRANCISCO CALVO SERRALLER

PAUL E. CERUZZI

CARLOS M. DUARTE

JOAN ESTEBAN

LUIS FERNÁNDEZ-GALIANO

JOHN B. HEYWOOD

GERALD HOLTON

ALEXANDER KIND

CAYETANO LÓPEZ

JOAN MASSAGUÉ

JOSÉ M. MATO

ROBERT McGINN

GINÉS MORATA

LUIS DE PABLO

NATHAN ROSENBERG

VICENTE SALAS FUMÁS

FRANCISCO SÁNCHEZ MARTÍNEZ

JOSÉ MANUEL SÁNCHEZ RON

ANGELIKA SCHNIEKE

SANDIP TIWARI

BBVA

This book is conceived as BBVA's contribution to an extraordinarily ambitious task: offering the lay reader a rigorous view of the state of the art, and new perspectives, in the most characteristic fields of knowledge of our time.

Prestigious researchers from all over the world, working on the «frontiers of knowledge,» summarize the most essential aspects of what we know today, and what we aspire to know in the near future, in the fields of physics, biomedicine, information and telecommunications technologies, ecology and climate change, economics, industry and development, analyzing the role of science and the arts in our society and culture.

science, innovation and society: shifting the possibility frontier

FRANCISCO GONZÁLEZ
BBVA CHAIRMAN

The book you hold before you, published by BBVA, stands in tandem with the BBVA Foundation Frontiers of Knowledge Awards, whose first edition is about to be decided as these lines are being written. These awards, with a monetary amount among the highest in the world, are to be granted annually to teams of researchers and creators in eight categories—precisely the eight fields covered in this publication.

BBVA earns no immediate return from this initiative to foster the generation and dissemination of knowledge. Our group is not a part of the pharmaceuticals, information technology or telecommunications sector, or of any other industry that might obtain direct commercial benefits from the results of scientific research. Nevertheless, its purpose connects with two key vectors of BBVA's culture, strategy and activity: to work for a better future for people; and to do so by means of innovation, anchored in turn on the best available knowledge. We are convinced that, in this way, BBVA is fulfilling one of the functions that companies in general, and leading multinationals in particular, are called on to perform in the global society of the twenty-first century.

The BBVA knowledge agenda: innovation and corporate responsibility

There are currently over 70,000 multinational corporations in the world and they represent 25% of its economic production. In the last two decades, the combined foreign investment of these companies has exceeded the total amount of official development aid. They can thus be considered the main instruments for the construction of a global economy and society, facilitating the worldwide spread of technology, values and more modern and efficient commercial and management practices. Moreover, large corporations have an enormous social presence and impact through their employees, customers and suppliers, so can act as powerful catalysts for innovation and the transition to a sustainable world.

Companies cannot be a part of the world's problems; they have to be a part of their solution, and a vital one at that. In the twenty-first century, any responsible company attuned to the legitimate demands of its stakeholders has the duty to work for the improvement of the societies where it does business. And it has two very important reasons for doing so: conviction and interest. Conviction, because its actions must be guided by ethical considerations and the core values of corporate culture. Interest, because, in an increasingly informed and demanding society, companies need greater legitimacy if they are to successfully carry forward a long-term project, and because a prosperous and stable society is both a result and pre-condition of corporate development.

If that is the case for all companies, it is even more so for banks, because the financial industry stands at the heart of the economy and society. Its function is to help companies and individuals realize their projects, by offering them basic payment, savings and investment services, as well as a growing range of other, increasingly specialized solutions. For this reason, we can see banking as a fundamental motor for development. And for this reason too, it must serve as a repository of trust for social agents in not one but two senses: it must work for the interests of its stakeholders (in the strict meaning of the word "trust"); and it must also exhibit the prudence and professional competence associated with the term "confidence". Ethics and competence are two essential attributes that each institution, and the financial system in its entirety, must zealously protect. In recent months, events have revealed the grave effects on the global economy and society—and on finance entities themselves—of a breakdown of confidence in the financial sector.

In keeping with its central position in the economy and society, banking is fully exposed to technological and social change. Our customers change, in their needs, expectations and demands, as do the channels and procedures through which they choose to operate. Responding to these shifting demand patterns requires a profound technological, organizational and cultural transformation that parallels the transformation being undergone by global society, in order to harness the immense potential held out by technological and scientific advances. The goal of this transformation should be to sustain economic growth, improve the wellbeing of society and restore the environmental balance that has been lost in the recent past.

The financial industry works with two main raw materials: money and information. Money, at the beginning of the twenty-first century, has been largely dematerialized. It has turned into book entries, that is, information that can be transmitted instantly and at practically no cost. Technological advances have given banks exceptional opportunities to improve their services, and to take more and better products and services to an incomparably wider public, with a maximum of convenience and at much lower prices.

Unquestionably, the financial industry has changed in recent decades, adapting its systems and processes to new technological capabilities, but the breadth and depth of these changes have been of limited scope compared to other industries and services. Still, the weight of global events and social and economic realities means a more profound transformation is both urgent and inevitable.

BBVA seeks to occupy a leadership position in this transformation of the financial industry, following a strategy based on three pillars: principles, innovation, and people.

This strategy, and the corporate culture from which it flows—and which it nourishes—is encapsulated in our vision statement, "*BBVA, working towards a better future for people*". People in their multiple dimensions as customers, employees and shareholders, as well as citizens in the societies where we conduct our business.

We believe that we are contributing to a better future when we apply firm ethical principles of honesty and transparency, when we place people at the center of our activity, and when we work for the betterment of the societies where we operate, sharing in their aspirations. And, finally, when we foster innovation as a key enabler of more flexible, individualized customer solutions, affordably priced, as a means to give more people access to financial services, as a source of additional value for our shareholders and as an outlet for the creative talent of our professional teams.

This combination of innovation and corporate responsibility provides the framework for our commitment to promote and disseminate science and learning.

Innovation is a mainstay of the BBVA strategy and culture, as reflected in our current innovation and transformation plan. This perspective sets us apart from more conventional competitors, while leveraging our capacity to generate recurrent value on a sustained basis. We are aware that science, research and the creative climate they both draw on and enrich are essential aspects of process and product innovation, and the search for new, more efficient solutions to the demands and challenges of contemporary societies.

Responsibility towards the societies in which we do business—currently over thirty countries and growing in various continents—is also an integral part of BBVA's strategy and culture.

We believe our primary responsibility is to do our work well, and that striving day by day to improve the quality, reliability and price of the services we offer is the best way we have of contributing to economic development and social stability.

But our commitment is more far-reaching: we want to help make that growth sustainable over time. That is why we have implemented pioneering policies in our sector with regard to the environment. And why we have developed a social strategy and policy with three basic lines of action, in which education and knowledge play a central role.

The first line is articulated through our "Financial Inclusion" program, to favor access to the basic financial services of payments, savings and loans.

Banking accessibility is a key means to prevent economic and social exclusion and to further the personal development of low-income sectors of society who tend to be let down by the mainstream finance industry. In some areas where BBVA operates, such as Latin America, this group represents a majority of the population.

Many of these initiatives are being carried out by the bank itself by means of innovative models whose intensive use of technology makes it possible to drastically reduce the cost of producing and distributing basic financial services. Moreover, our Group is strongly committed to the development of microfinances, and has set up the BBVA Microfinance Foundation, a non-profit organization funded with 200 million euros.

"Aid to Education", especially among the least-favored segments of the population, is another priority action line. BBVA's Social Action Plan for Latin America is funded with 1% of the profits of each Group bank in the region, with more than 80% of those resources earmarked for educational initiatives.

The third line of action, as stated at the start of these pages, is "Knowledge Promotion". This is carried out fundamentally by the BBVA Foundation through programs supporting scientific research and publications, with special emphasis on social sciences, biomedicine, the environment and basic sciences, and arts and humanities (particularly Spanish and Latin American literature and contemporary music).

The present book, *Frontiers of Knowledge,* and the awards of the same name, are a part of this endeavor. They seek to address the lack of visibility and explicit recognition —especially notable in Spanish and Latin American society—of the multiple achievements of the scientific community. The idea is to improve social awareness of the latest developments in science, technology and contemporary artistic creation, and their significance as a source of new opportunities and choices for individuals and groups. They will also cast a spotlight on those whose research and creative work have enlarged the space of knowledge and enriched the realm of culture.

It is a paradox that society's high esteem for scientists and researchers in the abstract—evidenced by multiple surveys since the decade of the 1950s—finds little expression in the public recognition, and far less acclaim, of individuals who have contributed decisively to the advancement of knowledge. This contrasts with the high-profile treatment of other professions far less central to society's wellbeing. Only a few international prizes—the best known being the Nobel Prize—bring a select group of researchers and creators to the attention of the wider public. The Frontiers of Knowledge Awards want to make their own contribution to raising the social visibility of modern scientific culture.

These new awards, funded by BBVA through its Foundation, differ in important respects from other major international honors. They seek to recognize and encourage research and artistic creation, prizing contributions of lasting impact for their originality, theoretical and conceptual implications and, where appropriate, their translation to innovative practices of a salient nature. The name of the scheme is intended to encapsulate both research work that successfully enlarges the scope of our current knowledge—continually pushing forward the frontiers of the known world—and the meeting and overlap of different disciplinary areas. Specifically, honors are given in eight categories coinciding with the eight chapters in this book: Basic Sciences (Physics, Chemistry, Mathematics); Biomedicine; Ecology and Conservation Biology; Climate Change; Information and Communication Technologies; Economics, Finance and Management; Development Cooperation; and Art (Music, Painting, Sculpture, Architecture).

Besides the number of categories and their respective contents, there are other elements that give these awards their unique profile. Firstly, they recognize the increasingly interdisciplinary nature of knowledge through the closing decades of the last century and up to the present day. Secondly, they acknowledge the fact that many seminal contributions to our current stock of knowledge are the result of collaborative working between large research teams. This is why, unlike others, these awards may be shared by any number of any size teams, provided the achievement being recognized is the result of collaborative or parallel working. Thirdly, honors in science and technology are joined by a category recognizing creative work of excellence in four fields decisive in shaping the culture and identity of this or any other era, namely contemporary music, painting, sculpture and architecture. And fourthly, recognition is extended to diverse endeavors (from research through to practical actions and initiatives) in the areas of climate change and development cooperation, two central and interlocking issues of the global society of the 21st century that are vital to the fight against the poverty and exclusion affecting large areas of our planet.

These awards respond to BBVA's vision of knowledge and innovation in our global society; a vision which we outline here by way of introduction to the central chapters of this book.

Knowledge society and global society

The term "knowledge society" and related terms like "information society" or "the knowledge economy" made their first appearance in the 1960s. They all refer to the one phenomenon or facets of the same: namely, the emergence of a society (an economy) in which scientific-technological advances—especially in information and telecommunications—and their rapid transfer to society become central planks of economic activity, as well as deeply transforming culture and lifestyles.

The first author to use the concept of "information society" may have been the economist Fritz Machlup 1962. As early as 1962, he noted that the number of employees engaged in handling information was greater that those carrying out physical or manual labor.

The expression "knowledge society" was first used by Peter Drucker in his 1968 book *The Age of Discontinuity.* In it, he postulates that the basic resource for the production of wealth in our time lies in knowledge and its productivity. Later, the same influential author emphasized the profound social transformation this would imply and the close links between "knowledge" and "globalization". In Drucker's view, the rise of information and communication technologies allows companies, products and consumers to transcend national borders, furthering the emergence of a global market (Drucker 1994).

In the decades since Machlup and Drucker's initial contributions, the trends they detected have gained more force than they could possibly have foreseen. It is no surprise, then, that a series of sociological and economic models have been proposed in recent decades to explain the transition from Industrial Society to what is known as Postindustrial Society, incorporating the main attributes of the Information Society. Two of the best-known authors of such models are A. Touraine and, especially, Daniel Bell.[1] The societies of the last third of the twentieth century have received many other similar labels, the most characteristic of which have been compiled by Beninger (1986). These labels emphasize the importance of the *technological* base—especially "*information* technologies"—in modeling the structure of advanced societies: *Computer Society, The Information Era, Compunications, Postindustrial Society, Electronic Revolution, The Wired Society, The Micromillennium* and *The Third Wave.* And the list could be further enlarged to include other expressions, such as Beninger's own *Control Revolution,* or, in the final decade of the twentieth century, *Network Nation, Virtual Community, The Network Society* and *Cybersociety 2.0.*

Technological advances, cultural change and innovation

The interaction of electronic technologies, new materials, computers and telecommunications, as well as developments underway in the fields of nanotechnology and biotechnology, have made it possible to replace the technological base that sustained various decades of uninterrupted growth from the end of the Second World War almost until the close of the twentieth century. One of the essential components of the current techno-scientific push—the association of computing and telecommunications—has the peculiarity of affecting processes and products in all economic sectors without exception. It also spans the entire economic cycle, from design to marketing, and everything in between, including engineering and production. Besides the direct economic impact of information technologies, over at least the last two decades, their fundamental effects are also measurable in a long series of areas such as scientific work itself, education, healthcare, leisure, associative practices (the emergence of interest groups, electronic associations and "virtual" communities) and the field of culture.

In the last half century, information technology has made formidable progress. "Moore's Law" (actually an empirical observation that storage and information-processing capacity will double every 18 months), has proved true ever since Gordon Moore formulated it in the 1960s. But that is not all. The universal spread of computers and the development of the Internet have been of fundamental importance. The Internet is a platform whose enormous power lies in its combination of both the information it stores and continuously updates, and its status as a *network.* In fact, "Metcalfe's Law" establishes that the value of a network can be expressed as the square of the number of persons connected to it.

It is clear that one of the factors explaining the rapid spread of information lies in scientific-technological advances that allow the sustained improvement of features at the same time that costs drop. But the computer's trajectory from the mid-1940s to the present day has also been affected by social dimensions, some as intangible as the *way of seeing* the computer: what possibilities it offers and how we conceive of the "cohabitation" of human processors and electronic ones, to quote the stimulating image of Nobel prizewinner Herbert Simon (1985). It is worth briefly considering those two aspects—technology and society—as a way of shedding light on the complexity of innovatory processes.

Today, we take it for granted that the computer is an "all-purpose" technology ("The Universal Machine"), which puts it at a very different level than the emblematic machines of the industrial revolution, which were designed to do one, or at

[1] In the 1976 prolog to the second edition of *The Coming of the Postindustrial Society,* Daniel Bell expressed his discomfort with the labels, "Information Society," "Knowledge Society" and "Service Society," which only refer to partial aspects of the model of emerging society. But the analytical focus and literal tenor of Bell's argument so clearly address the techno-intellectual dimension of new society, that he clearly merits inclusion in the group of authors of information society models.

most a very few previously-determined tasks. We see, and use, the computer as technology that can support and expand an ever greater variety of mental functions and specialized tasks. This versatility goes far beyond the function its name refers to ("computing" or "calculation"), leading to such varied functions as integral treatment of quantitative or categorical (qualitative) information and the creating of images, or even "virtual worlds", and much more, including interaction with telecommunications, a flexible and robust communications platform that accepts all formats from voice to text, images and video, and so on. Moreover, it spans the entire planet, thus shaping "Global Networks" (Harasim 1993).

Our perception of the breach with the recent past brought about by the universal spread of computers and the web it has woven is so strong that we must turn to historians of technology to realize that the dominant conception of early electronic computers, even among most of their creators, was that of a highly specialized technology destined to occupy a limited place in just a few large organizations: corporations, the military and scientific institutions. Paul Ceruzzi, who collaborates in the present book, points out that, as late as 1951, experts thought the United States' computer needs would be satisfied with four or five computers (Ceruzzi 1986). Such a narrow view of the computer's possible uses and destiny can be explained by technological and cultural factors.

Beginning with the technological questions, analysis of the history and processes of scientific-technological innovations has shown that the maximization of the potentialities of a specific technology requires the confluence of various technological developments (clusters of different advances) (Freeman and Soete 1997). For Rosenberg (1976), the length of time between an "invention" and its spread as an "innovation" is largely dependent on the time it takes to carry out *additional* inventive activities to refine the original design in order to meet the needs of the final users.

In the case of the computer, without the microelectronic revolution it would be impossible to explain the mass production of computers and their spread from large companies to private homes. Nor would we be able to explain what the late Michael Dertouzos, a computer scientist at MIT, called "hidden computers"—microprocessors built into a broad range of products and technologies (advanced machine tools, automobiles, household appliances, sound and image devices, clocks and watches, and many, many others) to improve their features, increase their reliability, considerably save materials and even allow remote diagnosis (Dertouzos 1984) and repair.

From a cultural standpoint, the early conception of the computer is similar to what happened with other radical technologies when they first appeared: the new technology was harnessed with the image of the tool or technology it "replaced". Thus, the car was seen as a "horseless carriage" rather than as the beginning of "the era of auto-mobility". And for a short period of time, even Alexander Graham Bell considered the telephone as a means of sending messages from a central point (like a radio), rather than as a conversation tool. In that sense, there is nothing strange about the initial conception of the electronic computer as a powerful tool intended to advantageously replace the only existing "computers" of the time: dedicated individuals with paper and pencil—or mechanical "tabulators"—who calculated tables used in different areas (navigation, insurance, etc.).

Indeed, the very possibility of a *personal* computer required a multitude of advances in programming languages, interfaces, operating systems and applications, all of which have practically eliminated the initial barriers to its use. But cultural factors—how the computer is viewed and how final users relate to it—have also been fundamental to its massive spread. Visionaries such as Vannevar Bush, Douglas Engelbart, the Stanford Research Institute's (SRI) "Augmented Human Intellect" project, the community of Artificial Intelligence researchers, the legendary Xerox research center in Palo Alto, the Computer Supported Cooperative Work (CSCW) research program, and the implementation of some of those communities' guidelines at the beginning of the Apple company, shaped a view of the computer as technology to "expand" (rather than replace) the capacity of human beings, broadening the possibilities of communication and cooperation in working groups and even among social groups and individuals.

The jump from a dozen or so large computers in the 1940s, each carrying out a few specialized tasks for a tiny segment of scientists and engineers, to the millions of all-purpose microcomputers accessible even to children in the 1990s represents a fundamental milestone in the history of technology and its social impact. But more recently, the closeness and merging of telecommunications and computing, characterized by the exponential growth of networks and the Internet, have marked a decisive before and after in the development of technology, as well as an unprecedented space for social experimentation.

This entire complex process shows how radical innovation can only be successful when accompanied by the interaction of numerous advances, from strictly technological ones to those of a social or cultural nature. And each of these advances develops at its own pace, making it difficult to foresee when or even how an innovation with significant effects will emerge.

Toward a true knowledge society?

Among the social effects of computer and telecommunications innovations, the most important may be that citizens now live in an information society. We have access to a universe of information in perpetual expansion. We have ever more powerful and widely accessible technology for creating and accessing that information, as well as transmitting, treating, analyzing, relating, and, eventually, converting it into knowledge for use in problem solving. Thus, in a little over three decades, we have reversed a centuries-old pattern characterized, first, by an absolute paucity of information, and second, but no less important, by the concentration of information in the hands of a tiny proportion of society.

Of course our society is by far the one that has generated and accumulated the largest amount of knowledge in the history of humanity. In 1990, David Linowes affirmed that the amount of knowledge existing at the time of Christ did not double until the mid-eighteenth century. It doubled again in the following 150 years—around the beginning of the twentieth century—and again in just fifty years, around 1950. Nowadays, the volume of knowledge doubles every four or five years.

Nevertheless, a true knowledge society is still no more than an aspiration, a goal towards which we are advancing, though it still escapes us. There are two fundamental reasons for this: first, a large portion of humanity continues to be excluded from this development. As Janet Abbate points out in her article in this book, over half the population of developed countries was using the Internet by 2005, compared to 1% of the population in the 50 least-developed countries. Today there is an immense digital divide that slows the collective possibilities offered by computers and the web (Norris 2001). This urgently requires the attention of public and private agents if we are to bridge the gap between advanced societies and those that have not managed to enter the path of sustainable growth.

The second reason why a knowledge society is still more of a goal than an empirically observable reality is that the immense majority of available information is an enormous and undifferentiated mass of quantitative data and categorical or qualitative information. The structure of most of this information remains hidden (we do not know what internal relations exist between the multiple pieces that constitute such a mass), and we do not have articulate explanations to make it globally coherent. In short, we have not mastered an essential technology: how to convert that data into knowledge and how to convert a significant part of knowledge into innovation, i.e. into new applications that are useful in people's lives and in solving the planet's main problems. A clear example of this situation is the recent explosion of genetic information (the decoding of the human genome) and the difficulty of interpreting it and applying it to new and more efficient therapies. But intense work is being done today to develop statistical algorithms and methodologies capable of helping us discover the meaning of gigantic volumes of information of diverse nature. The success of this effort will largely determine our capacity to transform information into knowledge, which we can use to generate innovations that satisfy needs and demands in a multitude of areas (Hastie, Tibshirani and Friedman 2003).

The technological revolution and the rapid growth of knowledge have undoubtedly led to a highly expansive phase of global economic growth. Nevertheless, the benefits of that growth are very unequally distributed. The planet's current productive capacity allows it to support a population three times as large as in the mid-twentieth century. The quality of life has also improved in much of the world, and extreme poverty is shrinking, not only in relation to the total population, but also in absolute terms (United Nations 2007).

And yet inequalities have increased in almost every country in the world, as well as between the most and least advanced areas (World Bank 2008). Not surprisingly, there has been a proliferation of reflections on the difficulties and problems of globalization, as well as a greater focus on the problems than on the opportunities of globalization and scientific-technological change.[2]

Moreover, accelerated population growth and productive activity bring very serious problems of environmental sustainability linked to the overexploitation of natural resources on land and sea, scarcity of fresh water, the accelerated loss of biodiversity (species and habitats) and climate change. Each of them will have its own impact on populations and their economies in the coming decades, and these fundamental questions are being addressed in the present book with contributions by outstanding specialists.

Still, there are reasons for optimism. The basic sciences continue to delve ever deeper into physical, chemical, and biological processes and structures, with consequences of all kinds, from strictly cognitive ones (improved knowledge) to technological ones (new instruments to cover necessities). We are at the beginning of the evolution of extremely powerful and

[2] In 1998, the sociologist from the University of Chicago, Saskia Sassen published *Globalization and Its Discontents* (New York: The New Press, 1998) and four years later, the Nobel Prize-winning economist, Joseph E. Stiglitz published his own work with the same title (*Globalization and Its Discontents.* New York-London: W. W.Norton & Company, 2002).

highly versatile technologies, including the Internet, whose third generation will offer much greater possibilities for cooperation and the active inclusion of everyone in electronic space. We are barely scratching the surface of a gigantic vein of wealth and wellbeing for humanity. And this is certainly the most "democratic" technological revolution ever. Not only for the obvious reason that democracy is the reigning political regime in most of the world right now—though this undoubtedly reinforces its positive effects—but also because this is the revolution that has most rapidly spread around the world, despite the barriers and limitations mentioned above. This is the one that has proved most accessible to people in all parts of the world.

How can the generation of knowledge be accelerated?

This context, in which serious and ever-more-pressing problems coexist with enormous potential for scientific and technological growth, poses questions as to the most appropriate mechanisms and procedures for empowering the generation and spread of knowledge. And those questions are key to the future of humanity.

Traditionally, there were two main catalysts to the generation of knowledge: economic advantage, which drove private agents; and war, which drove government initiatives and programs.

War requirements attained unprecedented efficiency as catalysts for scientific and technological advance in the context of the Second World War. Subatomic science and technologies, microelectronics and the computer were fostered by the war effort, and there were also fundamental advances in other areas, especially medicine, pharmacology, psychology, and operative research.[3]

The Second World War's impetus to scientific and technological research was reinforced in the last decades of the twentieth century by a growing participation of the private sector. This was a response to growing opportunities for commercial exploitation of advances in research.

In the last fifty years, there have been profound changes in how knowledge is generated and used. In the light of the unprecedented growth of scientific knowledge, governments, companies and scientific institutions have been debating the most efficient ways to apply scientific advances to make companies and countries more competitive, thus improving collective possibilities.

For decades, the dominant approach was for public powers and a few large corporations to unreservedly support basic research on the supposition that sooner or later the knowledge it generated would lead to practical but highly unforeseeable applications of the most radical sort: applications that competing countries and companies would have great difficulty imitating. This model was more or less related to the doctrine of Vannevar Bush and the experiences of the Second World War.[4]

Since at least the 1980s, though, conceptual and practical dissatisfaction with this inherited approach became patent. For example, MIT's Commission on Industrial Productivity, whose members included the Nobel Prize-winning economist, Robert Solow, published the influential interdisciplinary report, *Made in America* (Dertouzos, Lester, and Solow 1989), which sought to explain the paradox that the United States had the most advanced basic science and the best-trained scientists and technologists, yet its capacity to turn that cognitive gap into innovation had dropped in comparison to the decades immediately following the end of the Second World War. Yet Japan, where the focus was more on applied research, seemed capable of capturing markets for unmistakably American products (such as consumer electronics), brought new products onto the market in far less time in sectors traditionally associated with US industry (automobiles), and reached far higher levels of quality.

Such difficulties were hardly limited to the United States, as the authors of the mentioned report seemed to believe. Almost a decade later, the European Commission (re)discovered the same problem and did not hesitate to label it "the European paradox" in *The Green Book of Innovation* (1995) and the related *Made in Europe* project.

This type of analysis has led—not without criticism from some sectors of the scientific community—to a change of perspective in the drawing up of scientific policy. Support for research directed at the advancement of knowledge (known as "pure" or "basic" research) with no direct practical applications has lessened. This "type" of science has had to cede, or at least share, financial and human resources with so-called "strategic research", subject to external planning (by public agencies or institutions), that seeks to satisfy socioeconomic objectives—especially the competitiveness of national economies, defense, and health policies.

This focus has also taken root in large companies, leading to a strict alignment of research programs with economic objectives. In many cases, companies have reduced or even eliminated their own industrial laboratories, turning to the market instead. In other cases, R&D departments are replaced by contracts with public centers or private institutes dedicated exclusively to research. Uncertainty about the possible results and the difficulty of exclusive control of the

[3] The interaction between military needs and technological development is analyzed in the William H. McNeil's classic *The Pursuit of Power. Technology, Armed Force, and Society since A.D. 1000.* Chicago: Chicago University Press, 1982. The specific and emblematic case of the USA is addressed in the work edited by Merrit Roe Smith, *Military Enterprise and Technological Change. Perspectives on the American Experience.* London: MIT Press,1987.

[4] For a critical analysis of the strengths and limitation of the model of support for the sciences linked to this engineer from MIT, see: Claude E. Barfield,ed., *Science for the 21st Century. The Bush Report Revisited.* Washington: The American Enterprise Institute Press, 1997.

fruits of such research—especially in the case of pure research—have been important factors in this reconsideration of the scale and role of Research and Development and its financial support.

Since the 1980s, public administrations in advanced societies have considered the postwar model of pure science development exhausted.

But the underlying suppositions of models for the strategic direction of science are far from solid. The history of technology and innovation reveals the winding path that leads from purely theoretical progress to the improvement and introduction of new processes and products, and vice versa.

It is certainly difficult to predict and manage the transformation of theoretical advances into new applications. On the other hand, the links between theory and practical application have multiplied, and their roots have grown deeper, making it necessary to discard simplified notions of what is useful and what, theoretically, "only" contributes to greater knowledge of reality. In countries and regions culturally and institutionally committed to excellence and innovation, public and private agents share the view that economy and society are increasingly dependent on an infrastructure of intangibles—theories, information, and scientific knowledge—in which scientific activity and corporate strategies broadly overlap and continuously redefine themselves.

Europe, which is generally slow to incorporate concepts and experience from the other side of the Atlantic, needs to pay more attention to what is really going on in the United States. Literature and empirical evidence show that scientific research financed with public funds has played a leading role in the United States' industrial innovation.

Patterns of innovation in the United States indicate the need to increase public and private support for Research and Development on this side of the Atlantic, promoting science and technology of excellence, and, most of all, introducing a "market" culture of open competition and sustained effort to excel, with universities and research centers ranked according to their capacity to contribute to knowledge and innovation. There must be mobility and interaction among researchers and the private sector, and new interfaces for efficient communication between institutions dedicated to the creation and transmission of knowledge and the corporate world. This is a program whose development requires vigorous support and coordination by public administrations on every scale, from European to national to regional.

It is time to renegotiate the previous "implicit contract" between universities, industry and the administration, redefining what each can expect from, and give to, the others. As two outstanding experts on innovation—professors Rosenberg and Nelson—have pointed out, we must modify the *status quo.* But these changes must be based on a careful consideration of the functional specialization of each institution, seeking a better division of labor among all participants in the innovation system.

What seems clear, at any rate, is the need to establish a tight network of relations between industry and universities. This can take various forms, from fluid access by corporate R&D personnel to their university colleagues (and vice versa) to the development of specialized institutions halfway between corporations and public research centers, along with public financing of research areas whose goal is to improve competitiveness, supervised by composite advisory committees with representation of academia, corporations and the public administration. No matter what formulas are applied, what really matters is to create dynamic networks that transmit information and signals, and to generate trust and the exchange of tacit knowledge (which is difficult to codify) among the different participants, breaking down barriers still visible in much of Europe, which, succinctly put, separate academia and corporate enterprise.

Interactions between science and technology

The results of scientific research and technological innovation are increasingly present in all aspects of human life and activity. As Peter Drucker points out, "much more than a social transformation, [they are generating] a change in the human condition" (Drucker 1994). This creates growing interpenetration and cross-pollination between scientific research, innovation, and human productive activities and lifestyles. It also leads to a drastic reduction of the lag time between scientific discovery and the commercial exploitation of its results (Mowery 1989).

Increasingly, science and technology are advancing driven by their overlap and cross-fertilization, and also through the interaction of classic disciplines and the emergence of new ones, illustrating the obsolescence of the classical argument as to whether technology depends on previous scientific knowledge, whether the latter benefits from the former, or whether the two are completely independent of each other. During the twentieth century, especially the second half, relations between science and technology, and between both of them and society, changed in fundamental ways. Corporate and industrial sectors, as well as social demands in areas such as health, energy, agriculture and food, transportation, the environment, inequality and poverty are sources and signals for science. They call for the analytic power that can only be provided by scientific research, and the efficacious and efficient solutions offered by technology—an area that Nobel laureate Herbert Simon labeled "the sciences of the artificial" (1996).

The present complex framework also helps to explain the growing importance of multidisciplinary cooperation in contemporary scientific research, as well as the fact that most scientific research is carried out by large teams made up of researchers from different institutions, based in different parts of the world. Telecommunications and Internet innovations allow active and simultaneous participation in a specific project by researchers all over the world, including—and this is a motive for optimism—those from the least rich and advanced regions of the world.

Science's institutional architecture and cultural setting

Nowadays, science is a markedly social and highly institutionalized activity. It continues to require individual creativity and risk, but it is developed cooperatively in specialized organizational frameworks and a social setting from which it obtains not only the adequate human and material resources, but also signals (appreciation of science, research priorities) and conceptual and cultural influences. These can come from nearby fields, or from those of the Humanities, and from overall culture (the "worldviews" active in a society).

The alignment and positive interaction of all these elements has become critically important in recent decades. As Nathan Rosenberg explains in his excellent article in this book, a crucial challenge is the adaptation of science and research's institutional setting to our global society.

In that sense, we now face new organizational challenges for the development of research that is increasingly interdisciplinary, multipolar—even delocalized—and strongly cooperative, as well as increasingly interactive with its social medium.

How can we develop interdisciplinary research in universities divided into departments defined by specific disciplines, with researchers and scientists who—at least in the academic world—place great importance on the fact that they work in a recognized field? How can we combine the frameworks and disciplinary structures of knowledge, which are well defined and based on theoretical traditions and reasoning, with interdisciplinary institutions and centers that are closer to dealing with practical challenges? How can we reconcile the interests of governments and national public agencies—they are, after all, an integral part of the scientific world—with the configuration of multinational, highly flexible and changing research teams? How can we protect the incentive for companies to assign resources to research when projects have multiple participants and vital information can be divulged instantly *urbi et orbe*? And lastly, how can we ensure that this entire institutional structure focuses on the solving of problems of general interest, effectively contributing to the wellbeing of people, so that scientific and technological advances do not lead to increasing inequality and greater problems of global sustainability?

The mere enumeration of these challenges indicates that many of the answers must come from the field of the social and behavioral sciences, especially the "soft" technologies of organization and incentives, as well as the study of culture and attitudes.

Moreover, the redesign of science and technology's institutional architecture, of public policy for the promotion and strategic management of R&D and innovation by corporations, requires intangibles, values and perceptions—that is, science's cultural setting—to be sensitive to this, operating to foster and orient it.

Reconciling science, technology, and society

A positive social view of science is crucial in at least three fundamental ways. First, so that the citizenry, with its opinions, votes and even its buying power, can push public policy-makers and private industry decision-makers to support and invest in education and research. Rewarding innovation (the "high road") and discouraging competition based merely on low costs rather than on added value. Second, it attracts human capital to the sciences, so that talented youths feel motivated to undertake a demanding but thrilling career in research that is rewarding both economically and symbolically. Finally, the intellectual and cultural "appropriation" of science by society is crucial for both scientific creativity and for the efficient use and integration of new developments into the social tissue. In sum, the attitude that, in the case of the United States, was labeled by technology historian Thomas Hughes as "technological enthusiasm" is decisive for the advance of knowledge and of the society that fosters and welcomes it (Hughes 2004).

We might be tempted to think that, after various decades in which science and technology have made stunning contributions to humanity's economic progress and wellbeing, the general consideration of science as an unmistakably positive factor must be firmly established. But, as Gerald Holton points out in his excellent essay in the present volume, social attitudes towards science have historically been subject to strong oscillations. And the social status most desirable for science is not guaranteed. Following the optimistic sense of progress that reemerged after the Second World War—exemplified by Vannevar Bush's famous report, *Science, the Endless Frontier,* commissioned by President Franklin Roosevelt and published in 1945—voices and critical movements spoke out in the final part of the twentieth century against science's role in our society, recalling motives of Romantic resistance to science and modernization (Marx 1988). Such criticism attributes negative effects to scientific and technological progress, relating them to the development of weapons of mass destruction, the deterioration of the environment, inequalities within each society and between different parts of the world, and even the configuration of a uniform, dehumanized and excessively materialistic global culture lacking moral values.

Of course, concern for these questions, especially the serious worldwide deterioration of the environment, is not only legitimate, it is shared by many, many citizens. Leo Marx, the historian of American culture at MIT, has indicated that the belief in progress that characterizes modern Euro-American culture was eroded during the last decades of the twentieth century, mostly by pessimism about the human role in nature and the perception that the system of industrial production based on science and technology is having strong, unwanted effects on the global ecosystem (Marx 1998).

More-or-less systematic criticism of science seems to have lessened at the end of the first decade of the twenty-first century. Nevertheless, underlying concern about some of science's unwanted, though indirect, effects, as well as the complexity of global society, make it fundamentally necessary to promote and consolidate a favorable attitude towards science—a view based on the assumption that scientific and technological advances are, in fact, key elements in helping humanity deal with its largest challenges, and a recognition that scientific and humanistic aspects of our culture are not only fully compatible, but that together they can and must contribute to a sustainable improvement of the conditions of human existence. Three quarters of a century ago, the US philosopher and educator, John Dewey, made a recommendation that we would do well in recalling and applying today: use science to "cure the wounds caused by applied science" and, in particular, to foster the development of scientific culture and the transmission to general society of mental habits and attitudes that are characteristic of researchers: curiosity, objectivity, innovation, rational debate and a willingness to change one's mind on the basis of discussion and empirical evidence (Dewey 1934). This, in short, is the tradition of enlightened rationalism tirelessly defended by Karl R. Popper, undoubtedly one of the greatest philosophers and thinkers of the second half of the twentieth century.

To contribute, even in a modest manner, to this great task is the fundamental purpose of the present book, in which outstanding figures in science and the arts of our time—researchers on the frontiers of knowledge—review the state of the art and perspectives for our century's most characteristic scientific and artistic disciplines. These are the disciplines most responsible for advances visible to the overall citizenry, and these are the ones addressing the challenges most relevant to our future and that of our children: health, information and communications technologies, natural resources, the environment and climate change, the generation and fairer distribution of wealth, and of course, the arts, which are the expression of our culture and the sensors of social concerns in our time.

At BBVA, we are proud to contribute to the fostering of knowledge and creativity with the publication of this book and, in a more permanent way, through the BBVA Foundation Frontiers of Knowledge Awards. Our very sincere thanks to each and every one of the outstanding researchers and creators who responded to our request for first-hand reporting—with the intimate and authoritative experience conferred by their illustrious careers—on a selection of questions fundamental to their respective fields. It is our wish and desire that the readers of this book enjoy it as much as we have enjoyed publishing it, and that they join us in saluting the thousands of researchers who strive daily to advance our knowledge of the natural and social worlds, thus expanding our freedom to make decisions and our individual and collective possibilities.

what place for science in our culture at the “end of the modern era?”

GERARD HOLTON

Prefatory note

It is the very essence of democracy that any institution’s claim to a measure of authority invites, almost automatically, scrutiny by reasoned counter-argument. That is also true, and has been for centuries, for the authority that has been asserted on behalf of science and its place in Western culture.

But from time to time, those reasoned counter-arguments have been submerged under a flood of passionate, unreasoned, even sensationalist attacks on the place of scientific knowledge. (One thinks here, for example, of the “Bankruptcy of Science” movement in the nineteenth century.) The same process seemed to me to be beginning to happen some years ago, when the following pages were written in order to illustrate and to understand this social phenomenon, as well as to alert some among the usually placid scientific community to notice the challenge and to act upon it.

The hope was then also that—in part owing to the extraordinary advances continually being made in modern science and in its useful applications in daily life—those extreme voices would be muted. However, this has not happened. In fact, a combination of quite different forces have been at work (at least in the USA and some European countries) to swing the pendulum of antagonism against the authority of science—in academe, in popular culture, among very visible politicians, even among some theologians. There has been a continued increase in books with such titles as The End of Science*; in scholars’ publications with the central arguments that the scientific experimental method by its very essence “arose out of human torture transferred onto nature”; in highly-funded attacks on the biology of evolution; among some postmodern philosophers and sociologists, arguing that we are now “at the end of modernity,” and that the concept “nature,” having no validity, makes doing science an attempt at careerism; and in the suppression, at the highest level of government, of widely agreed-upon scientific findings regarding dangers to the environment and public health.*

In sum, the observations and findings presented below regarding the place of science in our culture have grown even more relevant in our time.

Behind every act in the life of a scientist—whether it be the choice of a research program, or interaction with students, the public and the media, or the never-ending search for funding, or advice asked by government officials—there is a hidden factor that in large part determines the outcome. That factor is how society at large regards the place of science in our culture. Most practitioners of science would claim they have little interest or expertise to concern themselves with such a seemingly complex and amorphous problem—at least not until the time comes, as it does periodically, when they begin to notice that their largely unconscious assumptions about the relations of science and the wider polity are being severely challenged.

Such a time has arrived once more. Here and there, intellectuals are waking up to the fact that increasingly such concepts as the "end of the modern era," the "end of progress," and the "end of objectivity," originating from parts of academe, from eloquent popularizers, and even from members of Congress, are making an unquestioned place for themselves in the public mind, with surprisingly little audible opposition from leaders of the scientific establishment. But far from being a passing phase, the movement—different from the anti-science phenomenon that I have tried to analyze elsewhere[1]—signals the resurgence of an old, recurring rebellion against some of the Enlightenment-based presuppositions of Western civilization, particularly against the claim of science that it can lead to a kind of knowledge that is progressively improvable, in principle universally accessible (i.e., intersubjective), and potentially valuable and civilizing. The impact of the resurgence of this reaction on the life of the scientist, on the public understanding of science generally, and on the legislation of science policy, is measurably growing and will become palpable even for the least attentive.

The aim of this essay is to help understand the movement, its main sources, and its driving ambitions. To this end it is well to begin with a survey of some of the chief theorists on the question of what role, if any, science may play in our culture, and its effects on key legislators in the US who are now redesigning the direction and conduct of science. In effect one must look back beyond the so-called implicit "contract" forged in the aftermath of World War II between science and society.

That contract, still the dominant myth among the majority of scientists even while it hardly corresponds to reality today, was the result of a more innocent phase, when for a few decades the pursuit of scientific knowledge was widely thought—above all by the scientists themselves—to embody the classical values of Western civilization, starting with the three primary virtues of truth, goodness, and beauty: when science tended to be praised as a central truth-seeking and enlightening process in modem culture—one might call it the Newtonian search for Omniscience; when science was thought to embody a positive ethos in an imperfect world, both through its largely self-correcting practice of honor in science, and through its tendency to lead to applications that might improve the human condition and ward off the enemies of our form of society—a Baconian search for a benign sort of Omnipotence; when the discovery of beauty in the structure, coherence, simplicity and rationality of the world was thought of as a Keplerian enchantment, the highest reward for the exhausting labor.

Before the euphoria ended

The last time the optimistic description just given could have been said to be generally taken for granted, at least in the US, was the period following the ending of World War II. It was embodied also in the famous Vannevar Bush report, *Science, the Endless Frontier*, of 1945, which became a main driving force of science policy in that country. Because it is such a convenient example of modem post-Enlightenment optimism about the role of science in culture, one that so many scientists tacitly assume to be still operative, it will be illuminating to look at the main thrust of that document.

In November 1944, President Franklin D. Roosevelt requested from Vannevar Bush, the head of the wartime Office of Scientific Research and Development, a report that would outline how, in the postwar world, research in the natural sciences—he called it "the new frontiers of the mind"—could be strengthened and put to service for the nation and humanity. Roosevelt was particularly interested in three results: waging a new "war of science against disease," "discovering and developing scientific talent in American youth," and designing a new system of vigorous federal support for scientific research in the public and private sectors. Beyond those, he argued that science's applications, so useful during the bitter war to preserve the world from fascist dictatorship (with the successes of the Allies' radar and antisubmarine devices the most striking examples at that time), now could be harnessed to "create a fuller and more fruitful employment, and a fuller and more fruitful life."

Vannevar Bush's detailed response came less than eight months later, the result of a crash program by an impressive brain trust of about forty experts from industry, academe, and government. Roosevelt had died, but with the war's successful end in sight, the

[1] Holton, Gerald. *Science and Anti-Science.* Cambridge, MA: Harvard University Press, 1993, chapter 6.

American administration proved generally hospitable to the report's ideas. While some of the details were too optimistic and others were modified in practice (often to Bush's dismay), his vision, it is generally agreed, set the stage for the development of new institutions for the support of science during the following decades, and paralleled the generally favorable popular attitudes that were prerequisites for the actions. The base was laid for global leadership in many branches of basic science. Not until the Vietnam war escalated was there substantial popular disenchantment both with governmental authority, with the widely visible use of sophisticated technology in a hopeless and unpopular war, and by implication with science that presumably could help give birth to such abuse. It signaled the end of what might be called a rather euphoric phase in the relation of science and society in this century.

The Bush report, together with the rival proposals by Senator Harley Kilgore, were historic exemplars of the science-based progressivism reigning in its time, which saw science and democracy as natural allies in the service of the ideal of empowerment and instruction of the polity as a whole. In this sense, they were part of the American dream as far back as Benjamin Franklin and his fellow statesmen-science amateurs. Vannevar Bush himself hinted as much in the brief preface to his report, taking courage from the fact that, as he put it, "the pioneer spirit is still vigorous within the nation." And to make the connection with the tradition of Condorcet even more explicit, he added a sentence that, while presenting the reigning opinion of a citizen of the mid-1940s, is likely to be rejected today by many who think of themselves as the children of the 1960s and 1970s. He wrote: "Scientific progress is one essential key to our security as a nation, to our better health, to more jobs, to a higher standard of living, and to our cultural progress." One could hear an echo of Thomas Jefferson's formula: "The important truths [are] that knowledge is power, knowledge is safety, knowledge is happiness."

Bush and his contemporaries could hardly have imagined that by the early 1990s those hopes had begun to be rejected, even at the highest levels—that, for example, a key person in the US Congress for science policy could imply (as we shall see in more detail later) that science and technology alone can be held to account for the whole sorry list of failures over decades of misdirected political leadership. He said: "Global leadership in science and technology has not translated into leadership in infant health, life expectancy, rates of literacy, equality of opportunity, productivity of workers, or efficiency of resource consumption. Neither has it overcome failing education systems, decaying cities, environmental degradation, unaffordable health care, and the largest national debt in history."[2] And another highly placed observer, formerly the Director of the National Science Foundation, exulted: "The days of Vannevar Bush are over and gone [...] the whole world is changing."

The changing balance of sentiments

After this reminder of a mid-century worldview predominant before the generation now in leadership positions came on the scene, we turn from the level of momentary vagaries to come closer to understanding the causal mechanisms responsible for the changes in the place assigned to science at significant stages in the intellectual history of the past hundred years. For if we know the general causes in the variation of the underlying ideology, we shall better understand the changes in policy toward science at a given moment.

Here we must confront at once the question of whether these changes are gradual, and part of an evolutionary development, or are so sudden that, as if in a political revolution, one passes discontinuously from the end of one age to the beginning of another. If the latter is the case, we would now be passing through a rupture of history, with "modern" behind us and "postmodern" right, left, and all before us. While I doubt this is the case—and it certainly is not visible in the *content* of science as against some of the current writings about science today—a fashion in history proper has for some time been trying to discern the arrival of a new age. Periodization, the arranging of the flow of events into clearly separate eras, is a common tool, although applied more wisely from the safe distance of retrospection. That is how we got such schoolbook chapters as "The Age of Reason" or "The Progressive Era in America" around the turn of the nineteenth century.

A chastening example of that whole genre was provided by the American historian Henry Adams. At the beginning of the twentieth century, he had been impressed by the publications of the physicist and chemist J. Willard Gibbs of Yale on the phase rule for understanding heterogeneous equilibria. Adams was also fascinated by the strange idea of some physicists of that day that the phase rule can serve, by analogy, as a means for putting into hierarchical order the following sequence: solid, fluid, gas, electricity, ether, and space—as if they formed a sequence of phases. Stimulated by such ideas, Adams believed that thought, too, passed in time through different phases, each representing a different period. In his essay of 1909, "The Rule of Phase Applied to History," Adams came to a remarkable conclusion about the imminent passing

[2] BROWN, George E. Jr., quoted in *Science*, vol. 260, May 7,1993, 735.

of modernity: "The future of Thought," he wrote, "and therefore of History, lies in the hands of the physicist, and [...] the future historian must seek his education in the world of mathematical physics [...] [If necessary] the physics departments will have to assume the task alone." Henry Adams' conclusion might fairly have been called in its own day a declaration of what the postmodern age would look like.

Today's formulation is likely to be exactly the opposite one. I cite this example—and many others come to mind—to signal my discomfort with trying to divide history into distinct periods. A less rigid and more workable notion is to recognize that at any given time and place, even during a period when a civilization appears to be in a more or less settled state of dynamic equilibrium, there exist simultaneously several competing and conflicting ideologies within the momentary heterogeneous mixture of outlooks. As Leszek Kolakowski noted, "It is certain that modernity is as little modern as are the attacks on modernity. [...] The clash between the ancient and the modern is probably everlasting and we will never get rid of it, as it expresses the natural tension between structure and evolution, and this tension seems to be biologically rooted; it is, we may believe, an essential characteristic of life."[3]

It is sometimes possible in retrospect to identify one of the competing worldviews as the most dominant one for a longer or shorter period. But what is also likely to occur when followed in real time are two effects. The first is that each of the different competing groups works fervently to raise its own ideology to a position where it would be accepted as the "taste of the time" or the "climate of opinion" which characterizes that particular age and region. The newest and most ambitious one *will also be trying as part of its agenda to delegitimate the claims of its main rivals*. Especially when the previously relatively stable equilibrium begins to crumble, the pandemonium of contrasting voices gets louder. Some partial victors rise to be major claimants above the rest, and one of them may even be generally recognized for a while as the embodiment of the new worldview or "sentiment" of the society. Secondly, in this constant seesaw of changing historic forces, mankind's inherent liability to engage in over-ambition or one-sidedness may infect some of these claimants (not excluding, on occasion, scientists). This is the tendency, as Hegel had warned, toward "the self-infinitization of man," or simply to *yield to excess*—which, in turn can generate, in reaction, the same sort of excess among the opposing claimants. Recognizing these two facts is, in my view, central for understanding the course of culture in our time.

In this general struggle, from that of Apollo vs. Dionysus in Greece to this day, the specific, more limited question of the place assigned to the scientific conception of the world has always played a part. Sometimes this place has been at the cherished core of the rising or victorious overall worldview, as noted above; sometimes it has found itself embedded in the sinking or defeated one, and then was even accused of nourishing a great variety of sins against the better interests of humanity.

Historians of ideas have mapped the changing forms of the general contrary trends. Wise political leaders, too, have at times watched with apprehension as the net balance of prevailing sentiments has taken a turn, for as Jefferson said, "It is the manner and spirit of a people which preserve a republic in vigor. A degeneracy in these is a canker which soon eats into the heart of its laws and constitution." Weighty scholarship has chronicled how one of the world conceptions, and the scientific position within it, gained predominance over the others for some decades in significant segments of Western culture—an example is Robert K. Merton's early study on science and seventeenth-century Puritanism. There is also much documentation that such sentiments subsequently gave ground, as the overall balance of benignity or distress moved the other way for some more decades. As to the practicing scientists themselves, most of them have paid little attention to this constant seesaw of sentiments, except to weigh in now and then as promoters of the positive swings, or occasionally to become victims of the negative ones.

Today, this oscillating spectacle, so engrossing to the scholar, has ceased to be merely the site for the research of historians. The general balance among the contending elements, and with it the attitude of traditional patrons, is changing before our eyes. Studying this current drama is as fascinating and fruitful for the historian of ideas, whose perspective I shall be taking here, as the appearance of a supernova may be for an astronomer. But in both cases, the present state is the product of an historic process, the latest member of a motley progression.

Toward a "Monistic Century"

Let us therefore look at some of those claimants for representing the climate over the past hundred years up to the present—a sequence of selected samples meant to be analogous to stages in the growth of a culture of cells seen under the microscope. Our first sample concerns an event as a new century was signaling its beginning: the World's Columbia Exposition at Chicago in 1893. The fair was intended as a triumphant celebration of human and social progress

[3] Kolakowski, Leszek. *Modernity on Endless Trial.* University of Chicago, 1960, 4.

in all fields—above all, industrial, scientific, and architectural. The big attractions were Machinery Hall, the Electricity Building, the Electric Fountain, and the halls on Transportation and Mines. On the opening day, US President Grover Cleveland was on hand to push a button that turned on an abundance of electric lights and motors. (Electric light bulbs and ac motors were still fairly new.) This caused such an excited forward surging of the thousands of spectators that many fainted in the crush. One may safely assume that few among the twenty-seven million attendees during the Exposition worried about, say, the ill effects of rapid industrialization. And few if any would have guessed that, just a century later, at a World's Fair held in South Korea, the official US exhibit, as if in obeisance to a new *Zeitgeist*, was reportedly dedicated entirely to the detritus of the post-industrial world, featuring mounds of broken machinery and views of festering nuclear disposal sites; or that the current and permanent exhibition at Washington's Smithsonian Institution Museum of American History, "Science in American Life," devotes the major part of its space to an exposé of the hazards of science and the public's alleged disillusionment with technology.

Another indication of how much the worldview changed during one century is that one of the major events of the Exposition of 1893 was a spectacular World's Parliament of Religions. Personal religion is, and always has been, close to the hearts of most Americans. But it now seems surprising that in a setting glorifying science and industry, hundreds of religious leaders from all parts of the world met to present their views in two hundred sessions during seventeen days. It was a colorful affair, with Hindus, Buddhists, Jains, Jews, Protestants, Catholics, adherents of Shinto and Zoroaster, and so forth, all meeting together in their robes "for a loving conference," in the words of the chairman of the Parliament, J. H. Barrows. The purpose was clear. As it was for the Exposition as a whole, the subtext of that Parliament of Religions was also progress and harmonious unity. Hence the Exposition, Barrows said, could exclude religion no more than it could exclude electricity. Science was invoked as an ally in reaching a higher unity while serving the needs of mankind.

One of the passionate believers that science, religion, and indeed cultural activities are aspects of one grand unification program was one of the organizers of the Parliament of Religions, Paul Carus, a publisher now remembered mainly for having brought the writings of Ernst Mach to readers in the US. The title of his presentation[4] was nothing less than "Science, a Religious Revelation." His was a sort of anticlerical post-Christian Deism, much of which would have appealed to some American statesmen-philosophers of an earlier century. Individual dignity, Carus thought, can only be found through the discovery of truth, and that is the business of science. Hence, he announced, "through science, God speaks to us." One did not have to choose between the Virgin and the Dynamo; rather, the laboratory was the true cathedral, and vice versa. As the masthead of his journal The Open Court put it, he was "devoted to the science of religion [and] the religion of science."

Carus typified a popular, science-favoring universalism of that time, which today is severely challenged, both from the right and from the left. I have chosen Carus because his world picture was a good example of the then prominent movement, *Modern Monism*, based on the belief in a "unitary world conception." It arose essentially as an anti-thematic response against the Cartesian dualism of the material vs. the mental, and against the multiplicity of common sense experience, with its starting point in personal individuality. The movement on behalf of Monism had the enormous ambition, in the words of Carus, "to direct all efforts at reform, and to regenerate our entire spiritual life in all its various fields." This meant of course replacing conventional religion with what Carus called the "Religion of Truth," where Truth is defined as "the description of fact [...] ascertainable according to the methods of scientific inquiry." In this sense, "science *is* revelation"; and in this way one would overcome the old, unacceptable dualism of scientific truths vs. religious truths.

The head of the small but ambitious international Monistic movement was the great German chemist Wilhelm Ostwald (Nobel Prize, 1909). Whereas most modern scientists are quite aware of the limits even within their research—as Max Planck said in 1931, "a science is never in a position completely and exhaustively to solve the problem it has to face"—the publications of the Monistic movement show that it hoped every aspect of culture, life, and society would be guided by Monistic ideas, from the education of children to the economy of nations, and of course within the research program of science itself. Thus Ernst Haeckel, another patron of the movement, predicted that physical science would eventually trace back all matter to a "single original element."

Despite the philosophical naïveté of the leaders, the movement attracted for a time an enthusiastic following. In Germany, it had branches in forty-one cities, and even organized public mass demonstrations against the Church. One must perhaps allow for the effects on them of having had to live under the

[4] Barrows, John Henry, ed. *The World's Parliament of Religions.* Chicago: The Parliament Publication Co., 1893. Vol. II: 978-981.

reactionary political clericalism of Germany. But I have intentionally chosen this case of "scientism," of excess on the side of a small minority of scientists, as my first example of *the rhetoric of a polarizing over-reaching by many movements, before and since, on either side*. Thus, caught up in this fervor, Ostwald, with hubris unequaled by the few remaining captives of scientism today, was propelled to the heights of over-ambition, with such claims as these in 1911: "We expect from science the highest that mankind can produce and win on this earth. [...] Everything that mankind, in terms of its wishes and hopes, its aims and ideals, combines in the concept God, is fulfilled by science." And finally, "Science, now and with immeasurable success takes the place of the divine." Ostwald added the prophecy that "we see arrive the Monistic Century. [...] It will inaugurate a new epoch for humanity, just as 2,000 years ago the preaching of the general love for humanity had inaugurated an epoch."[5]

But soon after this publication, neither the Monistic nor the Christian base for kindness and love of fellow man had triumphed. Instead, war, which William James called the inevitable "bloody nurse of history," had taken charge. Strangely enough, it was Henry Adams who had sensed that the trend would be ultimately against a Monistic Century. Writing in 1905 in his autobiography, *The Education of Henry Adams*, he identified the course of history as away from Unity and toward fragmentation and multiplicity. Indeed, in the aftermath of World War I, the idea of progress and optimism about the place of science in culture were war casualties. The balance had swung the other way. The only major movement with large political ambition that continued to claim a scientific basis was of course Marxism, especially as defended by Lenin in his 1908 book, *Materialism and Empirio-Criticism*. The assertion that Marxism-Leninism, the founding ideology of the Soviet Union, had anything to do with real science is a rhetorical device, one of that century's great delusions even if this propaganda was taught to every child in Communist countries. It is disproved, not least by the faulty analysis of science and its philosophy in Lenin's own book, and by the widespread mistreatment that Soviet scientists experienced when their theories did not please their government.

Spengler's prediction of the end of science

Perhaps the most widely read attack against the claims of optimistic science appeared as the war was ending in 1918, and later it deeply influenced such theoreticians of history as Arnold Toynbee and Lewis Mumford. The book was *The Decline of the West*, written by a German mathematics teacher, Oswald Spengler. No quick summary can do justice to that richly baroque work, but the point I want to focus on here is what it had to say about the topic before us. Spengler's key conception was that for every part of mankind, in every epoch since Egypt, Greece, and Rome, the history of a civilization has taken fundamentally the same course, and this will continue in the future. Thus our own inevitable destiny in the West is to go to dust according to a timetable that he thought he could calculate from the available precedents. Spengler predicted the very date of our undoubted demise: the year 2000.

The end stages of every civilization, he wrote, can be recognized by the ideas science treasures in its own progress—by the adoption of the notion of causality instead of destiny; by attention to abstractions such as infinite space and to cause and effect, rather than to "living nature." The primacy of the soul is replaced by intellect; mathematics pervades more and more activities; and nature is reinterpreted as a network of laws within the corpus of what Spengler calls "scientific irreligion." Here Spengler introduces his most startling idea, one that has become familiar in new garb also. He warns that it is characteristic of the winter phase of civilization that precisely when high science is most fruitful within its own sphere, the seeds of its own undoing begin to sprout. This is so for two reasons: the authority of science fails both within and beyond its disciplinary limits; and an antithetical, self-destructive element arises inside the body of science itself that will eventually devour it.

The failure of science's authority outside its laboratories, he says, is due in good part to the tendency to overreach and misapply to the cosmos of history the thinking techniques that are appropriate only to the cosmos of nature. Spengler holds that the thought style of scientific analysis, namely "reason and cognition," fails in areas where one really needs the "habits of intuitive perception," of the sort he identifies with the Apollonian soul and the philosophy of Goethe. By asserting that an unbridgeable contrast exists between a pure "rationality" of abstract science and the intuitive life as lived, Spengler commits the same error as all such critics before him and after, to this day, of whom few seem even to have come closer to science than through their school textbooks. Therefore they are ignorant of the vast difference between, on the one hand, "public science"—the final results of intersubjective negotiations to fashion at least a temporary consensus and globalization on the basis of experiment and logic, and on the other hand, the earlier, "private" stage of work in science, where the particular researcher's intuitive, aesthetic, thematic or other non-logical preference may be the key to

[5] Ostwald, W. *Monism as the Goal of Civilization*. Hamburg, International Committee of Monism, 1913, 37.
The section that follows is an abstract of much of Chapter 5, "The Controversy over the End of Science," in Ref. 1.

the individual's advance beyond the previous level of public science. The complementarity between these two quite different stages in the actual development of any scientific result explains why in any given field the findings by natural scientists, operating within vastly different cultures and styles, are eventually harnessed into common products with (for a time) global validity.

All this may be clear enough to practicing scientists. But, Spengler continues, even in the cosmos of nature there is an attack on the authority of science, arising from within its own empire: Every conception is at bottom "anthropomorphic," and each culture incorporates this burden in the key conceptions and tests of its own science, which thereby become culturally conditioned illusions. All our rushing after positive scientific achievements in our century only hides the fact, he thinks, that as in classical times, science is once more destined to "fall on its own sword," and so will make way for a "second religiousness."

What Spengler termed the orgy of two centuries of exact sciences would shortly be ending, together with the rest of what was valuable in Western civilization. As a kind of postscript, Spengler added his opinion in his later book, (*Man and Technics* 1931), that advancing technology, with its mindlessly proliferating products, will also turn out to undermine the society of the West—because, he prophesied, its interest in and support of science and engineering will decrease: the "metaphysically exhausted" West will not maintain advances in these fields. Instead, the previously overexploited races in the rest of the world, "having caught up with their instructors," will surpass them and "forge a weapon against the heart of the Faustian [Western] Civilization." The non-Caucasian nations will adopt the technical skills, excel in them, and turn them against the Caucasian originators. In short, as H. Stuart Hughes put it, Spengler's prediction was that the East will triumph through better technology, first in commerce, and then militarily.[6]

A "scientific world conception" —the Vienna Circle

The early response to Spengler's diagnosis was predictably bimodal—on one side there was wide and enthusiastic acceptance, which continues among people today who have never read Spengler but, so to speak, have imbibed his ideas with their mother's milk. On the other side, the opponents of Spenglerian scenarios included of course many prominent scientists. Some of these had joined in a study group that called itself the Vienna Circle, which met in the 1920s and early '30s for discussion and publication. It included Moritz Schlick, Rudolf Carnap, Philipp Frank, Kurt Gödel, and Otto Neurath. Among their active sympathizers, they could count Hans Reichenbach and Richard von Mises in Germany, and in America, B. F. Skinner, P. W. Bridgman, Charles Morris, and W. V. Quine.

The most influential publication of the core group was a slim pamphlet issued in October 1929 as a kind of manifesto of the movement, the main title being nothing less than *The Scientific Conception of the World*.[7] The very title was a trumpet blast in the fight to change the balance again, to put science back at the center of modern culture, and against what the booklet called, in the first sentence, the chief alternative, the tendency toward metaphysical and theologizing thought, those old helpmates of the Romantic movement.

Although most of the scholars involved in the Vienna Circle concerned themselves chiefly with the study of the epistemological and logical problems at the foundations of science, there was a clear undercurrent of wider cultural, social, political, and pedagogic ambitions as well. For, as the manifesto said, "The attention toward questions of life is more closely related to the scientific world conception than it might at first glance appear. [...] For instance, endeavors toward the unification of mankind, toward a reform of school and education, all show an inner link with the scientific world conception. [...] We have to fashion intellectual tools for everyday life. [...] The vitality that shows itself in the efforts for a rational transformation of the social and economic order permeates the movement for a scientific world conception, too." (Carnap et al. 1929, 304-305.)

The members of the Circle associated themselves explicitly not with the Platonists and Pythagoreans, but with the Sophists and Epicureans, "with those who stand for earthly being, and the Here and Now." A science free from metaphysics would be a unified science; it would know no unsolvable riddles; it would train thinking to produce clear demarcations between meaningless and meaningful discourse, between intellect and emotion, between the areas of scientific scholarship on the one hand and myth on the other. Just as this approach would, by this formulation, clarify the foundations of mathematics, of the physical sciences, of biology and psychology, it would also demystify the foundations of the social sciences, "and in the first place [...] history and economics." The empiricist, antimetaphysical attitude would help to reject such dangerous conceptions as "folk spirit," and would "liberate from inhibiting prejudices."

Thus, the "debris of millennia" would be removed, and "a unified picture of this world" would emerge, free from magical beliefs. The social and economic struggles

[6] Hughes, H. Stuart, and Oswald Spengler. *A Critical Estimate*. New York: Charles Scribner's Sons, 1952.

[7] Carnap, Rudolf, Hans Hahn, and Otto Neurath. *Wissenschaftliche Weltauffassung: Der Wiener Kreis.* Vienna: Artur Wolf Verlag, 1929. For an English translation see Otto Neurath, *Empiricism and Sociology* (Dordrecht, Reidel, 1973). The page references are to the English translation; I have made occasional corrections in the translation, as necessary.

of the time would be ameliorated because the "broad masses of people" would reject the doctrines that have misled them. (Carnap et al., 315-317.) Beyond that, the spirit of the scientific world conception would penetrate "in growing measure the forms of personal and public life, in education, upbringing, architecture, and the shaping of economic and social life according to rational principles." And the manifesto for a new modernity ended with the blazing formulation, in italics: "The scientific world conception serves life, and life receives it" (Carnap et al., 318).

Perhaps the most carefully developed of the many publications expressing the Circle's position on science and its rationality as the keys to a sane world picture was the major book by Richard von Mises, the Austrian scientist, mathematician, engineer and philosopher (as well as scholar of the poet Rainer Maria Rilke). Von Mises entitled his big volume, with a bit of irony, *Kleines Lehrbuch des Positivismus*.[8] The aim was not only to show what an empiricist-rational scientific world conception would consist of, what its tools would be, and what problems it could solve within the sciences, from mathematics and physics to biology and the social sciences. All this is done in great detail; but an equally motivating force was to present thereby a choice from the then-reigning alternatives in German-speaking Europe: the Kantianism in Germany and the clerical-metaphysical trend in Austria, both of which were then being interspersed with the growing totalitarian ideologies. Von Mises noted his quite explicit opposition to what he called "negativism," in which he includes systematic, philosophical, and political anti-intellectualisms that have remained part of the present landscape. Among the examples he cited were, in fact, Oswald Spengler, and the once-popular German philosopher Ludwig Klages, whose point of view was enshrined even in the title of his main work, *The Mind as Enemy of the Soul*.

As a sign that von Mises' main aim of the book was to put science at the center of a healthy culture in the largest meaning of the term, his volume dealt at length with the way the scientific world conception would illuminate the understanding of metaphysics, poetry, art, the law, and ethics. The underlying commonality of the various forms of cultural achievements was considered by von Mises to be due to the principal unity of their methods, if carried through rationally and soundly. The original readers of the book must have felt themselves to be in the presence of an updated follower of Auguste Comte. The very last sentence is, as it were, the summary of the whole project: "We expect from the future that to an ever-increasing extent scientific knowledge, i.e., knowledge formulated in a connectable manner, will regulate life and the conduct of man." (Von Mises 1951, 370)[9]

Freud: Instinctual passions versus reasonable interests

But now we shall see the lever of sentiments shift the balance once more, and indeed on the very issue of whether knowledge formulated in a scientific manner can lead mankind to saner and more rational conduct. In 1929, the same year in which the optimistic manifesto of the Vienna Circle was published, Sigmund Freud, writing in the same city, produced a book of his mature years, giving his somber and pessimistic answer. To the founder of psychoanalysis, the role of science in our culture had been a continuing preoccupation, and in 1911 he had still been optimistic enough to sign the *Aufruf* of the Society for Positivistic Philosophy. But in that book of late 1929, *Das Unbehagen in der Kultur*,[10] Freud found that science, while counting among the most visible manifestations of civilization, was at best an ameliorating influence in a titanic struggle on which the fate of our culture depended. That struggle, he said, was centered on mankind's often-doomed effort to master "the human instinct of aggression and self-destruction." Even at that time he saw, in the last paragraph of the book, that "Mankind has gained control over the forces of nature to such an extent that with their help it may have no difficulty to exterminate one another to the last man." (Freud 1929, 92.)

Freud held that the restrictions which civilization imposes upon the demands of our instincts produce an irremediable antagonism between those fetters versus the innate "Destructive Instinct," or "Death Instinct" (Freud 1929, 7, 8.), the drive that is constantly at odds with the civilizing project to elevate the moral condition of mankind. He wrote, "...man's natural aggressive instinct, the hostility of each against all, and of all against each, opposes this program of civilization. This aggressive instinct is the derivative and the main representative of the death instinct which we have found alongside of Eros, and which shares world-domination with it. And now, I think, the meaning of the evolution of civilization is no longer obscure to us. It must present the struggle between Eros and Death, between the instinct of life (*Lebenstrieb*) and the instinct of destruction (*Destruktionstrieb*), as it works itself out in the human species. This struggle is what all life essentially consists of, and the evolution of civilization may therefore be simply described as the struggle for life of the human species. And it is this battle of the giants that our nursemaids try to appease with their lullaby about Heaven." (Freud 1929, 69.)

[8] He allowed a simpler title, *Positivism: A Study in Human Understanding*, for the English translation (Harvard University Press, 1951).

[9] The word "control," used in the English edition, has been corrected to "regulate," which corresponds to the German edition.

[10] Seriously mistranslated into English as *Civilization and its Discontents*. New York: W.W. Norton, 1961.

In this conflict, scientific and other cultural activities result from the successful if incomplete "sublimation of instinctual aims," making science at first glance merely a "vicissitude which has been forced upon the instincts by civilization." The accomplishments of science and technology originated as welcome tools in the effort to protect men against the hostile forces of nature; they have now become "cultural acquisitions" that "do not only sound like a fairy tale, they are actual fulfillments of every—or almost every—fairy tale wish." They verge on our attaining the old ideas of "omnipotence and omniscience." Man "has, as it were, become a kind of prosthetic God." (Freud 1929, 38-39.)

But there's the rub: Happiness still eludes him. "Present-day man does not feel happy in his God-like character," either individually or in terms of the group. That again has its reason in the fact that "civilization is built upon a renunciation of instinct," such as sexuality and aggressiveness, and "presupposes precisely the nonsatisfaction (by suppression, repression, or some other means) of powerful instincts." Hence, the "cultural frustration" (*Unbehagen*) which dominates the whole field of social relationships between human beings (Freud 1929, 43–44, 62).

Freud's pessimistic conclusion follows: "In consequence of this primary mutual hostility of human beings, civilized society is perpetually threatened with disintegration. The interest of work in common would not hold it together; instinctual passions are stronger than reasonable interests. [...] In spite of every effort these endeavors of civilization have not so far achieved very much. [...] It is always possible to bind together a considerable number of people in amity, so long as there are other people left to receive the manifestations of their aggressiveness," as in religious or ethnic persecution (Freud 1929, 59, 61).

During the decades since this was written, modern history has all too often seemed to be the experimental verification of Freud's dark assessments, according to which neither science nor any other cultural activity can fully displace our animal nature from its central position, but can only delay the ultimate fate that threatens.

Scientists as "betrayers of the truth"

Let us now turn to the more recent period. We are familiar with the fluctuations, during the 1960s and 1970s, of opinion in academe and among the public regarding the interactions of science and society. But starting in the early 1980s, a new and powerful element entered into this discussion, which has recently been assuming ever greater attention and institutionalization, at least in the US. The new element, the new force adding to the derogation of the credibility of science, is the insistence from some quarters—which increasingly has fallen on receptive ears among the population—that to a previously completely unrealized degree the pursuit of science is, and has been all along, ever since the days of Hipparchus and Ptolemy, thoroughly corrupt and crooked. Consequently severe measures must be applied to the practice of science from outside. This assertion, which has become louder and louder over the past few years in books, official reports, and hundreds of articles, has spawned dramatic public hearings, the formation of specific government agencies, university bureaucracies, and quite a few careers. The safeguarding of ethical practices and uses of science, of which there has been a long tradition within the scientific community, is now to be put into better and wiser hands.

A striking, pace-setting example of this assertion was the book by two influential New York Times science editors, William Broad and Nicholas Wade. It states its intention in the title on the jacket, *Betrayers of the Truth: Fraud and Deceit in the Halls of Science*,[11] and follows up with the unqualified canon shot of the opening sentence: "This is a book about how science really works." Going far beyond the need to expose the relatively few rotten apples in any barrel, which the scientific community itself has long recognized as necessary, if only for the sake of its own health, this kind of rhetoric has become commonplace. As this book and its many followers proclaim, the relatively few, sad cases of real or alleged misbehavior are the litmus test for the whole enterprise. Fraud and deceit are depicted as being part of the very structure of scientific research.

Similarly, the report to Congress by the Congressional Research Service, entitled *Scientific Misconduct in Academia*, stated that, more and more, "the absence of empirical evidence which clearly indicates that misconduct in science is not a problem [...] suggests that significant misconduct remains a possibility." Among all the targets to preoccupy those who are charged with timely attention to misconduct damaging our republic, this formulation singles out the conduct of science as being guilty until proved innocent. Moreover, the tendency has recently been to include in the allegation of *scientific* misconduct not only falsification of data, plagiarism, and the like, but also the spectrum of misdeeds more common to flawed mankind generally, and for which sanctions have existed, e.g., "use of university resources for inappropriate purposes, sexual harassment, racial discrimination," etc.[12]

Similarly, the Office of Scientific Integrity Review (OSIR) of the Department of Health and Human Services made part of its proposed definition of

[11] New York: Simon & Schuster, 1982.

[12] *Science*, 1993. Vol. 26: 1203.

"misconduct" in science, apart from fabrication, falsification, and plagiarism, "practices that seriously deviate from those that are commonly accepted in the scientific community." (Federal Code: 42 C.F.R. 50.102.) The intention here may have been to parallel the way the Supreme Court defined obscenity by reference to the current standards of the local community. However, when it comes to making progress in science, some practices contrary to those common at the time have again and again been the very hallmark of needed innovations—from putting mathematics into physics in the seventeenth century, to the introduction of quanta, which pained even the originator, Max Planck himself, and to the more recent innovation of modern teamwork. The proposed definition of misconduct, with its potential for mischief, was one more example of the gap between the culture of science and the culture outside the lab. One should add that to her credit the director of the National Institutes of Health at the time intervened on that point, objecting that such a community standard "would have implicated even the discoverer of penicillin, who serendipitously found good use for bacteria growing in a contaminated lab dish."[13]

The power of the generalized allegations against the conduct of science has two components. The first is of course the astonishing claim that basic research scientists in considerable numbers are intentionally false to their own most fundamental avowed mission, namely, to the pursuit of truths; in other words, that not just a few apples are rotten, but that the whole barrel is.

Yet, even in the presence of the occasional scandalous misdeeds by a relatively small number of the world's millions of scientific researchers, the vastly overblown allegation of pervasive and ingrained fraud and deceit in science would not have been taken so seriously that in the US the newspapers, college courses, training courses for scientists and physicians, commissions, Congressional committees, scientific societies, and so on, are now massively and expensively preoccupied with the institutionalization of the prevention of misconduct in science. The unrelenting accounts of specific incidents, some outrageous, more of them sensationalized, have left some of the public and legislators feeling that a great plague of dishonesty had invaded all academic laboratories. As the journal *Nature* noted shrewdly, the current trend is resulting in "a slow—and Hollywood-assisted—erosion of [the scientists'] public image, [...] [replacing it] in the public mind by a money-grabbing plagiarizing con-artist."[14] *Time* magazine chimed in with an essay on scientists, beginning with, "Scientists, it seems, are becoming the new villains of Western society." A raft of best-selling books add up the allegations in diatribes that have the frank aim, in the words of Bryan Appleyard's polemic *Understanding the Present: Science and the Soul of Man*, that science must be "humbled." We are, it appears, standing only on the shoulders of dwarfs.

What is getting lost in this avalanche of excitement, and also in the generally poor, even self-flagellating, responses from some scientific institutions, is some thorough inquiry into the actual rate of serious misconduct among scientists, the kind of empirical research that would yield a reasonable estimate of the likely relative rate of incidents. I have found only some scattered, preliminary steps in this direction, but these suggest that in fact the actual rate of misconduct (rather than suspected, alleged, or "perceived" without hard evidence) is remarkably *low*. Among the available, reasonably quantifiable measures is, for example, the National Library of Medicine finding that for the period of 1977 to 1986, when about 2,780,000 articles were published in the world's biomedical literature, 41 of these had to be withdrawn because fraudulent or falsified data appeared in them—a rate of under two one-thousandths of one percent of scientific publications per decade.[15] Other data support the same point. Thus the Food and Drug Administration, responding to allegations or evidence of misconduct in clinical research with investigational new drugs research, submitted twenty cases of suspected fraud or other criminal violations to the US Attorney General's office. These resulted in thirteen convictions of clinical investigators—about one per year, on the average.[16]

Nobody does or should condone even a single case. But even if the actual rate were as much as a hundred times greater than these figures indicate, the intellectually most interesting questions would be, first, why science as a whole progresses so well despite being the work of mere human beings; second, how small the number of alleged misconduct is in this field compared with those in others, ranging from the world of finance, law, industry, journalism, and government at every level. And third, why the few cases of highly publicized charges of misconduct in science can so severely undermine the trust and confidence of the public and its representatives in the integrity of research in general.

Science as myth

The answer to those questions is in good part that there is indeed another, reinforcing reason for the widespread success of assaults on the credibility of scientific research. This second line of attack has been opened up by a loose assemblage made up of a branch of contemporary philosophy of science and other humanists, some of the so-called "strong-program" constructivist portion of sociology, of a small

[13] As reported in the *Washington Post*, March 20, 1992.

[14] *Nature*, January 6, 1994. Vol. 367: 6. Unlike most scientific journals in the US, *Nature* has been alert to the likely damage of the imbalance in reporting. See for example the editorial by John Maddox of March 17, 1994, in *Nature*, vol. 368: 185. It is noteworthy that another among the few who have spoken out against the growing tide of easy condemnation is also a trained science journalist, Barbara J. CULLITON, in her essay, "The Wrong Way to Handle Fraud in Science." *Cosmos,* 1994, 34–35. For an argument on the costs to science that may result from the excesses of distrust in science, see Steven SHAPIN, "Truth, Honesty, and Authority of Science," in the National Academy of Sciences report *Society's Choices: Social and Ethical Decision Making in Biomedicine* (Washington, DC, National Academy Press, 1994).

[15] The data were kindly furnished to me by Donald A. B. Lindberg, Director, National Library of Medicine. These cases are quite different from the laudable practice of scientists publishing correction notices when they find it necessary to draw attention to their own unintended errors. Eugene GARFIELD, "How to Avoid Spreading Error," *The Scientist,* 1: 9, 1987, reports that "of the 10 million journal items indexed in the *SCI* [Science Citation Index] since its inception, over 50,000 were coded as explicit corrections. [...] These vary from corrections of simple typographical errors to retractions of and outright apologies for 'bad' data or data that cannot be verified." This indicates a rate of 0.5 percent for such volunteered corrections of errors.

[16] In addition, the Office of Research Integrity of the US Public Health Service recently announced that starting about a year ago and looking back, it has found a total of 14 researchers guilty of some form of scientific misconduct out of about 55,000 researchers receiving PHS support per year. (Private communication of July 20,1993, from Lyle W. Bivens, Acting Director, ORI.) The

subset of the media, of a small but growing number of governmental officials and political aspirants, and of a vocal segment of literary critics and political commentators associated with the avant-garde of the postmodern movement. This is a potent and eloquent collective of just the sort that in the past has successfully challenged the worldview of their time and place.

The overall message evolving from that direction is no longer based only on stories of unacceptable behavior among a few scientists. The charge has been generalized and made even more serious: Put in starkest terms, the claim is that the most basic fraud committed by the members of the scientific community is their assertion *that there are any truths to be found at all*. For there really is nothing there even to betray and falsify; and consequently, science is inherently not corrigible, even if all misconduct were eliminated.

From that point of view, the business of science is mainly self-serving; for example, building and operating expensive institutions that claim to be looking for objectively ascertainable information about entities like quarks and bosons—which, however, are nothing more than "socially constructed" fictions. Against the naive realism that most scientists still embrace, and the agnosticism of the more sophisticated ones, the new critics counterpoise the radical solution: as one sociologist of science put it recently, "There is no Nature; there is only a communication network among scientists." The literature in academe is now full of statements such as "science is a useful myth," or "we must abolish the distinction between science and fiction," or "science is politics by other means."[17]

Scientists have tended to adopt the Baconian view that the acquisition of basic knowledge of causes and interrelations of phenomena—by processes not easily predictable or fully understood—can yield power over those of nature's forces that cause our burdens and ills. But now, the new consortium tells us, the arrow really goes the other way: not from knowledge to power, but from power to knowledge, and to a rather questionable knowledge at that. The attempts to find generally applicable, shareable knowledge about what might be called reality—through the use of both the rational and the intuitive faculties of individual scientists, and through their skeptical but collaborative attempt to achieve some consensus—were not only doomed exercises, but ironically have led to the disasters that have marked the century. The whole modern era, launched under the flag of progress, has only led to tragedy. The extreme over-optimism of a Herbert Spencer or a Friedrich Engels can never be replaced by a soberer conception. Progress is illusion. The globalizing program of science—to find basic unities and harmony transcending the level of apparent variety and discord—is held to be completely contrary to the post-modern drive that celebrates individual variety and the equality of standing of every conceivable style and utterance, every group and competing interest. Ours is the time to face the end of the search for foundations, the "End of the Modern Era." We are in a state called the "objectivity crisis"—a fashionable phrase found in the titles of learned conferences and in policy-setting documents to be examined shortly.

Together, these slogans of the newly emerging sentiment indicate that the aim is not merely a call for the improvement of practice or for increased accountability, which is appropriate and being pursued through earnest actions, but at bottom is, for the main branch of the movement of critics, the delegitimation of science as one of the valid intellectual forces, a reshaping of the cultural balance, as we shall see in more detail below. In this respect, there is a big difference here compared with the history of *internal* movements of protest, such as those of the logical positivists within philosophy, the Impressionists or Dadaists within art, the modern composers within music, etc. In all those cases, it was some of the best talent in the field that took up the task of renewal. Not so here—the motivating force is not renewal from within, but radical cultural politics from without.[18]

The Romantic Movement's challenge

Here we meet a clarifying fact: the contest before us is not new, but draws on historic forces of great strength and durability. Therefore it will be instructive to trace some of the individual steps and stages in this remarkable development of the growing new view, so as to make it easier to extrapolate and to preview the new terrain we may have before us. While I can here only point briefly to a few recent milestones, I shall seek documentation in the recent writings of some of the most distinguished thinkers, rather than, say, through representatives of the Dionysian undercurrent.

Our first informant and guide is Isaiah Berlin, widely regarded as a most sensitive and humane historian of ideas. The collection of his essays, published as the fifth volume of his collected papers,[19] opens with a startling dichotomy. He writes: "There are, in my view, two factors that, above all others, have shaped human history in this [the twentieth] century. One is the development of the natural sciences and technology, certainly the greatest success story of our time—to this great and mounting attention has been paid from all quarters. The other, without doubt, consists of the great ideological storms that have altered the lives of

cases involved work that ranged over a considerable period; for example, one of them began in 1977. To get a sense of the low yield of the allegations, and the pedestrian rather than sensational nature of most of the cases, see Office of Research Integrity, *Biennial Report 1991–92*, September 1993, US Dept. of Health and Human Services. To glimpse the enormous complexity, cost, and labor as well as the fragility of the process of adjudicating allegations of scientific misconduct, see for example the 63-page document, obtainable from the US Department of Health and Human Services, entitled: "Departmental Appeals Board. Research Integrity Adjudications Panel. Subject: Dr. Rameshwar K. Sharma, Docket No. A-93-50, Decision No. 1431, Date: August 6, 1993."

[17] For a scholarly and even-handed treatment of the spectrum of the varied interests of sociologists of science, see Zuckerman, Harriet. "The Sociology of Science" in Neil J. Smelser, ed., *Handbook of Sociology*. Beverly Hills, CA: Sage Publications, 1988, 511-574.

[18] For a thoughtful analysis, see Searle, John R. "Rationalism and Realism, What is at Stake?" *Daedalus*, 1993, no. 4, vol. 122: 55-83. A recent book that aims to answer the various types of critics is Gross, Paul R., and Norman Levitt. *Higher Superstition: The Academic Left and its Quarrels with Science*. Baltimore, MD: The Johns Hopkins Press, 1994. It is also useful for its extensive bibliography. Another stimulating resource is FARRELL, Frank B. *Subjectivity, Realism and Postmodernism.* New York: Cambridge University Press, 1994.

[19] Berlin, Isaiah. "The Crooked Timber of Humanity." *Chapters in the History of Ideas*. New York: Random House, 1992.

virtually all mankind: the Russian revolution and its aftermath—totalitarian tyrannies of both right and left and the explosion of nationalism, racism and, in places, of religious bigotry, which, interestingly enough, not one among the most perceptive social thinkers of the nineteenth century had ever predicted." (Berlin, 1991, 1.) He adds that if mankind survives, in two or three centuries' time these two phenomena will "be held to be the outstanding characteristics of our century, the most demanding of explanation and analysis."

What might the author intend by so juxtaposing these two "great movements"? One's first temptation may be to see a connection through the fact that during World War II the ingenuity and frantic work of scientists among the Allies, supporting the valor of the Allied soldiers, brought an end to the totalitarian tyranny of that period, which might well have triumphed over the democracies and established itself at least throughout Europe.

But such a response would not be to the point here. What is on Isaiah Berlin's mind is quite different. As we follow his eloquent and subtle analysis, it dawns on the reader that science and tyranny, the two polar opposite movements which he holds to have defined and shaped the history of this century, are somehow intertwined—that the development of the modem natural sciences and technology may, *through the reactions against them*, have unintentionally and indirectly contributed to the rise of those "totalitarian tyrannies."

This stunning connection, to be sure, is never explicitly spelled out by the author. But we can glimpse the implicit argument later in the book, in his chapter significantly entitled "The Apotheosis of the Romantic Will: The Revolt against the Myth of an Ideal World." There, Berlin summarizes the chronology of some basic concepts and categories in the Western world, specifically the changes in "secular values, ideals, goals." What commands his attention is the change away from the belief in the "central core of the intellectual tradition [...] since Plato," and toward a "deep and radical revolt against the central tradition of Western thought" (Berlin, 1991, 208), a revolt which in recent times has been trying to wrench Western consciousness into a new path.

The central core of the old belief system, one that lasted into the twentieth century, rested on three dogmas that the author summarized roughly as follows. The first is that "to all genuine questions there is one true answer, all others being false, and this applies equally to questions of conduct and feeling, to questions of theory and observation, to questions of value no less than to those of fact." The second dogma is that, "The true answers to such questions are in principle knowable." And the third: "These true answers cannot clash with one another." They cannot be incommensurate, but "must form a harmonious whole," the wholeness being assured by either the internal logic among or the complete compatibility of the elements. (Berlin 1991, 209-211.)

Out of these three ancient dogmas both institutionalized religions and the sciences developed to their present form (although one might add that modern scientists, in their practice, have become aware of the need for proceeding antidogmatically, by conjecture, test, refutation, and assaying probability). In their pure state, these systems are utopian in principle, for they are imbued by the optimistic belief, inherent in and derivable from the dogmas, that "a life formed according to the true answers would constitute the ideal society, the golden age." All utopias, Isaiah Berlin reminds us, are "based upon the discoverability and harmony of objectively true ends, true for all men, at all times and places"—and by implication the same is true for scientific and technical progress, which are aspects of our drive toward what he calls "a total solution: that in the fullness of time, whether by the will of God or by human effort, the reign of irrationality, injustice, and misery will end; man will be liberated, and will no longer be the plaything of forces beyond his control [such as] savage nature...." This is the common ground shared by Epicurus and Marx, Bacon and Condorcet, the Communist Manifesto, the modern technocrats, and the "seekers after alternative societies." (Berlin 1991, 212-123.)

But, Isaiah Berlin now explains, this prominent component of the modern world picture is precisely what was rejected in a revolt by a two-centuries-old counter movement that has been termed Romanticism or the Romantic Rebellion. From its start in the German *Sturm and Drang* movement of the end of the eighteenth century, it grew rapidly in Western civilization, vowing to replace the ideals of the optimistic program, based on rationality and objectively true ends, by the "enthronement of the will of individuals or classes, [with] the rejection of reason and order as being prison houses of the spirit."

My own favorite summary of the view of science and its disvalue in nineteenth-century literature is the anti-hero in Ivan Turgenev's gripping novel, *Fathers and Sons*. One of the greatest figures of Russian literature, together with Gogol, Dostoevski, and Tolstoy, Turgenev was a poet largely in the tradition of nineteenth-century Romanticism, inspired by Goethe, Schiller, and Byron, among others. *Fathers and Sons* was published in 1861. Its main figure is Yevgeny Vassilevich Bazarov, a university student of the natural sciences, expecting to get his degree as a physician shortly. Being a scientist who "examines everything from a critical point of view," he confesses himself also to be ideologically and politically a nihilist, the natural consequence of not acknowledging any external authority. All talk of love, or the "mystic relationship between a man and a woman," is to him just "romanticism, humbug, rot, art." It would be better to study the behavior of beetles. Even on his vacation he has brought along a microscope and fusses over it "for hours at a time." Reading Pushkin, he says, is for little boys. He thinks it would be much better to start with Ludwig Büchner's *Force and Matter*, a book published in 1855 and embodying such a flagrantly materialistic view that Büchner was forced to resign from his professorship in Germany. (It is, as it turned out later, the very book Albert Einstein singled out in his *Autobiographical Notes* as one of the two or three that most impressed him as a boy, and caused him to turn to the pursuit of science.)

What matters, Bazarov claims, "is that two and two are four—all the rest is nonsense." When he meets a clever and beautiful woman, he startles his friend by saying that hers would be a beautiful body to examine—on a dissection table. As if in revenge, fate brings him to the bedside of a villager dying of typhus, and he is made to help in the postmortem. But he cuts himself with his scalpel, and soon he is on the verge of delirium, a case of surgical poisoning. As he is dying, he tries to keep hold on his kind of reality by asking himself aloud, "Now, what is 8 minus 10?" In short, he is a caricature recognizable throughout literature—except that the figure of the emotionally dysfunctional scientist, from Dr. Frankenstein to the crew of Dr. Strangelove, causes surgical sepsis not only in each of them, but also in all those around them.

Returning to Isaiah Berlin's account, it is striking that, as he notes, no one predicted that a form of the worldwide Romantic Rebellion would be what dominated "the last third of the twentieth century." The Enlightenment's search for generalizability and rational order is depicted by the rebels of our time as leading at best to the pathetic Bazarovs of science, and those must be replaced by the celebration of the individual, by flamboyant antirationalism, by "resistance to external force, social or natural." In the words of Johann Gottfried von Herder, the rebel shouts: "I am not here to think, but to be, feel, live!" (Berlin, 1991, 223.) Truth, authority and nobility come from having heroically suffered victimization.

This assertion of the individual will over sharable reason has undermined what Isaiah Berlin had called the three pillars of the main Western tradition. The Romantic Rebellion of course has also given us enduring masterpieces of art, music, and literature. But it originated, as it were, as an antithetical mirror image, created in reaction to the very existence of the earlier Enlightenment-based conception. In the apotheosis of the Romantic Will in our time, it glows forth as the alternative, the "romantic self-assertion, nationalism, the worship of heroes, and leaders, and in the end [...] Fascism and brutal irrationalism and the oppression of minorities." (Berlin, 1991, 225.) Moreover, in the absence of "objective rules," the new rules are those that the rebels themselves make: "Ends are not [...] objective values. [...] Ends are not discovered at all but made; not found but created."

As a result, "this war upon the objective world, upon the very notion of objectivity," launched by philosophers and also through novels and plays, infected the modern worldview. The "romantics have dealt a fatal blow" to the earlier certainties, and have "permanently shaken the faith in universal, objective truth in matters of conduct" (Berlin, 1991, 236-237)—and, he might have added, in science as well. As any revolt does, this one puts before us seemingly mutually incompatible choices. Just as with quite antithetical cases of excess such as Ostwald's, it is again either/or, rather than the needed complementarity of mankind's rational, passionate, and spiritual functions. One is reminded here of the fact that extremes tend to meet each other. Thus the poet William Blake, the epitome of the Romantic Rebellion—who called the work of Bacon, Newton, and Locke satanic—composed in his *The Marriage of Heaven and Hell* (1790) one of the "Proverbs" that reveal the credo of so many of the opposing actors in this story to this day: "*The road of excess leads to the palace of wisdom.*"

The Romantic Rebellion infuses state policy

Other authors provide verification and elaboration of the implications of Berlin's findings, and especially so of the ominous joining of the extremes of a Romantic Rebellion with irrational political doctrines. This was evident in the "Cultural Revolution" in Mao's China, in the USSR, and in other totalitarian systems. To glance at least at one telling example, the historian Fritz Stern has written about the early phases of growth of Nazism in Germany when there arose in the 1920s, in his words, the "cultural Luddites, who in their resentment of modernity sought to smash the whole machinery of culture." The fury over an essential part of the program of modernity, "the growing power of liberalism and secularism," directed itself naturally also against science itself. Julius Langbehn was one of the most widely read German ideologues in the 1920s, and Stern writes of him, "Hatred of science dominated all of Langbehn's thought. [...] To Langbehn, science signified positivism, rationalism, empiricism, mechanistic materialism, technology, skepticism, dogmatism, and specialization..."

Long before the Nazis assumed governmental power, some German scientists and other scholars demanded that a new science be created to take the place of the old one, which they discredited—a new "Aryan science," based on intuitive concepts rather than those derived from theory; on the ether, the presumed residence of the "*Geist*;" on the refusal to accept formalistic or abstract conceptions, which were reviled as earmarks of "Jewish science;" and on the adoption as far as possible of basic advances "made by Germans."

In a classic study,[20] Alan Beyerchen identified some of the other main pillars of Aryan science. There we find themes uncomfortably similar to those that are again fashionable. A prominent part of Aryan science was, of course, that science, as some would now say, is basically a social construct, so that the racial heritage of the observer "directly affected the perspective of his work." Scientists of undesirable races, therefore, could not qualify; rather, one had to listen only to those who were in harmony with the masses, the "*Volk*." Moreover, this *völkisch* outlook encouraged the use of ideologically screened non-experts to participate in judgments on technical matters (as in the *Volksgerichte*). The international character of the consensus mechanism for finding agreement was also abhorrent to the Nazi ideologues. Mechanistic materialism, denounced as the foundation of Marxism, was to be purged from science, and physics was to be reinterpreted to be connected not with the matter but with the spirit. "The Aryan physics adherents thus ruled out objectivity and internationality in science. [...] Objectivity in science was merely a slogan invented by professors to protect their interests." Hermann Rauschning, president of the Danzig Senate, quoted Adolf Hitler as follows:

> We stand at the end of the Age of Reason. [...] A new era of the magical explanation of the world is rising, an explanation based on will rather than knowledge: There is no truth, in either the moral or the scientific sense. [...] Science is a social phenomenon, and like all those, is limited by the usefulness or harm it causes. With the slogan of objective science the Professoriat only wanted to free itself from the very necessary supervision by the State.

That which is called the crisis of science is nothing more than the gentlemen are beginning to see on their own how they have gotten onto the wrong track with their objectivity and autonomy. A simple question that precedes every scientific enterprise is: Who is it who wants to know something, who is it who wants to orient himself in the world around him?[21]

One issue was how technology, so useful to the state, could be fitted into the Romantic idea. In recent times, many antimodern movements, including Fundamentalist ones, have embraced technology. But Philipp Lenard, an outstanding physicist but a chief cultural hero of Nazi propaganda, spoke for at least a minority when he said that the tendency of scientific results to prepare the ground for practical advances has led to a dangerous notion, that of man's "mastery" of nature: Such an attitude, he held, only revealed the influence of "spiritually impoverished grand technicians" and their "all-undermining alien spirit." This idea, too, had its roots in the centuries-old history of the rise of Romantic thought. Alan Beyerchen summarizes this section with the observation that "the romantic rejection of mechanistic materialism, rationalism, theory and abstraction, objectivity, and specialization had long been linked with beliefs in an organic universe, with stress on mystery [and] subjectivity..."

Because all these excesses were couched in phrases so reminiscent of currently used ones to delegitimate the intellectual authority of science, it is necessary to keep in mind that there is only a common ancestry of these views, rather than a necessarily causal connection between them. This applies also to the next case, as I turn now to the position embraced by another distinguished contemporary icon among humanists, although an advocate rather than an analyst. His writings on this topic are—like those of Oswald Spengler, or the positivists—of interest here not because they represent majority positions, which they do not, but because they have the potential for wide resonance at a turning point of sentiments. Also, in this case we shall see that the relation between modern natural science and the rise of totalitarianism, which Isaiah Berlin considered to be only the result of an obscene historic counterreaction, now receives a much more sinister interpretation: the two become directly, causally linked.

This ominous linkage has been argued repeatedly in writings by the Czech poet, playwright, resistance fighter against Marxist-Leninist oppression, and statesman—Václav Havel. In the passages to be discussed, we will notice that he subscribes to many of the themes discussed in Isaiah Berlin's analysis; but Havel's key point is that totalitarianism in our time was simply the perverse extreme end result of a trend of ideas embodied in the program of science itself. In this sense, Western science gave birth to Communism; and with the fall of the latter, the former has also been irremediably compromised.

Looking back on the twentieth century, other Central Europeans might characterize it as the release of the forces of brutal irrationality and bestiality, a reversion to

[20] Beyerchen, Alan. *Scientists under Hitler: Politics and the Physics Community in the Third Reich.* New Haven, CT: Yale University Press, 1977.

[21] Rauschning, Hermann. *Gespräche mit Hitler.* New York, Europa Verlag, 1940, 210. Mussolini expressed himself similarly.

ruthless autocracies in which the fates of millions were sealed by the whims of Kaiser Wilhelm, Hitler, Stalin, and their henchmen—rather than being the offspring of organized skepticism and the search for reasoned consensus, which are at the heart of science. But Havel finds the chief sources of trouble in that century to have been the very opposite, namely, the habit—in his words—of "rational, cognitive thinking," "depersonalized objectivity," and "the cult of objectivity." He advises us to take refuge now in unrepeatable personal experience, in intuition and mystery, and the other mainstays of the Romantic Rebellion. I must let him put his case at some length in his own words; for while he eschews the documentation or balanced account of the scholar, he is instead in fine command of the rhetoric of persuasion, the ease of unspecified assertions and generalizations, and of the chief art of the dramatist, the suspension of disbelief. The result, for many of his readers, is hypnotic acquiescence without questioning the generalities and leaps in the prose. The "end of Communism," he writes in one of his most widely quoted essays,

> ...has brought an end not just to the 19th and 20th centuries, but to the modern age as a whole.
>
> The modern era has been dominated by the culminating belief, expressed in different forms, that the world—and Being as such—is a wholly knowable system governed by a finite number of universal laws that man can grasp and rationally direct for his own benefit. This era, beginning in the Renaissance and developing from the Enlightenment to socialism, from positivism to scientism, from the Industrial Revolution to the information revolution, was characterized by rapid advances in rational, cognitive thinking. This, in turn, gave rise to the proud belief that man, as the pinnacle of everything that exists, was capable of objectively describing, explaining and controlling everything that exists, and of possessing the one and only truth about the world. It was an era in which there was a cult of depersonalized objectivity, an era in which objective knowledge was amassed and technologically exploited, an era of systems, institutions, mechanisms and statistical averages. It was an era of freely transferable, existentially ungrounded information. It was an era of ideologies, doctrines, interpretations of reality, an era in which the goal was to find a universal theory of the world, and thus a universal key to unlock its prosperity.
>
> Communism was the perverse extreme of this trend. [...] The fall of Communism can be regarded as a sign that modern thought—based on the premise that the world is objectively knowable, and that the knowledge so obtained can be absolutely generalized—has come to a final crisis. This era has created the first global, or planetary, technical civilization, but it has reached the limit of its potential, the point beyond which the abyss begins.
>
> Traditional science, with its usual coolness, can describe the different ways we might destroy ourselves, but it cannot offer truly effective and practicable instructions on how to avert them.[22]

A listener might at this point begin by objecting that these passages are built on immense overgeneralizations and illogical jumps, just as flawed as those of the extreme Monists were on the other side; or that at least on factual grounds the self-designation of Communist ideology as "scientific" was indeed a fraud. On this last point, the scholar of the history and philosophy of Soviet science, Loren Graham, made the trenchant observation: "In 1992, the playwright and President of independent Czechoslovakia, Václav Havel, wrote that the fall of communism marked the end of an era, the demise of thought based on scientific objectivity. [...] Was the building of the White Sea Canal in the wrong place and by the most primitive methods, at the cost of hundreds of thousands of prisoners' lives, the blossoming of rationality? Was the disregard of the best technical specialists' advice in the construction of Magnitogorsk, the Dnieper dam and the Baikal-Amur Railway a similar victory for objectivity? Was the education of the largest army of engineers the world has ever seen—people who would come to rule the entire Soviet bureaucracy—in such a way that they knew almost nothing of modern economics and politics an achievement of science? [...] And even long after the death of Stalin, into the 1980s, what was the Soviet insistence on maintaining inefficient state farms and giant state factories, if not an expression of willful dogmatism that flew in the face of a mountain of empirical data?"[23]

But one may doubt if Havel would reconsider his position, for the object of his essay is the conclusion, presenting the "way out of the crisis of objectivism," as Havel labels it. Only a radical change in man's attitude toward the world will serve. Instead of the generalizing and objectifying methods that yield shareable, repeatable, inter- or trans-subjective explanations, we must now turn, he says, to the very opposite, which presumably "science" somehow has totally banished from this world, i.e., to "such forces as a natural, unique, and unrepeatable experience of the world, an elementary sense of justice, the ability to see things as others do, [...] courage, compassion, and faith in the importance of particular measures that do not aspire to be a universal key to salvation. [...] We must see the pluralism of the world. [...] We must try harder to understand than to explain." Man needs "...individual spirituality, firsthand personal insight into things [...] and above all trust in his own subjectivity as his principal link with the subjectivity of the world..."

Despite Havel's hint, in passing, of a possible blending of the "construction of universal systemic solutions" or "scientific representation and analysis" with the authority of "personal experience," so as to achieve a

[22] "Politics and the World Itself," *Kettering Review*, Summer 1992, 9–11. His essay was also printed on March 1, 1992, in the *New York Times* as Havel's OpEd, entitled "The End of the Modern Era."

[23] Graham, Loren R. *The Ghost of the Executed Engineer: Technology and the Fall of the Soviet Union.* Cambridge, MA: Harvard University Press, 1993.

"new, postmodern face" for politics, Havel's identification of the "End of the Modem Era" is not to be understood merely as a plea for some compromise or coexistence among the rival constructs; that much was announced in an earlier and even sharper version of his essay, one which dealt with the place of modem science quite unambiguously and hence deserves careful reading:

> [Ours is] an epoch which denies the binding importance of personal experience—including the experience of mystery and of the absolute—and displaces the personally experienced absolute as the measure of the world with a new, manmade absolute, devoid of mystery, free of the 'whims' of subjectivity and, as such, impersonal and inhuman. It is the absolute of so-called objectivity: the objective, rational cognition of the scientific model of the world.
>
> Modern science, constructing its universally valid image of the world, thus crashes through the bounds of the natural world, which it can understand only as a prison of prejudices from which we must break out into the light of objectively verified truth. [...] With that, of course, it abolishes as mere fiction even the innermost foundation of our natural world. It kills God and takes his place on the vacant throne, so that henceforth it would be science that would hold the order of being in its hand as its sole legitimate guardian and be the sole legitimate arbiter of all relevant truth. For after all, it is only science that rises above all individual subjective truths and replaces them with a superior, trans-subjective, trans-personal truth which is truly objective and universal.
>
> Modern rationalism and modern science, through the work of man that, as all human works, developed within our natural world, now systematically leave it behind, deny it, degrade and defame it—and, of course, at the same time colonize it.[24]

Here we see the giant step that Havel has taken beyond Berlin's analysis: It is modern science itself that has been the fatal agent of the modern era. As if to answer Ostwald's excesses, it is held responsible even for deicide.

Many have been moved by Havel's powerful mixture of poetical feeling, theatrical flourish, and the bold waving of an ancient, bloodstained shirt. The summary of his ideas, published conspicuously under the title "The End of the Modern Era,"[25] made an immediate and uncritical impression on readers of the most varied backgrounds. Among them was one person especially well placed to ponder the values of science, and to draw conclusions of great import for the life of science in the US. Here we arrive at the last of the stages on the road to the current understanding of the place of science in our culture.

The person so deeply affected by Havel's piece was none other than the distinguished chairman of the US Congress Committee on Science, Space, and Technology, and one of the staunchest and most effective advocates of science during his long tenure in the House of Representatives: George E. Brown, Jr. of California. He acknowledged that he had received "inspiration" from Havel's essay, "The End of the Modern Era," and decided to reconsider his role as a public advocate of science. He therefore first wrote a long and introspective essay[26] under the title "The Objectivity Crisis," and then presented it to a group of social scientists in a public session at the annual meeting of the American Association for the Advancement of Science, under the title "The Objectivity Crisis: Rethinking the Role of Science in Society."[27]

Persuaded by Havel's version of the Romantic Revolt, Brown cast about earnestly for the consequences it should have for the pursuit of science in his country. As a pragmatic political leader, he was primarily concerned with how scientific activity may hold on to some legitimacy—by service to the nation in terms of visible "sustainable advances in the quality of life," "the desire to achieve justice" (which he says "is considered outside the realm of scientific considerations"), and all the other "real, subjective problems that face mankind." He now saw little evidence that "objective scientific knowledge leads to subjective benefits for humanity." The privileging of the claim of unfettered basic research is void too, he said, because all research choices are "contextual" and subject to the "momentum of history."

Moreover, science has usurped primacy "over other types of cognition and experience." Here Brown quoted Havel's definition of the "crisis of objectivity" being the result of the alleged subjugation of our subjective humanity, our "sense of justice, [...] archetypal wisdom, good taste, courage, compassion, and faith," the processes of science "not only cannot help us distinguish between good and bad, but strongly assert that its results are, and should be, value free." In sum, Brown held, it would be all too easy to support more research when the proper solution is instead "to change ourselves." Indeed, he came to the conclusion that "the promise of science may be at the root of our problems." To be sure, the energies of scientists might still find use if they were properly directed, chiefly into the field of education or into work toward "specific goals that define an overall context for research," such as population control. Embracing a form of Baconianism, Brown thus rejected Vannevar Bush's more general vision for science, a rejection I quoted near the beginning of this essay (see note 2). Like Havel's, his answer to the question whether science can share a place at the center of modern culture was clearly *No.*

When George Brown presented his ideas to an audience of scientists at the session he had organized and for which he had selected a panel of social scientists,[28] only one of the panel allowed himself to disagree openly, while another of the panelists urged

24
Reprinted in Vladislav, Jan, ed. *Vaclav Havel, or Living in the Truth.* London: Faber & Faber, 1987, 138–39. The passage was written in 1984.

25
On July 4, 1994 Havel used the occasion to repeat at length much of his previous argument, in the service of explaining the present "state of mind [that] is called postmodernism," and the "crisis" to which science has led mankind. The only new part of his speech (published as an OpEd, July 8, 1994, *New York Times*) is that our "lost integrity" might paradoxically be renewed by "a science that is new, postmodern," such as the "anthropic cosmological principle" and "the Gaia hypothesis." This was too much even for Nicholas Wade, who wrote a devastating attack on Havel's essay (WADE, N. "Method and Madness: A Fable for Fleas." *New York Times Magazine*, August 14, 1994, 18), ending with: "A view of the world built on the anthropic principle and the Gaia hypothesis would not be post-modern science but rather a throwback to the numerology and astrology from which the era of rationalism has still not fully rescued us. [...] To subvert rationalism into mysticism would be a cure more pernicious than the disease." The seduction of being counted among the postmoderns has apparently attracted even a handful of scientists; the chief example given is their postmodernist interest in "the limits of science." However, the lively discussion of that topic began in the 1870s, led by Emile Dubois-Reymond, and it also preoccupied the logical positivists. For other examples of this old problem, see HOLTON, G. and R. S. MORISON, eds. *Limits of Scientific Inquiry.* New York: W. W. Norton, 1978.

26
Published in September 1992 in the *American Journal of Physics.* Vol. 60, no. 9: 779–781.

27
Tape recording of the session (February 12, 1993) obtainable from the American Association for the Advancement of Science. George Brown's own "Opening Remarks" were also distributed as a Press Release by his Washington, DC office.

28
At the February 12, 1993, American Association for the Advancement of Science annual meeting.

Brown to go even further still: Perhaps not realizing how close he was coming to the "*völkische*" solution tried earlier elsewhere, including in Mao's Cultural Revolution, he seriously suggested that to screen proposals for scientific research funding the federal government form a variation of the National Science Foundation's Board whose membership should contain such non-experts as "a homeless person [and] a member of an urban gang." No one there dared to raise an audible objection. One felt as if one glimpsed the shape of a possible future. But it is also important to note that later on Mr. Brown, apparently moved by the intellectual objections, such as those given above, voiced to him by one or two scientists, distanced himself from Havel's position. Indeed, no one can fail to agree with him that in the post-Cold-War context, it is "a moral imperative to enlist science and technology in a campaign for a more productive and humane society in which all Americans can enjoy the benefits of an improved quality of life."[29]

In this brief overview, ranging from the trembling pillars of the Platonic tradition of the West to the so-called "End of the Modern Era" and the "End of Progress," we have identified some of the chief historic trends that have risen and fallen and risen again in the mixture from which the predominant view of an epoch emerges. Today's version of the Romantic Rebellion, while strong in other fields, represents still only a seductive minority view among analysts and science policy makers, coming not up from the grass roots but down from the treetops. However, while it is held among prominent persons who can indeed influence the direction of a cultural shift, the scientists at large, and especially the scientific establishment, have chosen to respond so far mostly with quiet acquiescence. If those trends should continue, and the self-designated postmodernists rise to controlling force, the new sensibility in the era to come will be very different indeed from the recently dominant one.

Experts in science policy are now debating what they call the on-going renegotiation of the "social contract" between science and society.[30] One can argue that such a change has been overdue for many reasons, one being that the relatively protected position given to science for many decades had less to do with society's commitment than with the Cold War and with the implicit over-promises regarding spin-offs, which, as Don K. Price warned long ago,[31] would eventually come back to haunt scientists. Adding concerns about the state of the economy, and competitiveness, the lack of general scientific literacy, etc., there is much in such a list to help explain the public's readiness for a reappraisal. But by my analysis, such factors act only as catalysts or facilitators of the tidal change that historically are always potentially present in our culture.

Of course, it may turn out that the recent version of the Romantic Rebellion will peter out—although I doubt it will. Or it may gain strength, as it did in the nineteenth-century and again at various times in the twentieth, especially when the scientific community itself paid little attention to the course of events. Or at best a new accommodation might gradually emerge, a "third way," based on a concept analogous to complementarity (and also analogous to the complementarity of personal and public science within the practice of research itself). That is, it may at last be more widely recognized, by intellectuals and the masses alike, that the scientific and humanistic aspects of our culture do not have to be opposing worldviews that must compete for exclusive dominance, but are in fact complementary aspects of our humanity that can and do coexist productively (as Samuel Taylor Coleridge put it memorably in chapter 14 of his *Biographia Literaria*: "in the balance or reconciliation of opposite or discordant qualities"). At any rate, historians will watch the next stages of the old struggle to define the place of science in our culture with undiminished fascination—although perhaps also with an uneasy recollection of Oswald Spengler's prophecy, of Sigmund Freud's pessimism, and of Isaiah Berlin's analysis of the trajectory of our modern era.

[29] Brown, George E. "New Ways of Looking at US Science and Technology." *Physics Today*, 1994. Vol. 47: 32. In a talk on "The Roles and Responsibilities of Science in Post-modern Culture" (February 20, 1994, at another annual meeting of the American Association for the Advancement of Science), Mr. Brown remarked: "Let me begin by suggesting that the term 'post-modern culture,' sometimes used to describe the present era, is a rubric that comes from the arts and architecture where it had identifiable meaning. To my mind, if the term post-modern is used as a definitional period for policy, politics, or for economic eras, it leads to confusion; and it will not help us to define a point of departure for our discussion here. I hope today's discourse does not get sidetracked on a tedious dissection of post-modernism. I should note, however, that the editorial that appeared in the *New York Times* two years ago entitled 'The End of the Modern Era' by Czech philosopher and playwright Václav Havel, contained several points to which I agree, and have included in previous talks. Although Havel comes to the terms modernism and postmodernism from his artistic and philosophical orientation, I do not subscribe to those labels, in large part because I do not fully understand his use of them." Similarly, Mr. Brown is one of the few policy makers who has protested Senator Barbara Mikulski's recent edict that federal funding for basic, "curiosity-driven" research be cut back in favor of supposedly quick-payoff "strategic research."

[30] See especially Brooks, Harvey. "Research Universities and the Social Contract for Science," in Lewis Branscomb, ed., *Empowering Technology: Implementing a US Strategy* (Cambridge, MA: MIT Press, 1993). Brooks has all along been one of the most prescient and observant authors on the place of science in our culture. See for example his essay, "Can Science Survive in the Modern Age?" *Science*, 1971. Vol. 174: 21–30.

[31] E.g., in Price, Don K. "Purists and Politicians." *Science*, Jan. 3, 1969. Vol. 163: 25–31.

the structure and evolution of the universe

FRANCISCO SÁNCHEZ MARTÍNEZ

Previous page:

Figure 1. In this image, two spiral galaxies—NGC2207 and IC2163—meet in their initial phase of interaction. Over time, the two galaxies will merge to form just one. The forces of NGC2207's gravitational tide have twisted the shape of IC2163, which expulses stars and spills gasses in long snake-like shapes one hundred thousand light-years long (on the right side of the image). That is how the large spiral galaxies were formed. Our own galaxy, the Milky Way, has absorbed smaller galaxies in the past and is currently swallowing up another. In the future, we will collide with M31, the largest spiral galaxy in our Local Group. NASA.

The Great Adventure

Ever since man became "sapiens," he has looked questioningly to the heavens with great interest and awe. And that is quite natural, for what happens over our heads has a great effect on us. Primitive man had a direct sense of the influence of the heavens on our lives, for he was in permanent contact with nature and totally dependant on the birth and death of the Sun, Moon, and stars that marked the rhythm of day and night, and of the seasons. He learned prediction by looking at the situation of the heavenly bodies, as he needed to know what was going to happen in order to find food, not to mention avoiding being eaten himself by other predators. Moreover, we can imagine the stupor and fear our ancestors must have felt in the face of the unforeseen and dramatic phenomena they could observe in the sky: lightening, thunder, the polar auroras, shooting stars, meteorites, comets, and solar or lunar eclipses. How could they not see them as signs of far superior beings? It is thus logical that they would consider the heavens to be where their gods dwelled. Some quickly realized that knowing the secrets of such phenomena and making others believe in their capacity to use them to help or harm each other would bring them great power and stature as divine mediators. That is why even the most archaic civilizations had celestial myths, rites, and omens, which were kept by their priesthoods. And even today, in the twenty-first century's most developed and technological societies, these primitivisms emerge in the form of astrology, astral sects, and other such trickery. And the most complex and speculative scientific theories and cosmological models are today defended by many of our world's must erudite people with a fanaticism that borders on the religious. All of this must be kept in mind when trying, as I now am, to offer a necessarily condensed and accessible overview of what we know today about the structure and evolution of the immense Universe to which we belong.

We must start by remembering and emphasizing something that will put the following observations in context: all scientific knowledge is provisional, and completely subject to revision. Moreover, what I am going to explain below are speculations based on rigorous science—but speculations all the same—that try to explain what we have been able to observe with the most advanced telescopes and instruments of our time. The real and complete reality of the vast Cosmos cannot be grasped by us, due to our own finitude. In sum, while we have learned a huge amount, there is a much larger amount that we still do not know. And, in writing this overview, I have been sorely tempted to accompany each statement with the cluster of unanswered questions that envelop it. But such detail must be left for specialized books.

That being said, it is even more thrilling to contemplate the beautiful and unfinished adventure of humans moving blindly but decisively forward in search of the infinite mysteries, driven by their curiosity and innate desire to learn. Such are the characteristics that have made us able to surpass our own strong limitations, reaching previously inconceivable heights. For example, in the concrete case of the sense of sight—we are, after all, practically blind, incapable of making out distant objects, and only capable of seeing an extremely limited band of wavelengths of the electromagnetic spectrum—we have been able to invent marvelous "prosthetic devices", that is, telescopes. So we can now "see" celestial objects that are at such enormous distances that we have to measure them in thousands of millions of light years.

As our capacity to "see" farther and in more detail has grown, our idea of the Cosmos has changed, and that change has been dramatic in recent times (fig. 2). If we look back only as far as the Renaissance, we find the "Copernican Revolution," which removed the Earth from the center of the Universe, reducing it to a mere satellite of the Sun and showing us that heavenly matter was of the same nature as us—no more divine than the Earth's own dust. For a time, we were thrilled by such unexpected and fantastic things, and entertained by the precise mechanics of the heavens, studying the movement of the bodies that make up our Solar System. As a result, we believed in the perfection of the cosmic clock and the unchanging nature of the Universe.

It was less than a century ago when this perfection and serenity came crashing down. First, science confirmed the existence of many other "island universes" located outside our Milky Way—until then, we considered it a unique cluster of stars and nebulae surrounded by infinite emptiness. Then we discovered that, the farther away they were, the faster they were separating from each other. So finally, we had to accept that the Universe was expanding, that it was growing larger, and colder. It was enough to posit this expansion in reverse in order to arrive at the original singular moment, and from there to the idea of the "Big Bang" that was the origin of everything. Alongside this, there was the discovery of the energy that makes the stars shine, which forced us to accept that the Universe is in permanent evolution. It is neither static nor eternal, and everything it contains is "alive," including stars and galaxies. Thus, we can observe the birth and death of celestial objects in all parts of the Universe, in perpetual processes of transformation and recycling.

And suddenly, we are surprised to discover that the Universe's expansion is *accelerating!* So once again, our cosmology is in upheaval. This global acceleration forces us to rebuild the edifice of cosmic physics from the ground up, conceiving of some sort of mysterious energy linked to what we call the vacuum, which fills, and pushes, everything. An energy that produces anti-gravitational forces capable of resisting the foreseeable implosion, and with such enormous strength that it can accelerate the expansion of space.

So now, dear reader, let us explore the narrow trail blazed by science in search of the structure and evolution of the Universe to which we belong.

The Universe is accelerating

One of the great practical problems involved in exploring the edges of the Universe is the precise determination of distances. Without getting too technical, we can say that we use "type 1a" supernovas (terminal stars, all of which have practically the same intrinsic brightness) as marker beacons, allowing us to determine the distance of extremely faraway galaxies. The concept is simple: if all such stars are equally bright at their point of origin, then a given supernova's degree of brightness when observed from the Earth will reveal its distance, and consequently, that of the galaxy to which it belongs.

Using this technique, research groups directed by Saul Perlmutter and Adam Riess were able to independently determine the distance of galaxies in which these type of supernova explosions occurred. When they compared the data they obtained with those galaxies' red shift—which measures the expansion of the Universe—they were surprised to discover that, the further away the galaxies, the slower the rate of expansion. In other words, in the past, the Universe was expanding more slowly than it is today. So the Universe is now governed by an accelerated expansion.

These observations by Perlmutter's and Riess' groups constitute the first observational data that the Universe's expansion rate has not been uniform throughout its very long history. That fact is enormously important, as we will see later on.

But before that, let us think for a moment about the idea of the Universe's expansion. This is one of the fundamental concepts of modern science, but it is still one of the most poorly understood. Most people imagine the Big Bang as some sort of enormous bomb that exploded at some point in space, pushing matter outwards because of pressure differences caused by that explosion. But for astrophysicists, the Big Bang was not an explosion "in space" but rather, an explosion "of space," and that is a truly important nuance. In this type of peculiar explosion, density, and pressure are maintained constant in space, although they decrease over time.

The visual metaphor for this explosion is generally a balloon. As it is inflated, any details printed on the surface grow farther apart, so that everything gets farther away from everything else. This two-dimensional idea is very graphic, but it has a problem: it can lead us to believe that, like a balloon, the Big Bang also had a center, a single point from which it expanded. In fact,

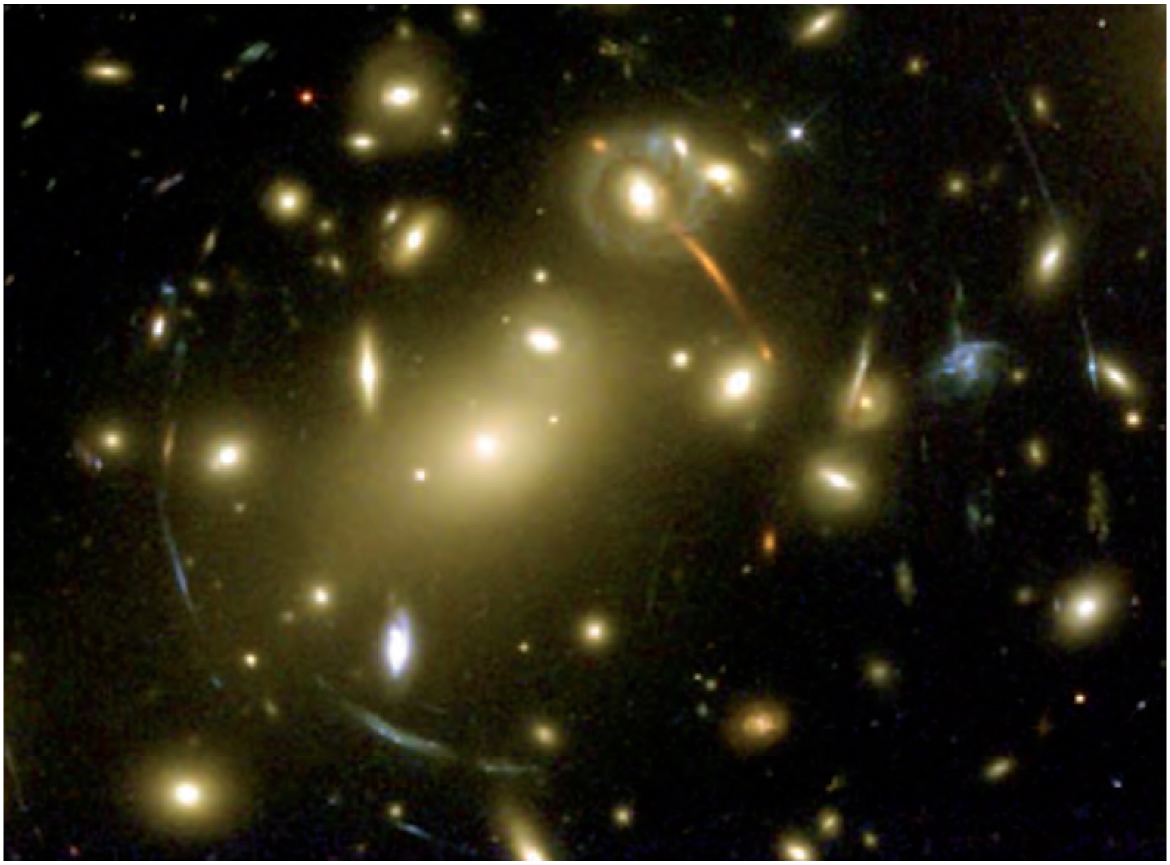


Figure 2. This is how we see the distant universe through a "gravitational lens." The galaxy cumulus, Abell 2218, is so massive that it is able to curve light rays, creating images the way a lens would. The arcs we can see are distorted images of galaxies much farther away than the cumulus. This amplifying effect makes it possible to penetrate even more deeply into the cosmos, seeing farther. NASA.

the Big Bang happened at all points in space at the same time, not in any specific one, which is in keeping with Einstein's General Theory of Relativity. According to the most accepted models, the Universe needs neither a center from which to expand, nor empty space *into which* it can expand. Those models do not even call for more than three dimensions, despite the fact that some theories, such as "string theory," call for a few more. Their theoretical base is Relativity, which establishes that space needs only three dimensions to expand, contract, and curve. Moreover, we should not imagine that the singular event at the origin of the Big Bang was something small, an "initial atom," as is sometimes said. Because this is generated no matter what size the Universe may have, be it finite or infinite.

Let us now recall that atoms emit and absorb light at very specific wavelengths, no matter whether they are in a laboratory or in a faraway galaxy. But in the latter case, what we see is a shift towards longer wavelengths ("red shift"). This is because, as space expands, electromagnetic waves stretch, becoming redder. This effect makes it possible to measure the speed with which galaxies are separating from each other, called recession speed. We should emphasize that cosmological red shift is not the normal Doppler effect that happens in space, and its formulae are also different.

Despite the overall expansion of space, there are galaxies, such as our neighbor Andromeda, that are drawing closer and seem not to obey the law of expansion. These are apparent exceptions caused by the fact that, near large accumulations of matter, gravitational energy becomes preponderate, leading those giant swarms of stars to turn around each other. Distant galaxies also present those local dynamic effects, but from our enormously distant perspective, they are overshadowed by their great recession speeds.

To make matters even more complex, the Universe is not only expanding, it is also doing so at an ever-faster rate. That situation has led scientists to recover the constant that Einstein introduced in his General Theory of Relativity to maintain the paradigm of a stationary Universe. When we thought we were living in a decelerating Universe, it was logical to think that, as time passed, we would be able to see more and more galaxies, but in an accelerating Universe, the opposite should be true. The cosmic horizon of events, determined by the finite velocity of light and space's increasing recession speed, marks a border beyond which the events that occur will never be seen by us, because the information they emit cannot reach us. As space's rate of expansion increases, we will gradually lose site of one galaxy after another, beginning with the most distant ones.

Dark energy

The fact that the Universe is accelerating has caught astronomers and physicists off guard and they are brimming with scientific speculations to explain it. The human imagination is capable of inventing many truly ingenious theories, but only those that can explain all of our observations will last. Astronomic observation continues to be the touchstone of any theory, and no matter how elegant it may be, it will have to be validated by observation. That is what makes scientists seem slow and conservative, moving extremely cautiously as they change well-established theories and paradigms.

One immediate form of interpreting accelerated expansion would be to consider that gravity does not follow the same laws in our nearby surroundings as on a super-galactic scale, and that such distances cannot brake expansion because gravity's power of attraction does not extend to an infinite distance. Another proposal that has already been formulated is that the acceleration observed is actually caused by time itself, which is gradually slowing down. But cosmologists prefer to maintain the universality of the physical laws established on planet Earth and its surroundings, and they have begun postulating the existence of a sort of cosmic fluid with contradictory properties that fills everything and appears in the form of an unknown energy—called "dark energy"—that repels, rather than attracts. Its power would be so great that it very efficiently overcomes the gravitational attraction of the enormous masses of galactic cumuli and supercumuli.

For the time being, available observations seem to favor this dark energy. As we mentioned above, the first observations were those made using photometry of type 1a supernovas, which showed that the oldest galaxies are expanding at a slower rate than at present. But measurements of radiation from the Cosmic Microwave Background (or "background radiation") point to the same conclusion.

Discovered in 1965 by Penzias and Wilson, this radiation is a background noise that fills everything and its discovery served to reinforce Big Bang models. Much later, it became possible to detect anisotropies (directional dependence) in the Cosmic Microwave Background, and even though they are extremely small—around 0.00001%—they are full of information about the origins of the structure of the gigantic Cosmos that we see. So now, the contribution of dark energy seems necessary to complete the density of the Universe as measured by the Cosmic Microwave Background. Since the sizes of irregularities in background radiation are a reflection of the global geometry of space, they serve to quantify the density of the Universe, and that density is considerably greater than the simple sum of ordinary and exotic matter. Moreover, the modifications that the gravitational fields of large cosmic structures cause in this radiation depend on how the rate of expansion has changed. And that rate agrees with the predictions made by dark energy models.

The distribution of galaxy swarms follows certain patterns that are in keeping with "stains" observed in background radiation and those stains can be used to estimate the Universe's total mass (fig. 3). It turns out that those models, too, require dark energy. And studies of the distribution of gravitational lenses (remember that very massive objects behave as lenses, curving the trajectories of light) seem to need dark energy to explain the growth over time of agglomerations of matter. But not everything confirms its existence. There are observations that cannot be explained by these models, including the abundance of the most distant galaxy cumuli.

Many researchers are trying to posit dark energy as the cause of aspects unexplained by previous models. Data is accumulating, and models are ever more refined, and more indications, and even proof, will undoubtedly

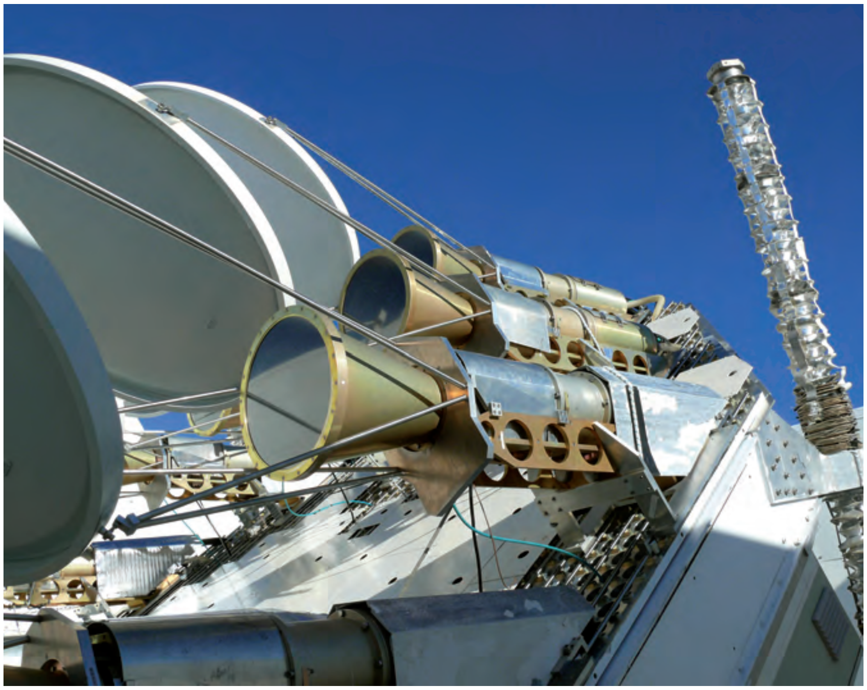


Figure 3. The Very Small Array, which belongs to the IAC and the Universities of Cambridge and Manchester, is one of the instruments installed at the Canary Islands' Astrophysical Institute's Teide Observatory for measuring anisotropies in the Cosmic Microwave Background. The observatory has been systematically measuring this primeval radiation with different instruments and techniques for over twenty years. IAC.

be added, given the feverish activity this surprising acceleration of the Universe has caused among scientists.

The omnipresence of dark energy is so subtle that, even though it fills everything, it has gone unnoticed until now. It is very diluted, and does not accumulate in "lumps" as matter does. In order for its effects to be noticeable, very large amounts of space and time are necessary, even though it is the most powerful energy in the Cosmos. Let me add that, in this energy, which acts as a repulsive force and thus has negative pressure, there are two possibilities: the so-called "phantom energy" and what has been dubbed as the "quintessence." All of this is very evocative, but difficult to digest from a scientific standpoint, as these are forces we do not understand and cannot observe.

It is totally logical to think that if such energy represents more than three quarters of our Universe, it must have had an enormous influence on the latter's entire evolution, determining its large-scale structure and the formation of galaxy cumuli. The very evolution of galaxies themselves must be marked by its omnipresence. We know that the formation of galaxies and their grouping in cumuli is determined by their own interactions, collisions, and merging—our own Milky Way is thought to be the result of the coalescence of perhaps a million dwarf galaxies—so dark energy must have played a significant role in all of this. Nevertheless, clear confirmation will come when we are able to determine whether the beginning of the predominance of accelerated expansion coincides in time with the end of the formation of large galaxies and supercumuli.

The Universe in four dimensions

I have thought a lot about how to illustrate what we know today about our Universe's structure. It is not at all easy for many reasons, and not only because of the difficulty of simplifying things for non-specialists without leaving any loose ends that we take for granted.

If we consider the Universe to be everything that exists, from the smallest to the most gigantic entities, one way of showing their structure would be to make an inventory of all such elements and order them hierarchically in space. But this would be incomplete unless we also listed their interconnections and interrelations. Moreover, none of this—neither the elements nor their interconnections—is static, all of it is interacting and changing on a permanent basis. We must realize that, as such, we cannot have a "snapshot" of what is in the Universe at the present time, because when we look in one direction with a telescope, the deeper our gaze looks, the farther back in time we go. Thus, we are looking at a wedge of the Universe's history, rather than a snapshot. Nevertheless, inasmuch as all directions in the Universe are statistically identical, what we see in any direction at a distance of thousands of millions of light-years must be a representation of how our own, or any other, region of space was, thousands of millions of years ago.

Let us take it a step at a time. First we should remember that, in keeping with what we have already said, more than three quarters of our Cosmos is now a form of that mysterious entity we call dark energy, and more than 85% of the rest is what is called "dark matter," which we cannot see because, though it interacts with gravity, it does not interact with radiation. In other words, not much more than three percent of the entire Universe is "ordinary matter." And we only manage to see a tiny part of the latter, concentrated in stars and galaxies. What we call ordinary matter is actually the baryonic matter—protons, neutrons, and so on—of which we ourselves are made. Most of such matter takes the form of ionized gas plasma, while only a tiny part of it is in solid or liquid state. How difficult it is to grasp that the immense oceans and solid ground of the Earth's surface, on which we so confidently tread, are incredibly rare in our Universe! But science has taught us to accept that we live in a very exotic place in an everyday part of the Cosmos.

On the other hand, the panorama could not be any more disheartening: despite our elegant scientific speculation, we do not have the slightest idea about the nature of 97% of what constitutes our Universe! Of course, just knowing that is already a great triumph for the grand human adventure in search of knowledge.

Our own nature leads us to move and understand things in three spatial dimensions plus time. And this space-time is the context in which most relativist models are developed. That is why I am going to describe the structure of the Universe in four dimensions. But first, I must at least mention models of "multiverses" derived from superstring theory. These are elegant physical-mathematical speculations about multiple universes in which our three-dimensional Universe would be just one projection of three dimensions installed in a global space of nine.

Below, I will try to offer an accessible description of how astronomers currently imagine the Universe to be at the present time in its history. Afterwards, I will focus on some of the most significant stages of its evolution.

On a large scale, the Universe we can now contemplate with our telescopes appears to be a little more that 13,000 million years old, and enormously empty. Matter appears to be very concentrated and hierarchically organized around the gravitational fields of the stars, with their planetary systems, of galaxies, galactic cumuli, and supercumuli (fig. 4). The enormous planetary, interstellar, and intergalactic voids are filled with very diluted matter, which actually adds up to the greater part of ordinary matter. Dark matter also accumulates, and is ordered in analogous fashion, for it, too, is ruled by gravity. Dark

Previous page:

Figure 4. The galaxy known as Triangle or M33. This extraordinary image obtained with the INT at the Canary Islands' Astrophysical Institute shows the smallest of the three spiral galaxies that make up the Local Group, which also includes Andromeda, the Milky Way and a few other smaller galaxies. Also visible are innumerable other components, from various sorts of interstellar clouds to stars in every phase of life, beautiful planetary nebulae and the remains of supernovas. IAC.

energy, however, does quite the opposite: it is uniformly spread throughout the Universe.

Were we to zoom in, drawing close to each part of our own galaxy—the Milky Way—we would find brilliant planetary systems with one or more suns, with their planets, satellites, comets, and myriad smaller objects orbiting around each other, and all around their respective centers of mass. And as there can be hundreds of thousands of millions of them in any galaxy, some will be new, emerging from clouds of interstellar dust and gas, among convulsions and readjustments, while others are in their final stages, imploding and exploding, expulsing incandescent matter, particles, and radiation in a very beautiful but nightmarish spectacle. Most galactic objects, of course, will be in the intermediate stages of their lives. A description of the lives and miracles of such multifaceted and variegated galactic "fauna" would exceed the scope of the present article, though it would be full of color, beauty, and drama. Another question that is very much more important to us here is that of the existence of life outside our planet. If there were life in different parts of the Universe, it would have to have an influence—though we do not yet know what that might be—on its structure. This might well lead to readjustments of our concept of the Cosmos even greater than those we now have to make as a consequence of the Universe's acceleration.

The enormous swarms of stars, gas, dust, and much dark material that make up the galaxies are not isolated in space. On the contrary, we can see that they are strongly linked to each other by gravity. And those links lead to groups, called galactic cumuli, which are, in turn, linked to form galactic supercumuli. Such enormous accumulations of material appear to be organized in meshes similar to spider webs, favoring filament-like directions of tens of millions of light-years. And all of it floats in enormous voids.

We must not forget that all this enormity is fully active and that all the celestial objects are moving at incredible speeds. Thus, to imagine a Universe that is mechanically regulated like some sort of perfect clock is the farthest thing from reality. Interactions are multiple, and collisions frequent. Those collisions—of external layers of stars with their surroundings, or interstellar clouds and super-clouds, or even between galaxies—turn out to be the most efficient mechanisms for fine-tuning the galaxies and mobilizing cosmic surroundings (fig.1). Energy seems to be unleashed in incredibly violent phenomena that we can observe all over the Universe, producing new celestial objects.

And every bit of this superstructure is also steeped in dark energy, which efficiently counteracts gravity, expanding space at an accelerating rate and undoubtedly generating direct or indirect activity at all levels of the cosmic structure. The fact that we do not yet know about it does not mean that it is not occurring.

We can retain this simplified image of a gigantic, violent Universe in accelerated expansion, with its matter—the ordinary matter of which we ourselves are made, and the dark matter—concentrated in islands full of action, pushed by gravity, uniformly steeped in dark energy, and bathed in electromagnetic radiation. And in one tiny spot, our miniscule Earth, filled with life, dancing in space.

Following this semi-cinematic portrayal of how we understand what must be the current structure of the Universe, we must say a little about the main stages of its life. Because what we see today, including the life that thrives on planet Earth, is a consequence of its general evolution, which is determined by laws we are trying to discover. In fact, the quest for knowledge about the birth and evolution of each and every part of the Cosmos is what presently underlies all astronomical research.

The evolution of the Universe

Much has been written about time's arrow, trying to discover where the evolution of our Universe is headed and, since it had a beginning, finding out what its end will be. Let us see what can be said about all this in an intelligible way, and with both feet on the ground.

Ever since the Universe stopped seeming immutable at the beginning of the past century, we have sought knowledge of its history and, most of all, its evolution. For that is the key to the origin of our own history, and of intelligent life on other planets in other star systems. But history is the story of events in time, and it seems that time, our time, began with the very Universe to which we belong. And we do not yet know with any certainty what the real essence of this physical parameter might be. Without entering into profound disquisitions, we can consider time to be the way we intuitively imagine it: a uniform continuity reaching from the Big Bang towards a distant future.

Almost all the information we received from outer space comes in the form of electromagnetic radiation, and the first retrospective snapshot of the Universe comes from the Cosmic Microwave Background. By then, the Universe was about 400,000 years old, and many things of enormous importance had already happened. We can infer what those things were using our cosmogonic models, the most accepted of which is known as the Standard Model. We must not forget that this model describes what happened after the Big Bang, but it does not offer information about that event, itself. We should also remember that the model was developed before the discovery of accelerated expansion, and the three pillars on which it stands are: decelerating expansion, the Cosmic Microwave Background, and primordial nucleogenesis, which produced the first light elements that continue to dominate matter. The key to this model is that, at the beginning, the Universe was very hot and very dense, becoming cooler and less dense as it expands.

Abundant new and highly speculative models continue to appear in order to explain the nature and birth of matter and its interactions, and they are so complex that they are only understood by those who work in that field. Their empirical underpinnings, however, are astronomic observations and experiments with large particle accelerators, all of which are still not sufficient to shed light in the midst of so much physical-mathematical speculation. I say all this in order to avoid the misconception that most of the disconcerting things being said about the Universe's first moments—including what I am going to say—are scientific fact.

Immediately following the Big Bang that started everything, in just the first 10^{35} seconds, when all the fundamental forces were still unified, space underwent a prodigious exponential expansion. It grew by a factor of 10^{26} in just 10^{33} seconds. That is what inflationary models suggest, and they rely on data from background radiation. That accelerated expansion rarified everything that preexisted, smoothing out possible variations in its density. This, then, is the first accelerated expansion, implying something as inconceivable as the idea that energy must be positive and remain almost constant—the "almost" is very important, here—while pressure is negative (fig. 5). This ends with a sudden drop in density. Obviously, we do not know how, nor why this inflation started and stopped.

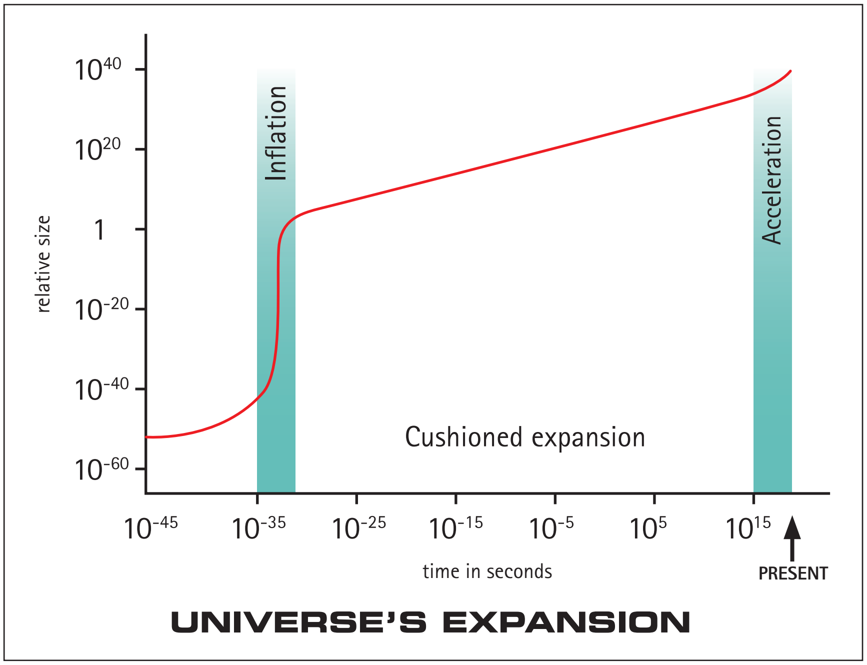


Figure 5. While it is not possible to realistically represent the expansion of space, much less its extraordinary "inflation," we can get a preliminary, though greatly simplified, idea of its expansion over time using the graph shown above. It uses logarithmic scales to make details visible. Both "inflation" and "accelerated expansion" share a negative pressure that opposes gravitational attraction. We have yet to discover the nature of these phenomena. IAC.

During the inflationary period, the density of space fluctuated minimally, due to the statistical nature of quantum laws that hold at subatomic levels. But those irregularities were exponentially expanded by inflation, leading to the anisotropies in the Cosmic Microwave Background. These are the seeds that mark the grandiose destiny of the Universe; they are the embryos of the macrostructures of galaxies and galactic cumuli we see today. The Universe emerged from this period in a heated state, with the potential energy of the void converted into hot particles.

To continue with this succinct description of the Universe's evolution based on the most accepted models; the different particles and antiparticles, along with their interactions, created themselves, as this was permitted by the Universe's ongoing expansion and cooling. Just 10^{5} seconds after the Big Bang, baryons already existed. This "soup" of particles in continuous birth and death continued to cool and almost all particles of matter and antimatter annihilated each other. But for unknown reasons, there was a slight excess of baryons that did not find particles of antimatter against which to annihilate themselves, and thus they survived extinction.

When the temperature fell to around 3,000 degrees, protons and electrons became able to combine and form electrically neutral hydrogen atoms. Matter thus stopped being linked to radiation, as photons stopped interacting with matter in such an intense way, and light spread all over. Those very first photons are what make up the radiation of the Microwave Background. By then, the Universe was about 400,000 years old and, as we have seen, some very important things had happened. One of these was the "primordial nucleosynthesis" that determined the absolute preponderance of hydrogen and helium in the Universe. That process, governed by expansion, must have happened in only a few minutes, which is why nucleosynthesis only generated the lightest elements.

This was followed by a grey period stretching from the freeing of the radiation that makes up the Cosmic Microwave Background to the re-emergence of light as the first galaxies and stars were born. We know very little about that period of the Universe because there are no radiations to be observed. Still, it was decisive, because that is when gravity began to assemble the objects that now inhabit the Cosmos. It ended in the first millions of years, when starlight became strong enough that its ultraviolet radiation could ionize the gas that now dominates intergalactic space. The stars that made that light were very peculiar—super-massive stars of one hundred solar masses or more, made up exclusively of hydrogen and helium. All of this is corroborated by observations of the spectra of the most remote quasars, galaxies, and explosions of gamma rays, as well as the discovery of distant galaxies less than one thousand million years after the Big Bang.

It is believed that galaxies and stars begin to form when a region of space with greater density than its surroundings begins to contract upon itself, as a result of its own gravity. Let us not forget that galaxies are made mostly of dark material, which cannot be directly observed. Even though such a region is subject to overall expansion, the excess of matter leads it to contract, creating a linked object, which may be a star, a stellar cumulus, or a galaxy. Of course many additions and nuances would have to be brought in to explain what we know about all this, which is quite a bit. And there are innumerable research projects underway to study the origin and evolution of stars and galaxies. Here, I can only indicate the basic mechanism that generated most of the objects we observe.

By all indications, during the Universe's first few thousand million years, there were frequent collisions among galaxies, gigantic outbreaks of star making inside them, and the generation of black holes of more that a thousand million solar masses. This, then, was an extraordinarily energetic and agitated period. That disorderly activity seems to be declining now, perhaps as a result of the accelerated expansion. In the nearest parts of the Universe, we only find such prodigious activity in the smaller galaxies. The larger ones, such as the Milky Way or Andromeda, are calmer and seem to have entered a mature stage.

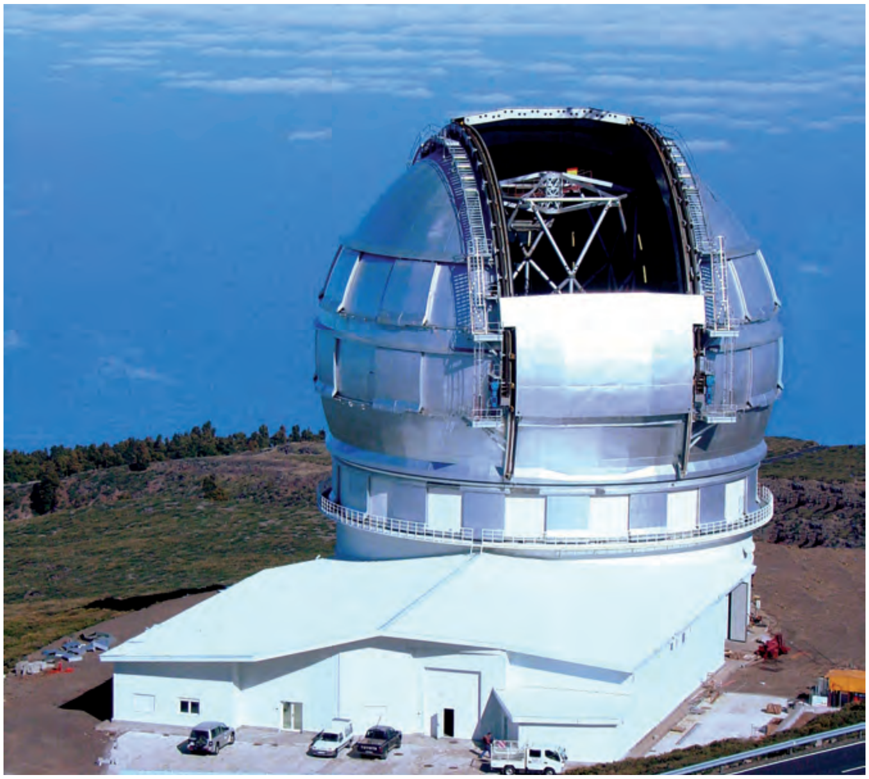


Figure 6. An outside view of the Gran Telescopio CANARIAS (GTC) at the Roque de los Muchachos Observatory of the Canary Islands' Astrophysical Institute. This is the largest and most advanced optical-infrared telescope in the world, with a segmented primary mirror of 10.4 meters in diameter. The low incidence of clouds—the ones visible in this image are actually below the telescope, although the photo's perspective does not make this clear—along with the transparence and stability of its atmosphere make this one of the extremely rare places on Earth where such an advanced telescope can be profitably installed. A large lens, excellent optics and powerful focal instrumentation, as well as the extraordinary astronomic quality of the sky over La Palma, make the GTC an extremely powerful tool for penetrating the secrets of the universe. IAC.

We cannot yet say when acceleration began to predominate over deceleration, although it has been pointed out that this must have happened when the Universe was around eight thousand million years old. Earlier, we mentioned the degree to which scientists speculate about the effects of dark energy. Undoubtedly, once the reality of the two competing energies is confirmed—gravity, which tries to bring material together; and dark energy, which tries to separate it—there will have to be new models to explain all our observations. But we will have to wait until the new telescopes on Earth and in space begin producing significant data before we can consider them trustworthy. In that sense, we have high hopes for the Gran Telescopio CANARIAS, which will be fully operational in 2009 (fig. 6). Telescopes are the only time machines, and the larger their mirrors, the more deeply they can look into the Cosmos, and the further back in time we can see. With them, we seek to observe the birth and evolution of the earliest objects.

Because of their importance to us, we should add that our Sun and its Solar System were born when the Universe must have been around nine thousand million years old. They stem from a cloud of recycled matter produced inside previous stars, and cast out into the interstellar void. The chemical elements of organic molecules that sustain all forms of life on our planet, could not have been created *in situ,* and must have already been in the protoplanetary disc, which allows us to say, with certainty, that "we are stardust." And this is much more than a pretty bit of poetry.

Until a very short time ago, the future of the Universe was predicted on the basis of gravitational energy and thermodynamics, as a function of the quantity of material it contained. If its mass was greater than the critical value calculated in keeping with common models, it was predicted that its expansion would grow slower until it imploded, as part of an oscillating process of Big Bangs and their posterior implosions. If that were not the case, we would continue to expand indefinitely. Now, we have to take into account the Universe's acceleration, and that leads us to imagine a different future.

The End

If accelerated expansion continues, the galaxies will begin disappearing from view as they cross the horizon of events, beginning with the most distant ones. In a few thousand million years, the nearby galaxies will have fused, forming a gigantic group of stars linked by gravity, a mega-super galaxy or "mesuga," enveloped in a dark, empty space. Radiation from the Microwave Background will be so diluted that it will be undetectable. Long before everything grows cold and ends, we will be isolated in space and our accessible Universe will be only our own "mesuga".

A bleak ending for our physical world. But is it really the end? What will the internal evolution of the myriad millions of "mesugas" be when they become isolated? Will there be a mechanism that connects them in some way, even though the separation between them continues to grow at an accelerating rate?

Clearly, we never lose hope, nor do we lose our will to be eternal! For that is how we are. Moreover, we are designed to be curious. We have a will to know, and in Sapiens-Sapiens, that drive seems as basic as the one to reproduce. Could this have something to do with the expansion of the Universe?

As things stand today, the joint existence of two opposite energies—gravity and dark energy—seems to have been necessary for the formation of the Universe, and of our Sun and Earth within it, so that, following a laborious process of evolution, our own parents could begat each of us. If dark energy had been just a little weaker, or a little more powerful, you would not be reading this book now, nor would I have been here to write it. Whether we descend to the minimum level of the most recently discovered sub-nuclear particles, or lose ourselves in the immensity of the Cosmos, we find everything in constant activity, moved by powerful forces. And the same can be said of life in all its facets, be they unicellular organisms or impenetrable forests—there, too, activity is incessant. This must be something consubstantial with our Universe: nothing is static, everything is action and evolution. Energy is abundant and seems to be wasted: cataclysmic episodes of immense power are frequent in galaxies and stars. This constant and often catastrophic agitation is the manifestation of what we could call "Cosmic impetus," which fills and pushes everything. That impetus is expressed as energies that we could try to systemize by cataloguing them in two large groups: "physical energies" and "life energies." At some point, it will be necessary to conceptualize all of this in mathematical form.

Anyone who has enjoyed the starry nights of the wilderness, far from the polluted hustle and bustle of our cities, and has let himself be carried away by his sensations, feelings and imagination in this state of grace, will have felt the profound force of the fundamental questions. Even if none of them has been answered in the previous pages, the careful reader will have sensed the brilliant spark of human intelligence successfully confronting the immensity of the Universe and its attractive mysteries. The truth is: even though our ignorance borders on the infinite, what we already know is considerable, and constitutes a notable triumph that dignifies all of humanity. While this incredible, evil, and stubborn, yet loving and intelligent species continues to exist, it will continue to look questioningly at the heavens.

Bibliography

Adams, Fred C. and Greg Laughlin. *The five ages of the universe: inside the physics of eternity.* Free Press, 2000.

Caldwell, Robert R. "Dark energy." *Physics World,* l.17/5, 2004, 37–42.

Caldwell, Robert R., M. Hamionkowski, and N. N. Weinberg. "The phantom energy and cosmic doomsday." *Physical Review Letters*, 91/071301, 2003.

Dine, M. and A. Kusenko. "Origin of the matter antimatter asymmetry." *Reviews of Modern Physics*, 76, 2004, 1–30.

Freedman, W. L. and Michael Turner. "Measuring and understanding the universe." *Review of Modern Physics,* 75, 2003, 1433–1447.

Harrison, Edward R. *Cosmology: the science of the universe.* Cambridge University Press, 2000.

Hertzberg, Mark, R. Tegmark, Max Shamit Kachru, Jessie Shelton, and Onur Ozcan. "Searching for inflation in simple string theory models: an astrophysical perspective." *Physical Review*, 76/103521, 2007, 37–45.

Kirshner, Robert P. *The extravagant universe: exploding stars, dark energy, and accelerating cosmos.* Princeton University Press, 2004.

Krauss, Lawrence and Glenn Starkman. "Life, the universe and nothing: life and death in an ever-expanding universe." *Astrophysical Journal,* 531/22, 2000, 22–30.

Peebles, P. J. E. and B. Ratra. "The cosmological constant and dark energy." *Review of Modern Physics,* 75, 2003, 559–606.

Primack, Joel and Nancy Ellen Abrams. *The View from the Centre of the Universe: Discovering Our Extraordinary Place in the Cosmos.* Riverhead, 2006.

Silk, Joseph. *The infinite cosmos: questions from the frontiers of cosmology.* Oxford University Press, 2006.

Srianand, T. A., P. Petitjean, and C. Ledoux. "The cosmic microwave background radiation temperature at a redshift of 2.34." *Nature*, 408/6815, 2000, 93–935.

Wilson, Gillian et al. "Star formation history since z = 1 as inferred from rest-frame ultraviolet luminosity density evolution." *Astronomical Journal*, 124, 2002, 1258–1265.

the world after the revolution: physics in the second half of the twentieth century

JOSÉ MANUEL SÁNCHEZ RON

The great revolutions of the twentieth century

During the first half of the twentieth century—actually, the first quarter—there were two major scientific revolutions. Those cognitive cataclysms took place in physics, and are known as the relativist and quantum revolutions. They are respectively related to the special and general theories of relativity (Einstein 1905a, 1915), and quantum mechanics (Heisenberg 1925, Schrödinger 1926).

Relativity

Much has been written, and will be written in the future, about the importance of those theories and their effect on physics as a whole, even before the middle of the century. Created to resolve the increasingly evident "lack of understanding" between Newtonian mechanics and the electrodynamics of James Clerk Maxwell (1831-1879), the special theory of relativity imposed radical modifications of ideas and definitions that had been in force ever since Isaac Newton (1642-1727) included them in the majestic structure contained in his *Philosophiae Naturalis Principia Mathematica* (1687)—concepts as basic from a physical, ontological and epistemological viewpoint as space, time and matter (mass). The result, in which measurements of space and time depend on the state of movement of the observer, and mass, *m,* is equivalent to energy, *E* (the famous expression $E= m\cdot c^2$, where *c* represents the speed of light), opened new doors for understanding the physical world. For example, this theory helped explain how it was possible that radioactive elements (uranium, polonium, radium, thorium) that had been studied for the first time by Henri Becquerel (1852-1908) and Marie (1867-1934) and Pierre Curie (1859-1906), emit radiation in a continuous manner with no apparent loss of mass.

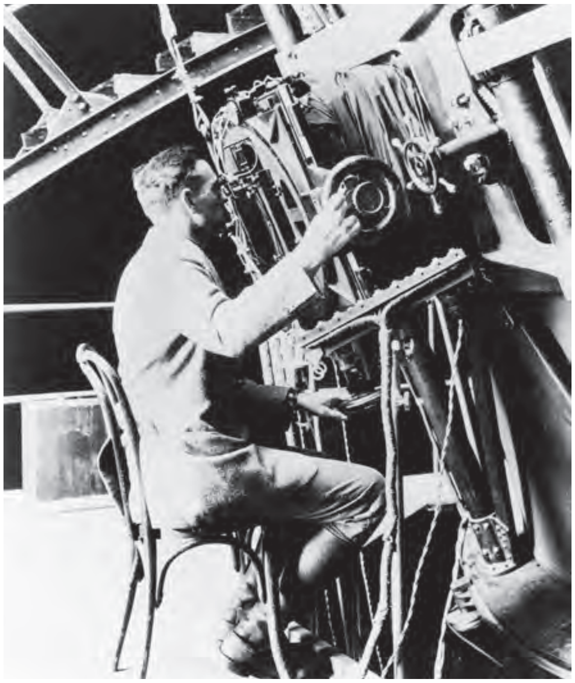
And then there was the general theory of relativity, which explained gravity by converting space—actually, four-dimensional space-time—into something curved, and with variable geometry! It was immediately apparent that, compared to Newton's universal gravitation, Einstein's new theory made it much easier to understand perceptible phenomena in the solar system (it solved, for example, a century-old anomaly in the movement of Mercury's perihelion). As if that were not enough, Einstein himself (1917) had the intellectual daring to apply his general theory of relativity to the overall Universe, thus creating cosmology as an authentically scientific and predictive field. While it is true that the model Einstein proposed at that time, in which the Universe is static, did not survive in the end; what matters is that it opened the doors to a scientific approach to the Universe, which makes it an almost unprecedented event in the history of science.[1]

[1] In order to construct a model of a static universe, Einstein had to modify the basic equations of general relativity, adding an additional term that included a "cosmological constant."

To find the exact solution to the equations of relativistic cosmology he was using, Einstein (1879-1955) employed physical considerations. Other mathematicians or physicists with special sensibilities and mathematical skills followed a different path, quickly finding new exact solutions—which implicitly represented other models of the universe—based exclusively on mathematical techniques, which they used to address the complexities of the equations of relativistic cosmology (a system of ten non-linear equations in partial derivatives). Alexander Friedmann (1888-1925), Howard Robertson (1903-1961) and Arthur Walker (b. 1909) found solutions implying that the Universe was expanding. In fact, another scientist obtained similar results: the Belgian Catholic priest, Georges Lemaître (1894-1966). This, however, should be mentioned separately because, as Einstein had done with his static model, Lemaître (1927) used physical considerations to defend his idea of a possible, real expansion of the Universe.

All of these models arose from solutions of cosmological equations; that is, they addressed theoretical possibilities. The question of how the Universe really is—static? expanding?—had yet to be elucidated, and for that, the only acceptable proof had to come from observation.

The lasting glory of having found experimental evidence indicating that the Universe is expanding belongs to the United States astrophysicist Edwin Hubble (1889-1953), who took advantage of the magnificent 2.5 meter-diameter reflector telescope at the Mount Wilson (California) observatory where he worked, along with excellent indicators of distance. Those indicators were cepheids, stars of variable luminosity in which it is possible to verify a linear relation between their intrinsic luminosity and the period of how that luminosity varies (Hubble 1929; Hubble and Humason 1931). And if, as Hubble maintained, the Universe is expanding, that would mean that there must have been a moment in the past (initially estimated as around ten thousand million years ago, later, fifteen thousand million, and now around thirteen thousand seven hundred million years) when all matter would have been concentrated in a small area: Lemaître's "primitive atom" or, the Big Bang, which turned out to be a very successful name.



Edwin Hubble taking photographs with the 2.5 meter Mount Wilson telescope (1924).

This was the birth of a conception of the Universe that is now a part of our most basic culture. But that has not always been the case. In fact, in 1948, as the first half of the twentieth century neared its end, three physicists and cosmologists working in Cambridge —Fred Hoyle (1915-2001), on one hand, and Hermann Bondi (1919-2005) and Thomas Gold (1920-2004) on the other (all three had discussed these ideas before publishing their respective articles)—published a different model of an expanding Universe: the steady-state cosmology, which held that the Universe has always had, and will always have, the same form, including the density of matter. This last aspect forced them to introduce the idea of the creation of matter, so that a "volume" of the universe would always have the same contents, even though it was expanding. According to them, the Universe had no beginning and would never end.[2]

Despite what we may think of it today —we are now fully imbued with the Big Bang paradigm—, steady-state cosmology was highly influential during the nineteen fifties. As we will see, it was not until the second half of the century that it was finally rejected (except in the minds of a few true believers, led by Hoyle himself).

Quantum Physics

The second major revolution mentioned above is quantum physics. While not rigorously exact, there are more than enough arguments to consider that this revolution's starting point was in 1900. While studying the distribution of energy in black-body radiation, the German physicist, Max Planck (1858-1947), introduced the equation, $E=h\cdot\upsilon$, where E is, as in the relativistic equation, energy, h is a universal constant (later called "Planck's constant") and υ is the frequency of the radiation involved (Planck 1900). Initially, he resisted this result's implication that electromagnetic radiation

[2] Hoyle (1948), Bondi and Gold (1948).

Werner Heisenberg in Goettingen (around 1924).

(that is, light, which was still considered a *continuous* wave at that time) could somehow also consist of "corpuscles" (later called "photons") of energy $h{\cdot}\upsilon$. But that implication of a "wave-corpuscle duality" eventually held sway, and Einstein (1905b) was decisive in its acceptance.

For a quarter century, physicists struggled to bring sense to those quantum phenomena, which eventually included radioactivity, spectroscopy and atomic physics as well. Here, it is impossible to offer so much as a list of the number of scientists involved, the ideas they handled and the concepts they introduced, let alone their observations and experiments. I can only say that a decisive moment in the history of quantum physics arrived in 1925, when a young German physicist named Werner Heisenberg (1901-1976) developed the first coherent formulation of quantum mechanics: *matrix quantum mechanics.* Soon thereafter, the Austrian Erwin Schrödinger (1887-1961) discovered a new version (they soon proved identical): *wave quantum mechanics.*

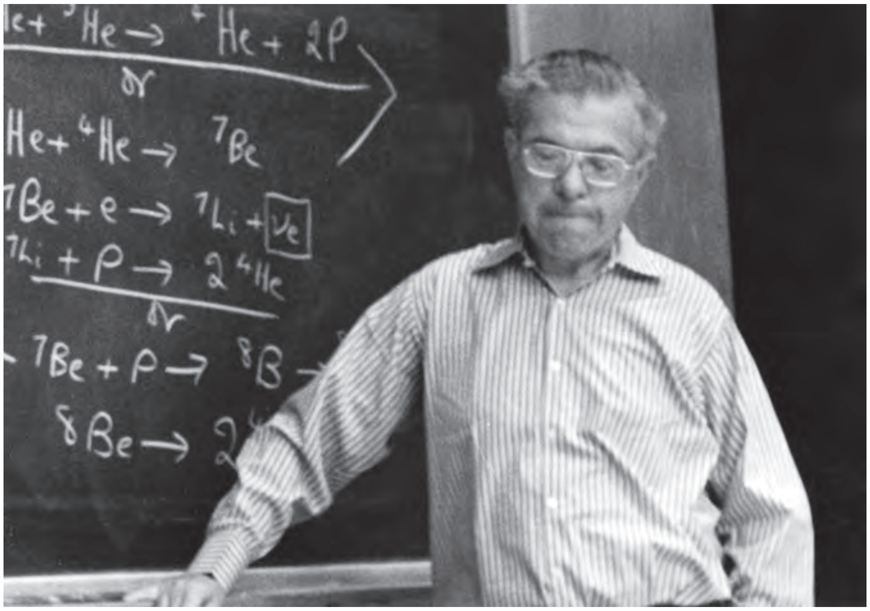
The stipulation by one of the two axioms of the special theory of relativity that the speed of light had to be constant, the dependence of space and time measurements on the movement of the observer, and the dynamical curvature of space-time, were already innovative and surprising findings, contradictory to "common sense." But the contents or deductions of quantum mechanics were even more shocking, including two that must be mentioned here: 1) Max Born's (1882-1970) interpretation of the wave function set out in Schrödinger's equation, according to which that function—the basic element used by quantum physics to describe the phenomenon under consideration—represents the probability of a concrete result (Born 1926); and 2) the principle of uncertainty (Heisenberg 1927), which maintains that canonically conjugated magnitudes (such as position and velocity, or energy and time) can only be determined simultaneously with a characteristic indeterminacy (Planck's constant): $\Delta x{\cdot}\Delta p \geq h$, where x represents position and p the linear momentum (the product of mass multiplied by velocity). At the end of his article, Heisenberg drew a conclusion from his results that has had lasting philosophical implications: "In the strong formulation of the causal law *'If we know the present with exactitude, we can predict the future,'* it is not the conclusion, but rather the premise that is false. *We cannot know,* as a matter of principle, the present in all its details." And he added: "In view of the intimate relation between the statistical character of quantum theory and the imprecision of all perception, it is possible to suggest that behind the statistical universe of perception there is a hidden "real" world ruled by causality. Such speculations seem to us—and we must emphasize this point—useless and meaningless. For physics must limit itself to the formal description of relations among perceptions."

Heisenberg and Schrödinger's quantum physics opened up a new world, both scientifically and technologically, but that was only the first step. There were still many challenges to be met, including making it compatible with the requirements of the special theory of relativity, and building an electromagnetic theory, an electrodynamics that would include quantum requirements. Einstein had shown, and later quantum physics agreed, that light, an electromagnetic wave, was quantized, that is, that it was simultaneously a wave and a "current" of photons. But the electrodynamics constructed by Maxwell in the nineteenth century described light exclusively as a wave, with no relation to Planck's constant. So, it was clear that something was wrong: the electromagnetic field also had to be quantized.

It was not necessary, though, to wait until the second half of the twentieth century for quantum electrodynamics. That theory, which describes the interaction of charged particles though their interaction with photons, took shape in the nineteen forties. It was independently developed and proposed by Japanese physicist Sin-itiro Tomonaga (1906-1979), and the Americans Julian Schwinger (1918-1984) and Richard Feynman (1918-1988).[3]

Quantum electrodynamics was a considerable theoretical advance, but it was nowhere near the culmination of quantum physics. At most, it was one more step up a ladder whose end was still far away. First of all, by the time the Tomonaga-Schwinger-Feynman theory came out, it was already clear that, besides the traditional forces of electromagnetism and gravity, there were two more: weak force, responsible for the existence of radioactivity; and strong force, which holds together the components (protons and neutrons) of atomic nuclei.[4] Therefore, it was not enough to have a quantum theory of electromagnetic interaction; quantum theories for the other three forces also had to be constructed.

Intimately linked to this problem was the proliferation of "elemental" particles. In 1897, Joseph John Thomson (1856-1940) discovered the electron as a universal component of matter. The proton (which coincides with the nucleus of hydrogen) was definitively identified thanks to experiments carried out by Wilhelm Wien (1864-1928) in 1898 and Thomson in 1910. And the neutron (a particle without a charge) was discovered in 1932 by the English physicist James Chadwick (1891-1974). In December of that same year, the United States physicist Carl Anderson (1905-1991) discovered the positron (identical to the electron, but with the opposite charge, that is, positive). That latter particle had already been predicted in theory by the relativistic equation for the electron, introduced in 1928 by one of the pioneers in the determination of the basic structure of quantum mechanics, the English physicist Paul Dirac (1902-1984).



Fred Hoyle at a seminar on nucleosynthesis at Rice University (United States, March 1975).

Electrons, protons, neutrons, photons and positrons were only the first members of an extended family (actually, families) that has not stopped growing since then, especially with the advent of machines called "particle accelerators." This branch of physics is the most characteristic of what has come to be known as *Big Science*, that is, science requiring enormous economic resources and very numerous teams of scientists and technicians. Its most distinguished founder was Ernest O. Lawrence (1901-1958), who began developing one type of accelerator at the University of Berkeley in California in the 1930s. Called "cyclotron," this type of accelerator causes "elemental" particles to move faster and faster, gaining energy with every revolution until they are forced to collide with each other. Such collisions are photographed in order to study the products, among which new "elemental" particles appear. I will further discuss this field—called "high-energy physics"—later on, when I cover the second half of the twentieth century. For the time being, it is enough to say that its origin lies in the first half of the century.

This, then, is the general context. Let us now address the second half of the century, which is the true subject of the present article. I will begin with the most general setting: the Universe, in which gravitational interaction plays a central role, though not, as we will see, an exclusive one—especially in the first moments of its existence.

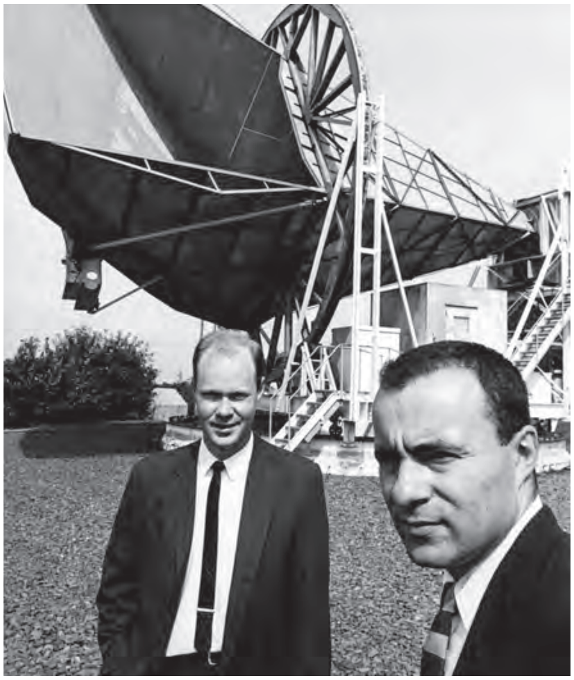
The world of gravitation

Evidence of the Universe's expansion: cosmic microwave radiation

I mentioned above that not all physicists, astrophysicists and cosmologists understood the expansion discovered by Hubble as evidence that the Universe had a beginning, a Big Bang. Hoyle, Bondi and Gold's steady-state cosmology offered a theoretical framework in which the universe had always been the same, and that idea was widely accepted. Nevertheless, in the decade following its formulation, the nineteen fifties, it began to have problems. This was not due to theoretical considerations, but to the new observational possibilities offered by technological development. This matter merits emphasis: what we call science is the product of a delicate combination of theory and observation. There can be no science without the construction of systems (theories) that describe groups of phenomena, but it is equally inconceivable without observations of what really happens in nature (we are simply unable to imagine how nature behaves). That observation requires instruments, and the more powerful they are—that is, the more they are able to improve the

[3] Fukuda, Miyamoto and Tomonaga (1949), Schwinger (1949) and Feynman (1949). For their work, the three sahred the Nobel Prize for Physics in 1965.

[4] Until then, it had been thought that atomic nuclei were made up of protons (positively charged) and electrons (negatively charged). This was considered the only possible explanation of the emission of electrons (beta radiation) that takes place in radioactive processes. Beta disintegration was finally explained using one of the most striking properties of quantum physics: the creation and annihilation of particles: electrons are not in the nucleus, they are *created* by beta disintegration.



Robert Wilson and Arno Penzias in front of the antenna with which they discovered the cosmic microwave background.

potential of our own senses— the better. This, then, is a matter of technological development, and the second half of the twentieth century was a period in which technology underwent gigantic development—much greater than any previous period—that very positively affected scientific advancement in general, and astrophysics and cosmology in particular. In that sense, the problems affecting steady-state cosmology, mentioned above, were revealed by the development of radio-astronomy, a field that began in the nineteen thirties, thanks to the work of Karl Jansky (1905-1950), an electrical engineer working for Bell Laboratories (strictly speaking: Bell Telephone Laboratories), the "department" of the American Telephone and Telegraph Corporation in charge of research and development. In 1932, while looking for possible sources of noise in radio transmissions, Jansky detected electrical emissions coming from the center of our galaxy. Despite the importance we assign to his observations in hindsight, Jansky did not continue exploring the possibilities they offered. After all, pure research was not his field.

Not immediately, but soon thereafter, primitive antennae evolved into refined radiotelescopes —usually dishes of ever-greater diameter—that pick up electromagnetic radiation from outer space. The importance of those instruments for the study of the Universe is obvious: the optical telescopes on which astronomy and astrophysics had been based until then could only study a very narrow range of the electromagnetic spectrum. They were, so to speak, almost "blind."

One of the first places that radio-astronomy flourished was Cambridge (England). It was there where Martin Ryle (1918-1984) decidedly to follow the path opened by Jansky. In doing so, he drew on knowledge he had obtained during World War II when he worked at the government's Telecommunications Research Establishment (later called the Royal Radar Establishment). He was also aided by improvements in electronic instruments brought about by the war. In 1950, using radio-telescopes that included components he designed himself, Ryle identified fifty radio-sources. That number grew radically, reaching two thousand in just five years. One of his discoveries was a radio-source in the Cygnus constellation, 500 light-years from the Milky Way. As he looked deeper into space, he was also looking farther back in time (the signals he was receiving had been emitted long ago—but it took them that long to reach the Earth). His observations were thus a look into the past history of the Universe. Hubble had taken the first great step *en route* to observational cosmology, and Ryle—who was awarded the Nobel Prize for Physics in 1974—took the second one.

Thanks to his observation of radio-sources, Ryle reached conclusions opposed to steady-state cosmology, thus favoring the Big Bang theory. In analyzing the curves that related the number of radio-stars per unit of solid angle with the intensity of their emissions, Ryle (1955) concluded that he saw no "way in which those observations could be explained in terms of steady-state theory."

A far more conclusive argument in favor of the existence of a major explosion in the past was provided by one of the most famous and important discoveries in the history of astrophysics and cosmology: microwave background radiation.

In 1961, E. A. Ohm, a physicist at one of the Bell Laboratory installations in Crawford Hill, New Jersey, built a radiometer to receive microwaves from NASA's Echo balloon (a reflector of electromagnetic signals launched in 1960). This was no coincidence: Bell Laboratories wanted to begin work in the field of communications satellites. In observations carried out on the 11-cm. wavelength, Ohm encountered a temperature excess of 3.3º (degrees, Kelvin) in his antenna, but that result was hardly noticed.[5]

Another instrument being developed at Crawford Hills at that time was an antenna whose horn shape was supposed to reduce interferences. The original idea was to use this antenna to communicate, via the Echo balloon, with the company's Telstar satellite. The antenna had to be very precise because the balloon's shape caused signals bouncing off it to be very diffused. A postdoctoral fellow at the California Technological

[5] The relation between temperatures and wavelengths can be obtained using laws such as those by Boltzmann, Wien or Planck. The relation between Celsius and Kelvin degrees is defined by the following relations: 0ºC equals 273.15ºK and 100ºC equals 373.15ºK.

Institute (Caltech), Robert Wilson (b. 1936), knew about this antenna and left his post for a job at Bell Laboratories. A Columbia University graduate, Arno Penzias (b. 1933), was three years older than Wilson and had already been at Bell for two years. Fortunately, that same year, the small but highly sensitive antenna became available for radio-astronomy use, as the company had decided to abandon the business of satellite communications. While making measurements on a wavelength of 7.4 centimeters, Penzias and Wilson found a temperature of 7.5°K that should only have been 3.3°K. Moreover, this extra radiation (or temperature), which they believed to be the effect of some sort of background noise, turned out to be constant, no matter which direction the antenna was pointing. The data indicated that the origin of what they were measuring was not in the atmosphere, or the sun, or even our galaxy. It was a mystery.

Having verified that the noise did not come from the antenna itself, the only possible conclusion they could draw was that it had something to do with the cosmos, although they did not know what its cause might be. The answer to that question came from their colleagues at nearby Princeton University, some of whom, like James Peebles (b. 1935), had already considered the idea that if there had been a Big Bang, there should be some background noise remaining from the primitive Universe. Such noise, in the form of radiation, would correspond to a much lower temperature—due to cooling associated with the Universe's expansion—than the enormously high one that must have coincided with the initial explosion. Peebles' ideas led his colleague at Princeton, Robert Dicke (1916-1995), to begin experiments intended to find that cosmic background radiation. Unwittingly, Penzias and Wilson stumbled upon it first. It was, however, the Princeton group that supplied the interpretation of Penzias and Wilson's observations (1965), which had been published by the them with no mention of their possible cosmological implications. According to current estimates, the temperature corresponding to that radiation in the microwave realm is around 2.7°K (in their 1965 article, Penzias and Wilson put it at 3.5K).

It is significant that Penzias and Wilson detected the microwave background at a center dedicated to industrial research, where new instruments were developed and available. It is a perfect example of what we mentioned above: the necessity for more precise instruments and new technology in order to advance our knowledge of the Universe. As such technology became available, the image of the cosmos grew, and this led to more discoveries, two of which I will discuss below: pulsars and quasars.

Pulsars and quasars

In 1963, Cyril Hazard, an English radio-astronomer working in Australia, precisely established the position of a powerful radio-source, called 3C273. With that data, Maarten Schmidt (b. 1929), a Dutch astronomer working at the Mount Palomar Observatory in California, optically located the corresponding emitter, discovering that the spectral lines of 3C273 were shifted towards the red side of the spectrum to such a degree that it was clearly moving away from the Earth at an enormous speed: sixteen percent of the speed of light. Hubble's law, which states that the distance between galaxies is directly proportional to their speed of recession, indicated that 3C273 was very far away. This, in turn, implied that it was an extremely luminous object—more than one hundred times as bright as a typical galaxy. Objects of this type are called *quasi-stellar sources,* or *quasars* for short, and are thought to be galaxies with very active nuclei.

Since 3C273 was discovered, several million more quasars have been found. They constitute ten percent of all light-emitting galaxies and many astrophysicists believe that many of the most brilliant galaxies pass briefly through a quasar phase. Most quasars are very far from our galaxy, which means that the light that reaches us must have been emitted when the Universe was much younger. That makes them magnificent instruments for the study of the Universe's history.

In 1967, Jocelyn S. Bell (b. 1943), Anthony Hewish (b. 1924) and the latter's collaborators at Cambridge built a detector to observe quasars at radio frequencies. While using it, Bell observed a signal that appeared and disappeared with great rapidity and regularity. Its cycle was so constant that it seemed to have an artificial origin (could it possibly be a sign of intelligent extraterrestrial life?). Following a careful search, however, Bell and Hewish concluded that those "pulsars," as they were finally denominated, had an astronomical origin (Hewish, Bell, Pilkington, Scott and Collins 1968).[6] But what were those highly regular radio sources? A theoretical interpretation was not long in coming, and was provided by Thomas Gold, one of the "fathers" of steady-state cosmology, who had now accepted the Big Bang. Gold (1968) realized that such short cycles (around one to three seconds in the first detected pulsars), could only come from a very small source. White dwarfs were too large to rotate or vibrate at such a frequency, but neutron stars could.[7] But did the origin of the signals being received lie in the vibration or rotation of such stars? Certainly not their vibrations, because neutron stars vibrate much too fast (around a thousand times a second) to explain the cycles of most pulsars. So pulsars had to be rotating neutron stars.

[6] In 1974, Hewish shared the Nobel Prize for Physics with Ryle. Jocelyn Bell, who had first observed pulsars, was left out.

[7] The possible existence of neutron stars—a sort of giant nucleus made entirely of neutrons linked by the force of gravity—was first proposed in 1934 (that is, just two years after Chadwick discovered the neutron) by the California-based (Caltech) Swiss physicist and astrophysicist, Fritz Zwicky (1898-1974). According to general relativity, the minimum mass that would allow a neutron star to exist is 0.1 solar masses, while the maximum seems to be around 6 solar masses. In the case of a neutron star of one solar mass, its radius would be about 13 kilometers and its density $2\cdot10^{17}$ kilos per cubic meter, which is about $2\cdot10^{14}$ times as dense as water.

Since then, scientists have discovered pulsars that emit X-rays or gamma rays (and some even emit light in the visible spectrum), so nowadays, scientists also accept the possibility of other mechanisms for the production of their radiation emissions, including the accretion of matter in double systems.

Besides their astrophysical interest, pulsars serve other functions. They have been used to test general relativity's prediction that accelerated masses emit gravitational radiation (a phenomenon analogous to that produced by electrical charges: electromagnetic radiation).

Confirmation that gravitational radiation does, in fact, exist came in 1974, with the discovery of the first system consisting of two pulsars interacting with each other (called PSR1913+16), for which Russell Hulse (b. 1950) and Joseph Taylor (b. 1941) received the 1993 Nobel Prize for Physics. In 1978, after various years of continuous observation of that binary system, they were able to conclude that the orbits of those pulsars vary and are growing closer together. That result was thought to indicate that the system is losing energy due to the emission of gravitational waves (Taylor, Fowler and McCulloch 1979). Since then, other binary pulsar systems have been discovered, but it is still not possible to detect gravitational radiation with instruments built and installed on Earth. This is extremely difficult, due to the extreme faintness of the affects involved. The gravitational waves that would arrive at the Earth from some part of the Universe where an extremely violent event had taken place would produce distortion in the detectors no greater than one part out of 10^{21}. That would be a tiny fraction the size of an atom. However, there are already devices designed to achieve this: the four-kilometer system of detectors in the United States known as LIGO (Laser Interferometric Gravitational wave Observatories).

Quasars are also very useful for studying the Universe in conjunction with general relativity. About one of every five-hundred quasars is involved in a very interesting relativist phenomenon: the diversion of the light it emits due to the gravitational effect of other galaxies situated between that quasar and the Earth, from which that effect is being observed. This effect is called "gravitational lensing",[8] and can be so powerful that multiple images of a single quasar are observable.

Actually, gravitational lenses are not produced exclusively by quasars, they are also produced by large accumulations of masses (such as cumuli of galaxies) which divert light from, for example, galaxies behind them (with respect to us) so that, instead of a more-or-less clear image, we see a halo of light, a "double image." They were first observed in 1979, when Walsh, Carswell and Weyman (1979) discovered a multiple image of a quasar in 0957+561. Since then, the Hubble space telescope has photographed a cumulus of galaxies about a thousand million light-years away in which, besides the light of the cumulus of galaxies itself, it is possible —though difficult because of their lesser luminescence— to detect numerous arcs (segments of rings). Those arcs are actually images of galaxies much farther away from us that the cumulus, but seen through the effect of the gravitational lens (the cumulous acts as a lens, distorting the light coming from those galaxies). Beside offering new evidence supporting general relativity, these observations have the added value that the magnitude of diversion and distortion visible in those luminous arcs is far greater than could be expected if the cumulus only contained the galaxies we see in it. In fact, evidence indicates that those cumuli contain between five and ten times more matter than we can see. Could this be the dark matter we will discuss further on?

For many scientists—at least until the problem of dark matter and dark energy took the fore—the background radiation, pulsars and quasars discussed in this section were the three most important discoveries in astrophysics during the second half of the twentieth century. What those discoveries tell us, especially pulsars and quasars, is that the Universe is made up of much more surprising and substantially different objects than were thought to exist in the first half of the twentieth century. Of course, when we speak of *surprising* or *exotic* stellar objects, we inevitably have to mention black holes, another "child" of the general theory of relativity.

Black holes

For decades after Einstein's theory was formulated in 1915 and its predictions about gravity with relation to the Solar System were exploited (the anomalous, with regards Newton's theory, movement of Mercury's perihelion, the curvature of light rays and the gravitational shift of spectral lines), general relativity was mostly in the hand of mathematicians—men like Hermann Weyl (1885-1955), Tullio Levi-Civita (1873-1941), Jan Arnouldus Schouten (1883-1971), Cornelius Lanczos (1892-1974) or André Lichnerowicz (1915-1998). This was partially due to the theory's mathematical difficulty and partially to the lack of almost any real situation in which to apply it. That theory mainly addressed the Universe, and exploring it would require technological means that did not even exist at the time, not to mention significant financial support. This problem began fading at the end of the nineteen sixties, and it can now be said that general relativity is fully integrated into experimental physics, including areas that are not even that close, such as the Global Positioning System (GPS). It is not only a

[8] As on other occasions, Einstein (1936) had already predicted the existence of this phenomenon.

part of experimental physics related to astrophysics and cosmology; as we will see further on; it is also a part of high-energy physics.

And here we must mention one of the most surprising and attractive stellar objects linked to general relativity discovered in the last few decades: black holes, whose existence has even reached beyond purely scientific circles and entered the social realm.

As I said, these object belong to the theoretical tenets of general relativity, although their Newtonian equivalents had already been proposed—and forgotten—much earlier by the British astronomer John Michell (c. 1724-1793) in 1783, and later by Pierre Simon Laplace (1749-1827) in 1795. Their exoticism derives from the fact that they involve such radical notions as the destruction of space-time at points called "singularities."[9]

Studies leading to black holes began in the nineteen thirties, when the Hindu physicist Subrahamanyan Chandrasekhar (1910-1995) and the Russian Lev Landau (1908-1968) demonstrated that in the Newtonian theory of gravitation, a cold body with a mass superior to 1.5 times that of the Sun could not support the pressure produced by gravity (Chandrasekhar 1931; Landau 1932). That result led scientists to ask what general relativity predicted for the same situation. In 1932, Robert Oppenheimer (1904-1967) and two of his collaborators, George M. Volkoff and Hartland Snyder (1913-1962) demonstrated that a star with that mass would collapse until it was reduced to a singularity; that is, to a point with a volume of zero and an infinite density (Oppenheimer and Volkoff 1939, Oppenheimer and Snyder 1939).

Stephen Hawking.

Oppenheimer and his collaborators' work received little attention or credence and it was ignored until interest in strong gravitational fields was spurred by the discovery of quasars and pulsars. In 1963, Soviet physicists, Evgenii M. Lifshitz (1915-1985) and Isaak M. Khalatnikov (b. 1919) took the first step and began studying the singularities of relativist space-time. Following the work of his Soviet colleges, the British mathematician and physicist Roger Penrose (b. 1931) and the physicist Stephen Hawking (b. 1942) applied powerful mathematic techniques to this question in the mid-nineteen sixties. They demonstrated that such singularities were inevitable when a star collapsed, providing certain conditions were met.[10]

A couple of years after Penrose and Hawking published their first articles, the physics of space-time singularities became that of "black holes," a felicitous term that has attracted considerable popular attention to this physical entity. The man responsible for this apparently insignificant terminological revolution was the United States physicist John A. Wheeler (1911-2008). He, himself, explained the genesis of that term in the following manner (Wheeler and Ford 1998, 296-297):

> In the fall of 1967, Vittorio Canuto, administrative director of NASA's Goddard Institute for Space Studies at 2880 Broadway in New York, invited me to give a lecture on possible interpretations of the new and stimulating evidence arriving from England about pulsars. What were those pulsars? Vibrating white dwarfs? Rotating neutron stars? What? In my lecture, I argued that we should consider the possibility that, at the center of a pulsar, we might find a completely collapsed gravitational object. I pointed out that we could not continue to say, over and over again, "completely collapsed gravitational object." We needed a much shorter descriptive term. "How about black hole" asked someone in the audience. I had been looking for the right term for months, ruminating in bed, in the bathtub, in my car, whenever I had a free moment. Suddenly, that name seemed totally correct to me. A few months later, on 29 December 1967, when I gave the more formal Sigma Xi-Phi Kappa lecture at the New York Hilton's West Ballroom, I used that term, and I later included it in the written version of the lecture published in spring 1968.

The name was catchy, and it stuck, but the explanation was mistaken (as I pointed out above, a pulsar is driven by a neutron star).

While the history of black holes began with the physics work of Oppenheimer and his collaborators,

[9] We must remember that from the standpoint of the general theory of relativity, space-time and gravity represent the same physical concept, as the curvature of the former is what describes the latter.

[10] See, for example, Penrose (1965), Hawking (1965, 1966a, 1966b) and Hawking and Penrose (1969). I will not go into more depth here, but I do want to point out that other scientists also took part in this project, including G. F. R. Ellis.

mentioned above, for some years, the field was dominated by purely mathematical studies like the previously mentioned ones by Penrose and Hawking. The underlying physical idea was that they must be very different than any other type of star, even though their origins were linked to them. They would occur when, after exhausting its nuclear fuel, a very massive star began to contract irreversibly, due to gravitational force. A moment would thus arrive when it would form a region (called "horizon") in which matter and radiation could only enter, without anything being able to get out, not even light (from whence the denomination, "black"). The larger such an object was, the more it would "eat", and the more it ate, the bigger it would get. The center of a black hole is its point of *collapse.* According to general relativity, there, the matter that once made up the star is compressed and expulsed, apparently "out of existence."

Clearly, "out of existence" is not an acceptable idea. However, there is a possible way out of such a paradoxical situation: the general theory of relativity is not compatible with quantum requirements, but clearly, when matter is compressed into a very reduced area, its behaviour will follow quantum rules. Thus, a true understanding of the physics of black holes calls for a quantum theory of gravitation (either by quantizing general relativity, or by constructing a new theory of gravitational interaction that can be quantized). At the present time, this has yet to be done, although some steps have been made in that direction, including one by Hawking himself, the grand guru of black holes. What is called "Hawking's radiation" (Hawking 1975), predicts that, due to quantum processes, black holes are not as black as we though, and are able to emit radiation.[11]

As a result, we do not really know what those mysterious and attractive objects are. Do they, in fact, exist at all? The answer is yes. There are ever-greater indications that they do. On 12 December 1970, the United States launched a satellite from Kenya to celebrate its independence. Called Uhuru—the Swahili word for "freedom"—this satellite carried instruments capable of determining the position of the most powerful sources of X rays. Among the 339 identified sources is Cygnus X-1, one of the most brilliant in the Milky Way, located in the region of the Swan. This source was later linked to a visible super-giant blue star with a mass 30 times that of the Sun and an invisible companion. The movement of the blue star indicated that its companion had a mass 7 times that of the Sun, a magnitude too great to be a white dwarf or a neutron star. It must be, therefore, a black hole. However, some argue that its mass is 3 solar masses, in which case it could be a neutron star. Anyhow, at least other 10 binary systems have been found in which one of its members seems to be a black hole: for example, V404 Cygni, formed by a star with 2/3 the mass of the Sun, and a black hole of 12 solar masses.

It is now generally accepted that there are super-massive black holes at the center of those galaxies whose nucleus is more luminous that all the rest of the galaxy (about 1% of all galaxies in the Universe are that way). In over two hundred cases, it has been possible to indirectly determine the masses of those super black holes, but a direct determination has only been possible in a few cases. One of the latter is in our own Milky Way.

Inflation and "wrinkles in time"

The study of the Universe is enormously puzzling. Obviously, measuring such basic data as distances, masses and velocities is extremely complex there. We cannot do so directly, nor can we "see" everything with precision. With the data then available, there was a time when the model that offered the Robertson-Walker-Friedmann solution to general relativity was sufficient. It represents a Universe that expands with an acceleration that depends on its mass-energy content. But there were increasingly clear problems with the cosmology of the Big Bang.

One of these was the question of whether mass-energy is such that the Universe will continue to expand forever, or if it is large enough that gravitational attraction will eventually overcome the force of the initial explosion, reaching the point where it begins to contract and finally arrives at a *Big Crunch.* Another problem lay in the considerable uniformity with which mass appears to be distributed throughout the Universe. This is observable using units of measurement of some 300 million light-years or more (of course, on a small scale, the Universe, with its stars, galaxies, cumuli of galaxies and enormous interstellar voids, is not homogeneous). Background microwave radiation is good proof of this macro-homogeneity. Now then, using the standard Big Bang theory, it is difficult to explain this homogeneity in terms of known physical phenomena: moreover, considering that information about what happens cannot be transmitted between different points in space-time any faster that the speed of light, it turns out that during the first moments of the Universe's existence it would not have been possible for different regions to "reach a consensus," so to speak, about what the mean density of matter and radiation should be.[12]

To resolve this problem the idea of an inflationary Universe was proposed. It hypothesizes that, during the Universe's first instants of existence, there was a gigantic, exponential increase in the speed of its

[11] Such an emission would lead to a slow decrease in the mass of a black hole. If that decrease were continuous, the black hole could eventually disappear. For normal black holes (those of just a few solar masses), however, that would not happen. For example, a black hole of just one solar mass would have a lower temperature than that of the radiation coming from the microwave background, which means that black holes of such mass would absorb radiation faster than they could emit it, so they would continue to increase in mass. If, however, there were very small black holes (made, for example, during the first instants of the Universe by fluctuations in density that must have happened at that time), then they would have a much higher temperature, emitting more radiation than they could absorb. They would lose mass, which would make them even hotter, and would finally blow up in a large explosion of energy. Their life would be such that we might be able to observe such explosions how. None has yet been detected, however.

[12] This difficulty is called the "horizon problem."

expansion. In other words, the mini-universe must have experienced a growth so rapid that there was not enough time to develop physical processes that would have led to non-homogeneous distributions. Once that inflationary stage ended, the Universe must have continued evolving according to the classic Big Bang model.

Among the scientists responsible for this inflationary theory, we should mention the American, Alan Guth (b. 1947) and the Soviet, Andrei Linde (b. 1948).[13] But, more than specific names, what I want to point out is that it is impossible to understand this theory without recourse to high-energy physics—what used to be called elementary-particle physics, which I will discuss further on—especially the Grand Unified Theories (GUT), which predict that there would have to be a phase shift at temperatures around 10^{27} degrees Kelvin.[14] Here, we have an example of one of the most important phenomena to take place in the field of physics during the second half of the twentieth century: the encounter between cosmology (the science of "the big") and high-energy/elemental-particle physics (the science of "the small"). Naturally, their meeting place is the first instants of the Universe's existence, when the energies involved were gigantic.

So, inflation lies at the origin of a uniform Universe. But then, what caused the miniscule primordial non-homogeneities that, with the passage of time and the effect of gravitational force, gave birth to cosmic structures such as galaxies?

One possible answer is that inflation may have enormously amplified the ultramicroscopic quantum fluctuations that occurred as a result of the uncertainty principle applied to energies and time ($\Delta E{\cdot}\Delta t{\geq}h$). If that were the case, what better place to look for non-homogeneities than the microwave radiation background?

The answer to this question appeared in the work of a team of US scientists led by John C. Mather (b. 1946) and George Smoot (b. 1945). In 1982, NASA approved funding for the construction of a satellite—the Cosmic Background Explorer (COBE), which was put into orbit 900 kilometers above the Earth in the fall of 1989—to study the cosmic microwave background. The entire project was coordinated by Mather, including the experiment (in which he used a spectrophotometer cooled to 1.5°K) that showed that the shape of the microwave radiation background corresponds to that of the radiation of a black body at a temperature of 2.735°K. Meanwhile, Smoot measured the miniscule irregularities predicted by inflation theory. Ten years later, following the work of over a thousand people and a cost of 160 million dollars, it was announced (Mather et al. 1990; Smoot et al. 1992) that COBE had detected what Smoot called "wrinkles" in space-time, the seeds that led to the complex structures—such as galaxies—we now see in the Universe.[15]

Just how thrilled those researchers were when they confirmed their results is clear in a book for lay readers published by Smoot soon thereafter. *Wrinkles in Time* (Smoot and Davidson, 1994, 336):

> I was looking at the primordial form of the wrinkles, I could feel it in my bones. Some of the structures were so huge that they could only have been generated when the Universe was born, no later. What was before my eyes was the mark of creation, the seeds of the present Universe.

Consequently, "the Big Bang theory was correct and the notion of inflation worked; the wrinkles model fit in with the formation of structures from cold dark matter; and the magnitude of the distribution would have produced the larger structures of the current universe under the influence of gravitational collapse over the course of 15,000 million years."

COBE was a magnificent instrument, but it was by no means the only one. There are many examples of astrophysics and technology working hand in hand, not only with Earth-based instruments, but also spacecraft. At this point, scientists have been exploring our Solar System for quite some time using satellites with refined instruments that send us all sorts of data and images: space probes such as Mariner 10, which observed Venus from a distance of 10,000 kilometers in 1973; Pioneer 10 and Voyager 1 and 2, which approached Jupiter, Saturn, Uranus and Pluto between 1972 and 1977, and Galileo, aimed at Jupiter and its moons.

A very special type of vehicle is the Hubble space telescope, which NASA put into orbit following a long process in the spring of 1990.[16] A telescope in an artificial satellite has the advantage of being outside the Earth's atmosphere, which is the greatest barrier to the reception of radiation. Since it was launched, and especially since its defects were corrected, Hubble has sent, and continues to send, spectacular images of the Universe. Thanks to it, we have the first photos of regions (such as the Orion nebulous) where it appears that stars are being born. It would not be a complete exaggeration to say that Hubble has revolutionized our knowledge of the Universe.

Extrasolar planets

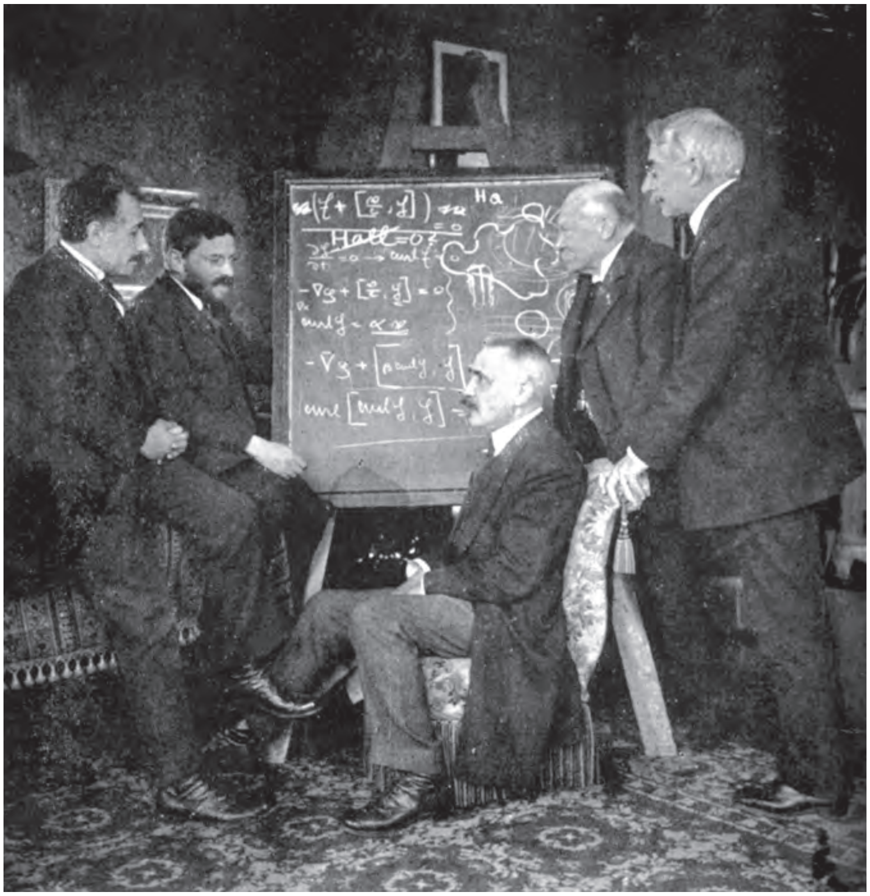
Thanks to technological advances, scientists are starting to be able to see new aspects and objects in the cosmos, such as planetary systems associated with stars other than the Sun. The first discovery of this sort took place in 1992, when Alex Wolszczan and Dale Frail found that at least two Earthlike planets were orbiting around a pulsar (Wolszczan and Frail 1992). Three years later, Michel Mayor and Didier Queloz announced their

[13] Guth (1981), Linde (1982).

[14] In a phase shift, there is a sudden change in the state of the system in question. On example is when water (liquid) turns into ice (solid).

[15] For their work, both received the Nobel Prize for Physics in 2006.

[16] Until the early nineteen nineties, the largest mirrors on optical telescopes installed on Earth were between five and six meters in diameter. The largest telescope at that time, with a primary (or collector) mirror six meters in diameter, was in the Russian Caucasus. It was followed by the telescope at the Mount Palomar Observatory, inaugurated soon after World War II, with a diameter of 5 meters, and then a long list of telescopes with mirrors around 4 meters in diameter. Now a series of large telescopes have been completed or are being built whose characteristics and employment of the most modern technology is making an important quantitative leap—and in many senses, a qualitative one, as well—in astrophysical research. These are telescopes of up to ten meters, such as the one already in use at Mauna Kea, Hawaii, which belongs to the California Institute of Technology and the University of California. Another of similar characteristics is being built at the same location. With the one already in use, which is the largest in the world, it has been possible to observe a brown dwarf (PPL15) in the Pleiades cumulus. This kind of star is so small that it does not shine like others, so none had ever been seen before, although they could sometimes be detected because of their gravitational effects. Another instrument of this size is the Grand Telescope at the Canary Islands Astrophysics Institute. Installed at Roque de los Muchachos, it has seen its "first light" recently.



Albert Einstein with Paul Ehrenfest, Paul Langein, Heike Kamerlingh Onnes and Pierre Weiss at Erenfest's house (Leiden, 1920).

discovery of a planet of the same size and type as Jupiter (a gaseous giant) orbiting around the star 51 Peasi (Mayor and Queloz 1995). Since then, the number of known extrasolar planets has grown considerably. And if such planets exist, life may have developed on some of them as well. Now, while the biology that addresses the problem of the origin of life supports the possibility that in sufficiently favorable environments combinations of chemicals could produce life through synergic processes, most probably such life would be of a different type than human life. Evolutionist biology, supported by geological data, has shown that the human species is the product of evolutionary chance. If, for example, an asteroid or comet approximately ten kilometers in diameter had not collided with the Earth some 65 million years ago—it hit the Earth at a speed of about thirty kilometers a second, producing energy equivalent to the explosion of one hundred million hydrogen bombs—then an enormous number of plant and animal species might never have disappeared (or certainly not then). These included the dinosaurs that impeded the rise of those small mammals that later evolved into *homo sapiens* and other species.

It is that element of chance that makes it impossible to be certain there is intelligent life on other planets —in our galaxy or others—or that such a life form might be trying, or have tried, to understand nature, build scientific systems, and attempt to communicate with other living beings that may exist in the Universe. Still, for quite some time, research programs have been scanning the Universe in search of signs of intelligent life—programs such as the Search of Extra-Terrestrial Intelligence (SETI), which has used 250-million-channel receivers that carry out around twenty thousand million operations per second.

Dark matter and dark energy

The existence of extrasolar planets certainly thrills and moves us, but it is not something "fundamental." It does not shake the foundations of science. But other discoveries relative to the contents of the Universe are a very different matter. For example, we have good reasons to believe that the cosmos contains a large amount of invisible matter that exercises gravitational force. The most immediate evidence comes from rotating disk-shaped galaxies (such as our own Milky Way). When we look at the outer part of such galaxies, we see that their gas moves at a surprising speed —much faster than it should, given the gravitational attraction produced by the stars and gasses we can detect inside it. Other evidence comes from the internal movement of galaxy cumuli. This "dark" matter is thought to constitute thirty percent of all the matter in the Universe, but what is its nature? That is one of the problems. It could consist of barely luminescent stars (such as brown dwarfs), or exotic elemental particles, or black holes. We cannot really understand what galaxies are, or how they came into being, until we know what this dark matter is. Nor will we be able to know what the ultimate destiny of the Universe is.

Along with dark matter, another similar question came to the fore in the last decade of the twentieth century: dark energy. While studying a type of supernova —stars that have exploded, leaving a nucleus—a group led by Saul Perlmutter (at the Lawrence Berkeley National Laboratory in California) and another by Brian Schmidt (at the Mount Stromlo and Siding Spring observatories in Australia) arrived at the conclusion that, contrary to previous suppositions, the Universe's expansion is accelerating (Perlmutter et al. 1998; Schmidt et al. 1998). The problem was that the Universe's mass could not explain such an acceleration; it was necessary to assume that gravity was behaving in a surprising new way: pushing masses away from each other rather than attracting them to each other. It had been assumed that the Big Bang must have been driven by a repulsive energy during the creation of the universe, but no one had imagined that such energy could continue to exist in the now-mature Universe.

Thus, a new energy came into play, a "dark" energy residing in empty space. And since energy is equivalent to mass, that dark energy signified a new contribution to the total mass of the Universe, thought not the same as dark matter. It is now thought that around 3% of the Universe consists of ordinary mass, 30%, of dark mass, and the other 67%, of dark energy. In other words: we thought we knew what the Universe is, and it turns out to be practically unknown to us, because we know the nature and make up of neither dark matter nor dark energy. One possible explanation of the latter could be found in the term introduced by Einstein in 1916 in his field equations for general relativity. As we saw, when applying his theory of gravitational interaction to the entire Universe, Einstein sought a model that would represent a static Universe. That obliged him to introduce a new term into his equations, the previously mentioned cosmological constant, which actually represents a field of repulsive forces that compensate for the attractive effects of gravitation. When relativistic cosmology found solutions that represent an expanding Universe, and that expansion was demonstrated by observation (Hubble), Einstein thought that it was no longer necessary to maintain that constant, although it could be included without any difficulty in theoretical expansive models. Now, it seems necessary to resurrect this term, but it will not be enough to include it in relativist cosmology again; it has to find its place and meaning in quantum theories that attempt to make gravity a part of quantum system. After all, dark energy is the energy of the void, and from a quantum viewpoint, vacuum has a structure. And given that quantum physics has again entered the picture here, let us discuss how the quantum revolution developed and solidified during the second half of the twentieth century.

A quantum world

High-energy physics: from protons, neutrons and electrons to quarks

When discussing the quantum revolution that emerged during the first half of the twentieth century, I mentioned the search for the basic components of matter, the so-called "elemental particles." There, we saw that moving beyond protons, electrons and neutrons, the most basic of those particles, required more elevated energy than could be supplied by the "projectiles"—alpha particles, for example—coming from the emissions of radioactive elements (especially, radium). We also saw that it was Ernest Lawrence who found a new way forward, developing instruments called particle accelerators (in his case, cyclotrons), which functioned by accelerating particles to high energy levels and then making them collide with each other (or with some predetermined target). The idea was to examine what was produced by such collisions, that is, what new and smaller components make up such particles if, in fact, there are any.[17]

The physics of elemental particles, also called high-energy physics, as I indicated above, became one of the main protagonists of the second half of the twentieth century. This is very expensive science (it is the epitome of *Big Science,* which requires large teams of scientists and technicians and large investments), and is becoming ever more expensive, as the size of accelerators grows, making it possible to reach higher energy levels.

After World War II, especially in the United States, high-energy physics drew on the prestige of nuclear physics, which had supplied the powerful atomic bombs. Here, I will mention only the most important accelerators. In 1952, the Cosmotron entered service in Brookhaven, New York. It was for protons and reached 2.8 GeV;[18] It was followed, among others, by the Bevatron (Berkeley, protons; 1954), with 3.5 GeV; Dubna (USSR, protons; 1957), 4.5 GeV; the Proton-Synchroton (CERN, Geneva, protons; 1959), 7 GeV; SLAC (Stanford, California; 1966), 20 GeV; PETRA (Hamburg, electrons and positrons; 1978), 38 GeV; Collider (CERN, protons and antiprotons; 1981), 40 GeV; Tevatron (Fermilab, Chicago, protons and antiprotons), 2,000 GeV, and SLC (Stanford, electrons and positrons), 100 GeV, both in 1986; LEP (CERN, electrons and positrons; 1987), 100 GeV, and HERA (Hamburg, electrons and protons; 1992), 310 GeV.

The initials, CERN, correspond to the *Centre Européen de Recherches Nucleaires* (European Nuclear Research Center), an institution created by twelve European nations in Geneva in 1954 to compete with the United States. CERN now includes more countries (including Spain) and with its accelerators it has played an outstanding role in the development of high-energy physics. In fact, in difficult times for this field, like the present, CERN has just completed (2008) construction of a new one in which protons will collide with an energy of 14,000 GeV: the Large Hadron Collider (LHC). Thus, old Europe carries the torch and "keeps the fire" for this costly branch of physics.

So why do I speak of "difficult times for this field?" Because due to its high cost, this branch of physics has been having difficulties in recent years. In fact, it was recently dealt a serious blow by what had been, until then, its strongest supporter: the United States. I am referring to the Superconducting Super Collider (SSC). This gigantic accelerator, which U.S. high-energy physicists considered indispensable for continuing

[17] Strictly speaking, it was not Lawrence who opened the door to elemental-particle physics using non-radioactive sources, although it is true that he did find the most adequate technical procedure. At Cambridge in 1932, John D. Cockcroft (1897-1967) and Ernst T. S. Walton (1903-1995) used a voltaic multiplier to obtain the 500 kV (1 kV = 1000 volts) that allowed them to become the first to observe the artificial disintegration of lithium atoms into two particles. And there were more precedents, such as the generators developed by Robert J. van de Graaff (1901-1967).

[18] 1 GeV = 1000 million electron-volts. 1 electron-volt is the motion energy a single electron would gain when subjected to the potential difference of one volt.

to develop the structure of the so-called standard model, was going to consist of an 84 kilometer tunnel to be dug near a small town of 18,000 inhabitants about thirty kilometers southeast of Dallas, in Waxahachie. Inside that tunnel, thousands of magnetic superconductor spools would guide two proton beams. After millions of laps, they would reach levels twenty times higher than could be attained with existing accelerators. At various points along the ring, protons from the two beams would collide and enormous detectors would track the results of those collisions. The project would take ten years, and its cost was initially estimated at 6,000 million dollars.

Things got off to a rocky start, but the tunnel excavation was completed. However, on 19 October 1993, following prolonged, difficult and changing discussions in both houses of Congress, the House of Representatives finally cancelled the project. Other scientific programs—especially in the field of biomedicine—were more attractive to American congressmen, senators and—why deny it?—society, which was more interested in health-related matters.

However, let us abandon the subject of accelerators, and discuss their products, those particles that appear to be "elemental." Thanks to those accelerators, their number grew so great that it wound up drastically undermining the idea that most of them could really be elemental in the fundamental sense. Among the "particles" discovered, we can recall pions and muons of various sorts, or those called Λ, W or Z, not to mention their corresponding antiparticles.[19] The number—hundreds—of such particles grew so great that scientists began speaking of a "particle zoo," a zoo with too many occupants.

One of its inhabitants was particularly striking: quarks. Their existence had been theorized in 1964 by U.S. physicists, Murray Gell-Mann (b. 1929) and George Zweig (b. 1937). Until quarks appeared in the complex and varied world of elemental particles, it was thought that protons and neutrons were indivisible atomic structures, truly basic, and that their electrical charge was an indivisible unit. But quarks did not obey this rule, and they were assigned fractional charges. According to Gell-Mann (1964) and Zweig (1964), hadrons—particles subject to strong interaction—are made up of two or three types of quarks and antiquarks called *u* (*up*), *d* (*down*) and *s* (*strange*), that respectively have electrical charges of 2/3, -1/3 and -1/3 of that of an electron.[20] Thus, a proton is made up of two *u* quarks and one *d,* while a neutron consists of two *d* quarks and one *u.* Therefore, they are composite structures. Later, other physicists proposed the existence of three other quarks: *charm* (*c;* 1974), *bottom* (*b;* 1977) and *top* (*t;* 1995). To characterize these quarks, scientists say they have six *flavors*. Moreover, each of the six types comes in three varieties, or *colors:* red, yellow (or green) and blue. And for each quark there is, of course, an antiquark.

Needless to say, terms like these—*color*, *flavor*, *up*, *down*, and so on—do not represent the reality we normally associate with such concepts, although in some cases there can be a certain logic to them, as happens with *color.* This is what Gell-Mann (1995, 199) had to say about that term:

> While the term "color" is mostly a funny name, it is also a metaphor. There are three colors, called red, green and blue, like the three basic colors in a simple theory of human color vision (in the case of painting, the three primary colors are usually red, yellow and blue, but when mixing light instead of pigment, yellow is replaced by green). The recipe for a neutron or a proton calls for a quark of each color, that is, one red, one green and one blue, so that the sum of the colors cancels out. As in vision, where white can be considered a mixture of red, green and blue, we can metaphorically state that neutrons and protons are white.

In short, quarks have *color* but hadrons do not: they are white. The idea is that only *white* particles are directly observable in nature, while quarks are not; they are "confined," that is, grouped to form hadrons. We will never be able to observe a free quark. Now in order for quarks to remain confined, there have to be forces among them that are very different than electromagnetic or other kinds of forces. "Just as electromagnetic force between electrons is mediated by the virtual exchange of photons," as Gell-Mann put it (1995, 200), "quarks are linked together by a force that arises from the exchange of other quanta: gluons, whose name comes from the fact that they make quarks stick together to form observable white objects such as protons and neutrons."[21]

About ten years after quarks appeared, a theory, quantum chromodynamics, was formulated to explain why quarks are so strongly confined that they can never escape from the hadron structures they form. Of course the name *chromodynamic*—from the Greek term *chromos* (color)—alluded to the *color* of quarks (and the adjective "quantum" to the fact that this theory is compatible with quantum requirements). Inasmuch as quantum chromodynamics is a theory of colored elemental particles, and given that color is associated with quarks, which are, in turn, associated with hadrons—"particles" subject to strong interaction—we can say that this theory describes that interaction.

With quantum electrodynamics —which, as I already stated, emerged in the first half of the twentieth century—and quantum chromodynamics, we have quantum theories for both electromagnetic and strong interactions. But what about the weak interaction, responsible for radioactive phenomena? In 1932, Enrico

[19] Each particle has its antiparticle (although they sometimes coincide): when they meet each other, they disappear—annihilating each other—producing energy.

[20] There are two types of hadrons: baryons (protons, neutrons and hyperons) and mesons (particles whose mass have values between those of an electron and a proton).

[21] It is also interesting to quote what Gell-Mann (1995, 198) wrote about the name "quark": "In 1963, when I gave the name "quark" to the elemental parts of nucleons, I based my choice on a sound that was not written that way, sort of like "cuorc." Then, in one of my occasional readings of James Joyce's *Finnegans wake,* I discovered the word "quark" in the sentence "Three quarks for Muster Mark." Given that "quark" (which is used mostly to describe the cry of a seagull) was there to rhyme with "Mark," I had to find some excuse to pronounce it like "cuorc." But the book narrates the dreams of an inn-keeper named Humphry Chipden Earkwicker. The words in the text often come from various sources at the same time, like the "hybrid words" in Lewis Carroll's *Through the Looking Glass.* Sometimes, sentences partially determined by bar slang appear. I thus reasoned that one of the sources of the expression "Three quarks for Muster Mark," might be "Three quarts for Mister Mark," in which case the pronunciation, "cuorc," would not be totally unjustified. At any rate, the number three fits perfectly with the number of quarks present in nature."

Fermi (1901-1954), one of the greatest physicists of his century, developed a theory for weak interaction, which he applied primarily to what was called "beta disintegration," a radioactive process in which a neutron disintegrates, leaving a proton, an electron and an antineutrino. Fermi's theory was improved in 1959 by Robert Marshak (1916-1992), E. C. George Sudarshan (b. 1931), Richard Feynman and Murray Gell-Mann, but the most satisfactory version of a quantum theory of weak interaction was put forth in 1967 by the US scientist, Steven Weinberg (b. 1933) and a year later by the English-based Pakistani, Abdus Salam (1929-1996). They independently proposed a theory that unified electromagnetic and weak interactions. Their model included ideas proposed by Sheldon Glashow (b. 1932) in 1960.[22] For their work, Weinberg, Salam and Glashow shared the Nobel Prize for Physics in 1979. This happened after one of the predictions of their theory—the existence of what they called "weak neutral currents"—was experimentally corroborated at CERN in 1973.

The electroweak theory unified the description of electromagnetic and weak interactions. But could it be possible to take a farther step on the path to unification, formulating a theory that would also include the strong interaction described by quantum chromodynamics? The affirmative answer to this question was provided by Howard Georgi (b. 1947) and Glashow (Georgi and Glashow 1974), who presented the first ideas of what came to be called, as we mentioned earlier, Grand Unified Theories (GUT).

This family of theories had the most impact on cosmology, especially on the description of the Universe's first instants. From the perspective of GUTs, in the beginning there was only one force, which contained electromagnetic, weak and strong forces. However, as the Universe cooled, they began to separate.

Such theoretical tools make it possible to explain questions such as the existence (at least in appearance, and fortunately for us) of more matter than antimatter in the Universe. This is due to something the different GUTs have in common: they do not conserve a magnitude called the "baryonic number," meaning that processes are possible in which the number of baryons—remember, these include protons and neutrons—produced is not equal to the number of anti-baryons. The Japanese physicist, Motohiko Yoshimura (1978) used this property to demonstrate that an initial state in which there was an equal amount of matter and antimatter could evolve into one with more protons or neutrons than their respective antiparticles, thus producing a Universe like ours, in which there is more matter than antimatter.

Thanks to the group of theories mentioned above, we have an extraordinary theoretical framework in which to understand what nature is made of. Its predictive capacity is incredible. These theories accept that all matter in the universe is made up of aggregates of three types of elemental particles: electrons and their relatives (those called muon and tau), neutrinos (electronic, muonic and tauonic neutrinos) and quarks, as well as the quanta associated with the fields of the four forces we recognize in nature:[23] photons, for electromagnetic interaction, Z and W particles (gauge bosons) for weak interaction, gluons for strong interaction; and even though gravitation has yet to be included in this framework, the as-yet-unobserved gravitons, for gravitational interaction. The subset formed by quantum chromodynamics and electroweak theory (that is, the theoretical system that includes relativistic and quantum theories of strong, electromagnetic and weak interactions) proves especially powerful in its balance of predictions and experimental confirmation. It is called the *Standard model* and, according to the distinguished physicist and science historian, Silvan Schweber (1997, 645), "the formulation of the Standard Model is one of the great achievements of the human intellect—one that rivals the genesis of quantum mechanics. It will be remembered—together with general relativity, quantum mechanics, and the unravelling of the genetic code—as one of the most outstanding intellectual advances of the twentieth century. But much more so than general relativity and quantum mechanics, it is the product of a communal effort." Allow me to emphasize that last expression, "communal effort." The attentive reader will have easily noticed in these pages that I have only mentioned a few physicists, no more than the tip of the iceberg. That is inevitable: the history of high-energy physics calls not for an entire book, but for several.

Of course, notwithstanding its success, the Standard model is obviously not the "final theory." On one hand because it leaves out gravitational interaction, on the other, because it includes too many parameters that have to be determined experimentally. Those are the always uncomfortable yet fundamental "why" questions. "Why do the fundamental particles we have detected exist? Why do those particles have the masses they have? Why, for example, does the tau weigh around 3,520 times as much as an electron? Why are there four fundamental interactions, instead of three, five, or just one? And why do those interactions have the properties they do (such as intensity or range of action)?"

A world of ultra-tiny strings?

Let us now consider gravitation, the other basic interaction. Can it be unified with the other three? A central problem is the lack of a quantum theory of gravitation that has been subjected to experimental

[22] Glashow (1960), Weinberg (1967), Salam (1968).

[23] To understand the idea of the quantum of an interaction it is enough to consider the case of electromagnetic radiation mentioned above. According to classic theory, it propagates in fields (waves), while quantum physics expresses that propagation in terms of corpuscles (photons), which are *quanta* of *h*· energy as proposed by Einstein in 1905.

testing. There are, however, candidates for this splendid unifying dream: complex mathematical structures called string theories.

According to string theory, basic particles existing in nature are actually one-dimensional filaments (extremely thin strings) in spaces with many more dimensions than the three spatial and single temporal one we are aware of. Although, rather than saying that they "are" or "consist of" such strings, we would have to say that they "are manifestations" of the vibrations of those strings. In other words, if our instruments were powerful enough, what we would see are not "points" with certain characteristics—what we call electrons, quarks, photons or neutrinos, for example—but tiny vibrating strings, with open or closed ends. The image this new view of matter calls to mind is thus more "musical" than "physical." In his best-seller, *The Elegant Universe* (2001, 166-168), Brian Greene, a physicist and outstanding member of the "string community" explains: "Just as the different vibratory patterns of a violin string generate different musical notes, *the different vibratory models of a fundamental string generate different masses and force charges…* The Universe—which is made up of an enormous number of these vibrating strings—is something similar to a cosmic symphony."

It is easy to understand how attractive these ideas can be: "Strings are truly fundamental; they are 'atoms,' that is, *indivisible components,* in the most authentic sense of that Greek word, just as it was used by the ancient Greeks. As absolutely minimum components of anything, they represent the end of the line—the last and smallest of the Russian 'matrioshka' nesting dolls—in the numerous layers of substructures within the microscopic world." (Greene 2001, 163). So what kind of materiality do these one-dimensional theoretical constructs have? Can we consider them a sort of "elemental matter" in a way similar to our customary concept of matter, including particles that are as elemental (though maybe only in appearance) as an electron, a muon or a quark?

I said before that string theories are complex mathematical structures, and that is certainly true. In fact, the mathematics of string theory are so complicated that, up to the present, no one even knows the equations of this theory's exact formulas—only approximations to those equations. And even those approximate equations are so complicated that, to date, they have only partially been solved. So it is no surprise that one of the great leaders in this field was a physicist with a special gift for mathematics. I am referring to the American, Edward Witten (b. 1951). The reader will get an idea of his stature as a mathematician when I mention that, in 1990, he received one of the four Fields medals (alongside Pierre-Louis Lions, Jean-Christophe Yoccoz and Shigefumi Mori) that are awarded every four years and are the mathematical equivalent of the Nobel Prize. In 1995, Witten launched "the second string revolution" when he argued that string (or super-string) theory could only become all-encompassing—a Theory of Everything—if it had ten spatial dimensions plus a temporal one. This eleven-dimensional theory, which Witten called M Theory, has yet to be completely developed.[24]

Faced with these string theories, it is reasonable to wonder whether we have reached a point in our exploration of the structure of matter in which "materiality"—that is, matter—disappears, becoming another thing altogether. But what is that other thing? If we are speaking about particles that appear as string vibrations, wouldn't that "other thing" actually be a mathematical structure? After all, a vibration is the oscillation of some sort of matter, but as a permanent structure, it is probably more of a mathematical than a material entity. If that were the case, we could say that one of Pythagoras' dreams had come true. Physicists would have been working very hard for centuries, or even millennia, only to discover that matter has finally slipped between their fingers, like a net, turning into mathematics, that is, mathematical structures. In sum, string theory unearths age-old problems, and maybe even ghosts: problems such as the relation between physics (and the world) and mathematics.

Independently of those essentially philosophical aspects of nature, there are others that must be mentioned here. Up to now, string theory has demonstrated very little, especially in light of the fact that science is not only theoretical explanation, but also experiments in which theory is subjected to the ultimate arbiter: experimental testing. String theories are admired by some, discussed by many, and criticized by quite a few, who insist that its nature is excessively speculative. Thus, the distinguished theoretical physician, Lee Smolin (2007, 17-18), pointed out in a book about these theories:

> In the last twenty years, a great deal of effort has gone into string theory, but we still do not know if it is certain or not. Even after all the work that has been done, the theory offers no prediction that can be tested through current experiments, or at least, experiments conceivable at the present time. The few clean predictions they propose have already been formulated by other accepted theories.
>
> Part of the reason why string theory makes no new predictions is that there seem to be an infinite number of versions. Even if we limit ourselves to theories that coincide with some of the basic facts observed in our universe, such as its vast size or the existence of dark energy, there continue to be something like 10^{500} different string theories; that is a one with five hundred zeros behind it, which is more than all the known atoms in the universe. Such a quantity

[24] There is no consensus about why the letter "M" was chosen. Some think it signifies Mother Theory, others, Mystery Theory, other Membrane Theory, and still others, Matrix Theory.

of theories offers little hope of identifying the result of any experiment that would not fit any of them. Thus, no matter what experiments show, it is not possible to demonstrate that string theory is false, although the opposite is equally true: no experiment can demonstrate that it is true.

In that sense, we should remember that one of the most influential methodologies in science continues to be the one put forth by Karl Popper (1902-1994), an Austrian philosopher who wound up at the London School of Economics. Popper always insisted that a theory that cannot be refuted by any imaginable experiment is not scientific. In other words, if it is not possible to imagine any experiment whose results contradict the predictions of a theory, then that theory is not truly scientific. In my opinion, that criterion is too strict to be invariably true, but it is certainly a good guide. At any rate, the future will have the final say about string theory.

Stellar Nucleosynthesis

Above, I dealt with the basic aspects of the structure of matter, but science is not limited to a search for the most fundamental, the smallest structure. It also seeks to understand what is closest to us and most familiar. In that sense, we must mention another of the great achievements of twentieth-century physics: the theoretical reconstruction of the processes —nucleosynthesis—that led to the formation of the atoms we find in nature, those of which we, ourselves, are made. These are questions addressed by nuclear physics, a field naturally related to high-energy physics —though the latter is more "fundamental," as it studies structures more basic than atomic nuclei.

In fact, high-energy physics supplies the basis for nuclear physics, which studies stellar nucleosynthesis. And it was the high-energy physicists who addressed the question of how the particles that constitute atoms emerged from the undifferentiated "soup" of radiation and energy that followed the Big Bang.[25]

As the universe cooled, the constituent parts of this soup underwent a process of differentiation. At a temperature of around 30,000 million degrees Kelvin (which was reached in approximately 0.11 seconds), photons—in other words, light—became independent of matter and were uniformly distributed through space. It was only when the temperature of the universe reached 3,000 degrees Kelvin (almost 14 seconds after the original explosion), that protons and neutrons began joining to form some stable nuclei, basically hydrogen (one proton around which one electron orbits) and helium (a nucleus of two protons and two neutrons with two electrons as "satellites"). Along with photons and neutrinos, those two elements, the lightest ones existing in nature, were the main products of the Big Bang, and they represent approximately 73% (hydrogen) and 25% (helium) of the universe's makeup.[26]

Consequently, we believe that the *Big Bang* generously supplied the universe with hydrogen and helium. But what about the other elements? After all, we know there are many more elements in nature. One does not have to be an expert to know of the existence of oxygen, iron, nitrogen, carbon, lead, sodium, zinc, gold and many other elements. How were they formed?

Even before high-energy physicists began studying primordial nucleosynthesis, there were nuclear physicists in the first half of the twentieth century who addressed the problem of the formation of elements beyond hydrogen and helium. Among them, we must mention Carl Friedrich von Weizsäcker (1912-2007) in Germany, and Hans Bethe (1906-2005) in the United States (Weizsäcker 1938; Bethe and Critchfield 1938; Bethe 1939a, b).[27] Almost at the very beginning of the second half of the twentieth century, George Gamow (1904-1968) and his collaborators, Ralph Alpher (1921-2007) and Robert Herman (1914-1997), took another important step (Alpher, Herman and Gamow 1948). They were followed two decades later by Robert Wagoner (b. 1938), William Fowler (1911-1995) and Fred Hoyle, who used a much more complete set of data on nuclear reactions to explain that in the Universe lithium constitutes a small fraction (10^{-8}) of the mass corresponding to hydrogen and helium, while the total of the remaining elements represents a mere 10^{-11} (Wagoner, Fowler and Hoyle 1967).[28]

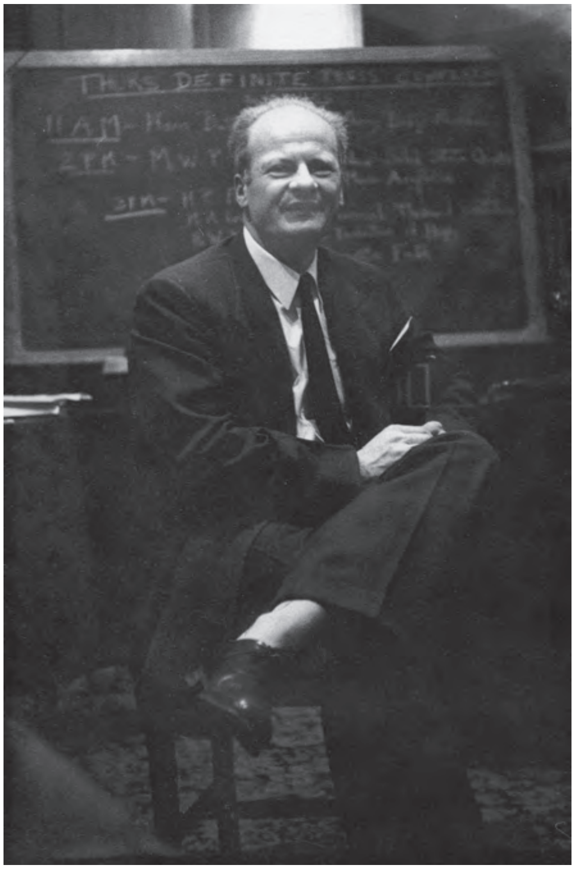
Thanks to their contributions—and those of many others—it has been possible to reconstruct the most important nuclear reactions in stellar nucleosynthesis. One of those reactions is the following: two helium nuclei collide and form an atom of beryllium, an element that occupies fourth place (atomic number) on the periodic table, following hydrogen, helium and lithium (its atomic weight is 9, compared to 1, for hydrogen, 4, for helium, and 6, for lithium). Actually, more than one type of beryllium was formed, and one of these was an isotope with an atomic weight of 8. It was very radioactive and lasted barely one ten-thousand-billionth of a second, after which it disintegrated, producing two helium nuclei again. But if, during that instant of life, the radioactive beryllium collided with a third helium nucleus, it could form a carbon nucleus (atomic number 6, atomic weight, 12), which is stable. And if the temperatures were high enough, then carbon nuclei would combine and disintegrate in very diverse ways, generating elements such as magnesium (atomic number 12), sodium (11), neon (10) and oxygen (8). In turn, two oxygen nuclei could join to generate sulphur and phosphorus. That is how increasingly heavy

[25] A magnificent and pioneering exposition is that of Weinberg (1979).

[26] I have not had occasion to mention that neutrinos, which were long thought to lack any mass (like photons), actually do have some. That is another of the important findings of physics from the second half of the twentieth century.

[27] Bethe received the Nobel Prize for Physics for this work in 1967.

[28] Fowler obtained the Nobel Prize for Physics for this work, which he shared with Chandrasekhar. Surprisingly, Hoyle, who initiated much of that work, was left out of the Swedish Academy's choice.



Hans Bethe (1957).

elements are made, up to, and including, iron (26).

Events like this raise another question: how did those elements reach the Earth, given that the place where they were made needed energy and temperatures unavailable on our planet? And if we suppose that there must not be too much difference between our planet and others —except for details such as their makeup and whether or not they have life— then, how did they arrive at any other planet? Some of the elements (up to iron) that were not produced during the universe's first instants were made primarily inside stars. They could then reach outer space in three different ways: through the lost of the mass in old stars in the so-called "giant" phase of stellar evolution; during the relatively frequent stellar explosions that astronomers call "novas;" and in the dramatic and spectacular explosions that take place in the final phase of a star's existence, called a "supernova" (one of these explosions was detected in 1987: the supernova SN1987A. It had actually occurred 170,000 years earlier, but it took the light that long to reach the Earth).

Supernova explosions are what most spread the heavy elements generated by stellar nucleosynthesis through space. It is not too clear why such explosions occur, but it is though that, besides expulsing elements that have accumulated inside them (except for a part that they retain, which turns into very peculiar objects, such as neutron stars); in the explosion itself, they synthesize elements even heavier than iron, such as copper, zinc, rubidium, silver, osmium, uranium, and so on, including the greater part of over a hundred elements that now make up the periodic table and are relatively abundant in star systems such as our Solar System.

It is precisely this abundance of heavy elements that makes it reasonable to assume that the Sun is a second-generation star, formed somewhat less than 5,000 million years ago by the condensation of residues of an earlier star that died in a supernova explosion. The material from such an explosion assembled in a disk of gas and dust with a proto-star in the center. The Sun "lit up" when the central nucleus was compressed so much that the hydrogen atoms melted into each other. The planets we now know as the Solar System—Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto (though the latter has recently lost its planetary status) with their satellites, such as the Moon—formed around the Sun, along elliptical bands, following a similar but gravitationally less intense process.

From that perspective, the Earth (formed around 4,500 million years ago), like the other planets, is something similar to a small cosmic junk heap (or cemetery); an accumulation of star remains not important enough to give life to a new star, that is, agglomerates of elements in such small quantities that they were not able to trigger internal thermonuclear reactions like those occurring in stars. But just as life finds its place in garbage dumps, so too, it found its place on our Earth, 12.700 kilometers in diameter and about $6{\cdot}10^{21}$ (6 followed by 21 zeros) tons in weight. We are both witnesses and proof of that phenomenon.

About 7,500 million years from now, the central zone of the Sun, where hydrogen turns into helium, will increase in size as the hydrogen is used up. And when that helium nucleus grows large enough, the Sun will expand, turning into what is called a red giant. It will become so huge that its diameter will reach the Earth's orbit, destroying the planet. But before that happens, the Earth's surface will have become so hot that lead melts, the oceans boil and all traces of life disappear. Thus, the very nuclear processes that gave us life will take it away.

Beyond the microscopic world

The physics theories discussed in previous sections are certainly quantum theories, but the world of quantum physics is not limited to them, and it would be a grave error not to mention other advances in this world during the second half of the twentieth century. Given the difficulty of deciding which of them is most important, I have chosen two groups. The first includes developments that have strengthened quantum physics in the face of criticism formulated by Einstein, Podolsky and Rosen, among others. The second has to do with work that has revealed the existence of quantum phenomena at a *macroscopic* scale.

A non-local theory: quantum entanglement

The goal of science is to provide theoretical systems that permit the relation of as many natural phenomena as possible, and that have a predictive capacity. That is what we call "explaining nature." Now, "to explain" does not mean finding familiar answers that do not contradict our most common explicatory categories: why should nature conform to such patterns? Above, I mentioned that some of quantum physics' most successful theories quite forcefully show that reality can be profoundly different than our intuition would seem to indicate. If this was already clear when quantum mechanics began in 1925-1926, it is even more so today. Let us consider this, now.

In 1935, Albert Einstein, along with two of his collaborators, Boris Podolsky (1896-1966) and Nathan Rosen (1910-1995), published an article (Einstein, Podolsky and Rosen 1935) arguing that quantum mechanics could not be a complete theory, that new variables had to be added. It would take a long time to explain their arguments, which extend beyond pure physics and enter clearly philosophical areas (they offered a definition of what "physical reality" is). What I can say is that their analysis led John Stewart Bell (1928-1990)—a physicist from Belfast working in CERN's theory division—to demonstrate the existence of a series of relations (inequalities) that could be used in experiments to determine which type of theory was correct. The candidates were, on one hand, a "complete" theory (which would include some "hidden" variables for quantum formulation) that would obey the requirements proposed by Einstein, Podolsky and Rosen in 1935, and on the other, traditional quantum mechanics (Bell 1964, 1966). On the basis of Bell's analysis, John Clauser, Michael Horne, Abner Shimony and Richard Holt (1969) proposed a concrete experiment through which Bell's inequality test could be applied. This experiment was carried out at the Institute of Theoretical and Applied Optics of Orsay, on the outskirts of Paris, by a team led by Alain Aspect (b. 1947). The result (Aspect, Dalibard and Roger 1982) supported quantum mechanics. It might be rare, counterintuitive, have variables that cannot be determined simultaneously, and undermine our traditional idea of what reality is, but it is true. Bell's analysis and the experiment by Aspect's team also brought out a trait of quantum mechanics that, while known, had gone practically unnoticed: its nonlocality. All of the elements of a quantum system are connected, *entangled.* It does not matter that they might be so distant from each other that transmitting a signal to one element about what has happened to another is not even possible at the speed of light, which is the maximum allowed by special relativity. In other words, an element "finds out," and reacts instantly to, what has happened to another, no matter how much distance separates them. Nonlocality—which Einstein always rejected as contrary to common-sense physics—unquestionably poses a problem of compatibility with special relativity, but there is no reason to think that we will be unable, at some future date, to find a generalization of quantum mechanics that solves it. Still, it is certainly not going to be easy.

Moreover, nonlocality offers possibilities that would seem to belong to the realm of science fiction. Science writer Amir Aczel (2004, 20) put it this way: "Through entanglement, the state of a particle can also be 'teleported' a great distance, as happened whenever captain Kirk of the *Star Trek* TV series asked to be beamed up to the *Enterprise.* To be precise, no one has yet been able to teleport a person, but the state of a quantum system has been teleported in a laboratory. And this incredible phenomenon is beginning to be used in cryptography and (could be used) in future quantum computing."

Ideas, and to some degree realities, such as these show that science can even surpass science fiction. At any rate, these consequences of quantum physics are

John Bardeen.

more a matter for the twenty-first century than for the one that recently ended.

Macroscopic quantum phenomena: The submicroscopic becomes macroscopic

We are accustomed to thinking that the domain of quantum physics is exclusively the ultramicroscopic, that of elemental particles, atoms and radiation. But such is not the case, even though historically those phenomena were responsible for the genesis of quantum theories. The two main manifestations of macroscopic quantum physics are Bose-Einstein condensation and superconductivity.

Bose-Einstein condensates

From a theoretical standpoint, Bose-Einstein condensates (or condensation) come from an article published by the Hindu physicist, Satyendranath Bose (1894-1974) in 1924. There, he introduced a new statistical method (a way of counting photons) to explain the law of black-body radiation that had led Max Planck to formulate the first notion of quantization in 1900. It was Einstein who recognized and helped publish Bose's work (1924), which he completed with two articles (Einstein 1924, 1925) in which he expanded Bose's conclusions. He pointed out, for example, that condensation could occur in photon gas: "One part 'condenses' and the rest continues to be a perfectly saturated gas" (Einstein 1925). With the term "condensation," Einstein meant that a group of photons acts like a unit, even though there do not appear to be any interactive forces among them. He also predicted that "if the temperature drops enough," the gas will experience "a brutal and accelerated drop in viscosity around a certain temperature." For liquid helium—where there were already indications of such superfluidity—he estimated this temperature to be around 2°K.

The next advance in Einstein's prediction of the existence of superfluidity did not arrive until 8 January 1938, when the English magazine, *Nature,* published two brief articles—one by Piotr Kapitza (1894-1984) and the other by Jack Allen (1908-2001) and Don Misener (1911-1996). Kapitza had been a senior professor at the Cavendish Laboratory in Cambridge until 1934, when he returned to Russia on vacation. Stalin refused to let him leave, and he became director of the Physics Problems Institute in Moscow. Allen and Misener were two young Canadian physicists working in Cambridge at the Mond Laboratory sponsored by the Royal Society. Those articles (Kapitza 1938; Allen and Misener 1938) announced that, below 2.18°K, liquid helium flowed with almost no viscosity-induced resistance. But the theoretical demonstration that this phenomenon constituted evidence of superfluidity came from Fritz London (1900-1954) and Laszlo Tisza (b. 1907).[29]

Of course, this was the old idea put forth by Einstein in 1924, which had drawn very little attention at the time. Now, it was more developed and had been applied to systems very different than the ideal gasses considered by the father of relativity.

It should be pointed out, however, that despite the importance we now give to those 1938 discoveries as macroscopic examples of quantum behavior, that aspect was less evident at the time. In order to better understand the relation between Bose-Einstein condensation and macroscopic aspects of quantum physics, it was necessary to deal with atoms, producing "superatoms," that is, groups of atoms that behave like a unit and are perceptible macroscopically. That achievement arrived much later, in 1995, when Eric Cornell (b. 1961) and Carl Wieman (b. 1951), two physicists in Colorado, produced a superatom of rubidium. A few months later, Wolfgang Ketterle (b. 1957) did the same with sodium at MIT (all three shared the Nobel Prize for Physics in 2001). This is how the first two described their work (Cornell and Wieman 2003, 82):

> In June 1995, our research group at the Joint Institute for Laboratory Astrophysics (JILA) in Boulder created a tiny, but marvellous drop. By cooling 2000 rubidium atoms to a temperature less than a hundred thousand-millionths of a degree above absolute zero (100 thousand-millionths of a degree Kelvin), we got those atoms to lose their individual identities and behave like a single "superatom." The physical properties of each one, their movements, for example, became identical. The Bose-Einstein condensate, the first to be observed in a gas, is materially analogous to a laser, except that, in a condensate, it is atoms, not photons, that dance in unison.[30]

Further on, they add (Cornell and Wiemann 2003, 82-84):

> We rarely see the effects of quantum mechanics reflected in the behavior of a macroscopic amount of matter. The incoherent contributions of the immense number of particles in any portion of matter obscure the wavelike nature of quantum mechanics; we can only infer its effects. But in a Bose condensate, the wavelike nature of every atom is in phase with the rest in a precise manner. Quantum-mechanical waves run through the entire sample and are plainly visible. The submicroscopic becomes macroscopic.
>
> The creation of Bose-Einstein condensates has shed light on old paradoxes of quantum mechanics. For example, if two or more atoms are in a single quantum-mechanical state, which is what happens with a condensate, it will be impossible to tell them apart, no matter how they are measured. The two atoms will occupy the same volume of space, move at the same speed, disperse light of the same color, and so on.
>
> In our experience, based on the constant treatment of matter at normal temperatures, nothing can help us understand this paradox. For one reason: at the normal temperatures and scales of magnitude in which we generally work, it is possible to describe the position and movement of each and every one

[29] London (1938), Tisza (1938).

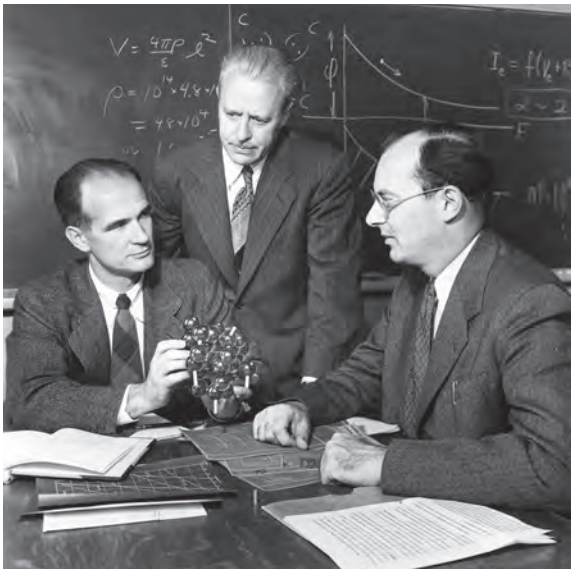
[30] The temperature called absolute zero (0°K) corresponds to -273.15°C. At that temperature, molecules do not move.

> of the objects in a group… At extremely low temperatures, or small scales of magnitude, classical mechanics no longer holds… We cannot know the exact position of each atom, and it is better to imagine them like imprecise stains. The stain is a package of waves, the region of space where one can expect that atom to be. As the group of atoms cools, the size of such wave packages increases. As long as each atom is spatially separate from the others, it will be possible, at least in principle, to tell them apart. But when the temperature gets low enough, the wave packages of neighbouring atoms overlap. Then, those atoms 'Bose-condense' in the lowest possible energy state and the wave packages merge to form a single macroscopic package. The atoms suffer a quantum identity crisis: we can no longer tell them apart.

Superconductivity

Superconductivity is another of the physical phenomena in which quantization appears on a macroscopic scale. The phenomenon itself was discovered long ago, in 1911, by Heike Kamerlingh Onnes (1852-1926), a Dutch physicist and the world's leading expert on low temperatures. In his Leiden laboratory, he discovered that cooling mercury to 4ºK entirely annulled its resistance to the passage of electric current (Kamerlingh Onnes 1911). Once the current began, it would continue indefinitely even if no power difference was applied. It was later discovered that other metals and compounds also became superconductors at temperatures near absolute zero. Of course experimental evidence is one thing and a theory capable of explaining it is quite another. It was not until 1957 that US scientists, John Bardeen (1908-1991), Leon Cooper (b. 1930) and John Robert Schrieffer (b. 1931) arrived at such a theory (known as the BCS theory, for the initials of their last names).[31]

Its explanation (Bardeen, Cooper and Schrieffer 1957) is that below a certain temperature the electrons that transport electric current in a superconductive element or compound form pairs that act as bosons; that is, particles like photons that are not subject to certain quantum requirements. Cooper (1956) had reached this supposition before, which is why they are now called "Cooper pairs." This grouping occurs at very low temperatures and is due to the interaction between electrons and the network of metal atoms in the superconductive compound. Once the pairs are formed, they march like a harmonious army of bosons, ignoring atomic impediments. That is how this quantum effect is manifested on a macroscopic scale.



The inventors of the transistor: W. Shockley, W. Brattain and J. Bardeen.

The BCS theory was a formidable success for quantum physics, but it is not totally satisfactory, as was revealed by its incapacity to predict the existence of superconductivity in ceramic materials at much higher temperatures than had previously been employed. It was in 1986, at the IBM laboratories in Zurich, that Georg Bednorz (b. 1950) and Alexander Müller (b. 1927) discovered that an oxide of lanthanum, barium and copper was superconductor at temperatures as high as 35ºK (which is certainly not high by everyday human standards, of course).[32] The following year, Paul Chu (1987) raised the scale of superconductor temperatures when he discovered an oxide of yttrium, barium and copper that became superconductor at 93ºK, a temperature that can be reached simply by bathing that oxide in liquid nitrogen—unlike helium, the latter is abundant and cheap. Since then, the number of such materials and the temperature at which they become superconductors has increased continually.

Bednorz and Müller's discovery (1986),[33] for which they received the Nobel Prize for Physics in 1987, offers new perspectives, not only for physics, but even more so, for technology. Materials that are superconductors at temperatures that can be achieved in everyday settings (that is, outside the laboratory) might revolutionize our lives some day.

Quantum devices: transistors, chips, masers and lasers

Our previous observation about the relevance of quantum physics to technology extends far beyond superconductivity. Superconductors may someday change our lives, but there is not doubt at all that other materials—semiconductors—have already done so.[34] The first major use of semiconductors arrived after John Bardeen, William Shockley (1910-1989) and Walter Brattain (1902-1987) invented the transistor while working in Bell Laboratories' department of solid-state physics.[35] In 1956, the three were awarded the Nobel Prize for Physics—the first of two for Bardeen (as we saw above, he received the second for superconductivity).

[31] The three shared the Nobel Prize for Physics in 1972.

[32] Until then, an alloy of niobium and germanium had the highest known temperature for superconductivity (23ºK).

[33] See also, Müller and Bednorz (1987).

[34] As its name indicates —although it is not an especially illustrative definition— a semiconductor is a material that conducts electricity to a degree that falls somewhere between the conductivity of a metal and that of an insulating material. The conductivity of semiconductors can normally be improved by adding small impurities, or through other factors. Silicon, for example, is a very poor conductor at low temperatures, but its conductivity increases with the application of heat, light or a difference of potential. That is why silicon is used in transistors, rectifiers and integrated circuits.

[35] See Shockley (1947, 1948) and Bardeen and Brattain (1948, 1949).

Aleksandr Prokhorov and Nikolai Basov showing Charles Townes (in the middle) their laboratory in Moscow (1965).

A transistor is an electronic device made from a semiconductor material that can regulate a current passing through it. It can also act as an amplifier or as a photoelectric cell. Compared to the vacuum tubes that preceded them, transistors need only tiny amounts of energy to function. They are also more stable and compact, work instantly, and last longer.

Transistors were followed by integrated circuits, tiny and very thin devices on which the digital world is based. Integrated circuits are made with a substrate (usually silicon), on which are deposited fine films of materials that alternately conduct or insulate electricity. Assembled according to patterns drawn up beforehand, these films act as transistors (each integrated circuit can hold millions of transistors) that function like switches, controlling the flow of electricity through the circuit, or *chip.*

As part of these chips, transistors carry out basic functions in the billions and billions of microprocessors installed to control car engines, cell phones, missiles, satellites, gas networks, microwave ovens, computers and compact disc players. They have literally changed the way we communicate with each other, relate to money, listen to music, watch television, drive cars, wash clothes and cook.

Until the advent of transistors and integrated circuits, calculating machines were gigantic masses of electronic components. During World War II, one of the first electronic calculators was built: the Electronic Numerical Integrator And Computer (ENIAC). It had 17,000 vacuum tubes linked by miles of cable. It weighted 30 tons and consumed 174 kilowatts of electricity. We can consider it the paradigm of the first generation of computers. The second generation arrived in the nineteen fifties, with the advent of transistors. The first computer to emerge from solid-state physics —a branch of quantum physics—was called TRADIC (*Transistor Digital Computer*). Bell Laboratories built it in 1954 for use by the United States Air Force. It used 700 transistors and was as fast as ENIAC. The third generation of computers arrived in the late nineteen sixties, with the advent of integrated circuits. It was followed by a fourth generation, which used microprocessors and refined programming languages. There is now talk of quantum computers. Rather than *bits,* which have defined values of 0 or 1, they will use *qubits,* that is, *quantum bits,* which can take values between 0 and 1, just as quantum states can be the superposition of photons with horizontal and vertical polarizations. But if quantum computers are ever successfully made, they will probably belong to the second half of the twenty-first century.

Thanks to all these advances, we are now immersed in a world full of computers that carry out all kinds of functions with extraordinary speed and dependability. Without them, our lives would be very different. And it is very important to emphasize that none of this would have happened without the results obtained in one branch of quantum physics: solid-state physics (also known as condensed-matter physics).

Another positive aspect of this branch of physics is the way in which it has generated closer relations between science and society. In 1955, for example, Shockley, one of the transistor's inventors, left Bell Laboratores to found his own company in the Bay Area of San Francisco. The Shockley Semiconductor Laboratory opened for business in February 1956 and recruited an excellent group of professionals. Though not especially successful, it was the seed that led to the development of numerous high-technology companies in a part of California that came to be called Silicon Valley.

Science and technology are allied in this techo-scientific world in such an intimate way—so to speak—that we cannot really say that fundamental innovation occurs only in scientific enclaves and business in technological ones. In that sense, let us recall that the fundamental techniques (the "planar" process) for manufacturing chips were conceived in 1957 by Jean Hoerni (1924-1997) at the Fairchild Semiconductors company. The first integrated circuit was built at the same place by Robert N. Noyce (927-1990) in 1958. Ten years later (1968), Noyce left Fairchild to found Intel along with Gordon Moore (b. 1929). There, he and Ted Hoff (b. 1937) directed the invention of the microprocessor, which launched a new revolution.

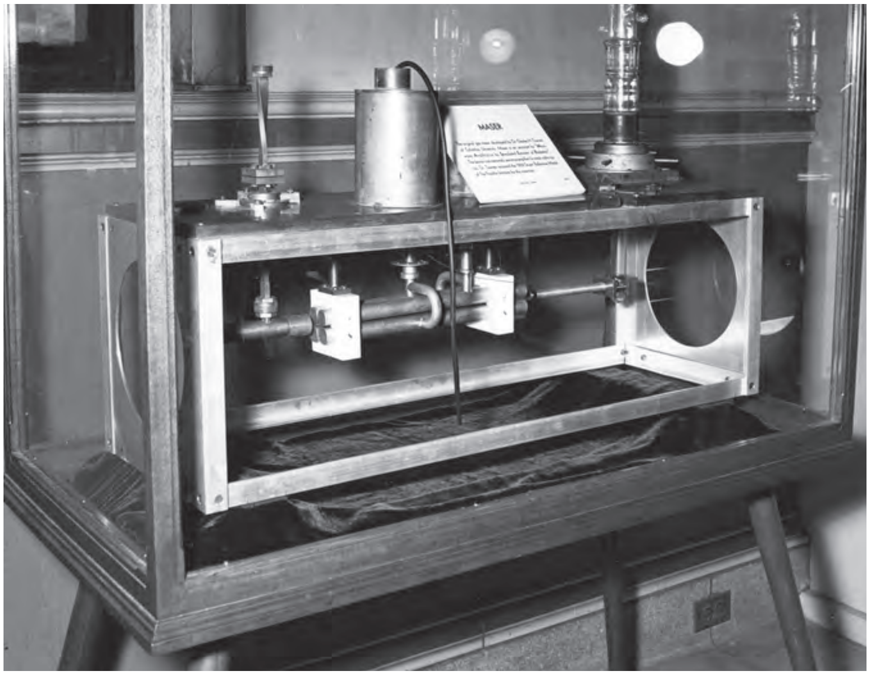
In that same sense, I should add that the development of electronic microprocessors has stimulated—and simultaneously benefited from—what is called "nanotechnology." The latter seeks to control and manipulate matter at a scale of between one and one-hundred nanometers (one nanometer equals 10^{-9} meters). Nanotechnology is more a technique (or group of techniques) than a science, but it can be expected to lead to developments (to a degree, it already is) that contribute not only to our material possibilities, but also to the most basic scientific knowledge.

Masers and lasers

I have yet to mention the maser and the laser although chronologically they are earlier than some of the advances mentioned above. Those terms are acronyms for **m**icrowave **a**mplification by **s**timulated **e**mission of **r**adiation and **l**ight **a**mplification by **s**timulated **e**mission of **r**adiation, respectively.

From a theoretical standpoint, these instruments or procedures for amplifying waves of the same frequency (wavelength) are explained in two articles by Einstein (1916a, b). Their practical development, however, with all the new theoretical and experimental elements involved, did not arrive until the nineteen fifties. This achievement was carried out, independently, by physicists from the Lebedev Physics Institute in Moscow—Aleksandr M. Prokhorov (1916-2002) and Nicolai G. Basov (1922-2001)—and the United States scientist, Charles Townes (b. 1915), at Columbia University in New York (the three shared the Nobel Prize for Physics in 1964).

In May 1952, at a conference on radio-spectroscopy at the USSR Academy of the Sciences, Basov and Prokhorov described the maser principle, although they did not publish anything until two years later (Basov and Prokhorov 1954). They not only described the principle; Basov even built one as part of his doctoral dissertation, just a few months after Townes had done so.

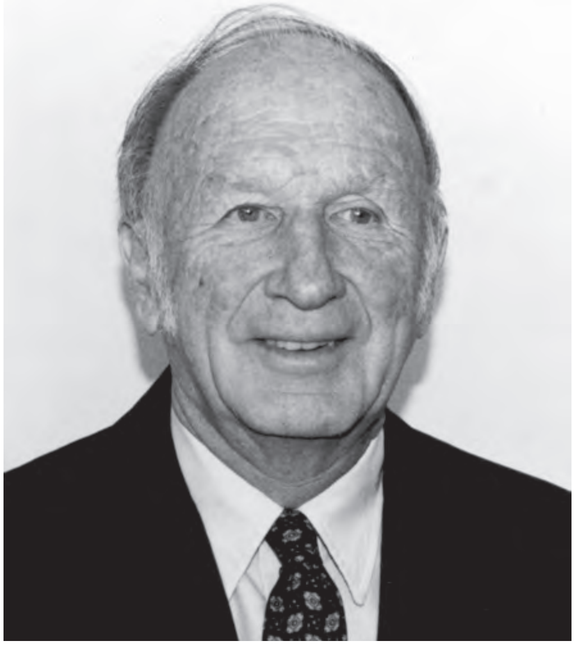


The first maser built by Townes and his collaborators, exhibited at the Franklin Institute (Philadelphia).

It is worth telling how Townes arrived independently at the same idea of a maser, as it shows how very diverse the elements making up a process of scientific discovery can actually be. After working at Bell Laboratories between 1939 and 1947, where he carried out research on radar, among other things, Townes moved to the Columbia University Radiation Laboratory, created during World War II to develop radars, instruments essential to the war effort. As with other institutions, this one continued to receive military funds after the war, and it dedicated 80% of its funding to the development of tubes able to generate microwaves. In the spring of 1950, Townes organized an advisory committee at Columbia to consider new ways of generating microwaves shorter than one centimeter for the Naval Research Office. After thinking about this question for a year, he was about to attend one of the committee sessions when he had an idea about a new way to approach it. That new idea was the maser. When, in 1954, Townes, a young doctor named Herbert J. Zeiger and a doctoral candidate named James P. Gordon managed to make the idea work, using a gas of ammonia molecules (Gordon, Zeiger and Townes 1954), it turned out that the oscillations produced by the maser were characterized not only by their high frequency and power, but also by their uniformity. In fact, the maser produced a coherent emission of microwaves; that is, highly concentrated microwaves with just one wavelength.

Even before the proliferation of masers, some physicists began attempting to apply that idea to other wavelengths. Among them were Townes himself (as well as Basov and Prokhorov), who began work in 1957 to move from microwaves to visible light. On this project, he collaborated with his brother-in-law, Arthur Schawlow (1921-1999), a physicist from Bell Laboratories. Together, they wrote a basic article explaining how a laser could be built, although they still called it an "optical maser" (Schawlow and Townes 1958). We might add that Bell Laboratories' lawyers thought that the idea of a laser was not sufficiently interesting to bother patenting it. They only did so at the insistence of the two scientists (Schawlow and Townes 1960).

From that moment, the race was on to build a laser. While later history has not always been sufficiently clear on this matter, the first successful one was built by Theodore Maiman (1927-2007) at Hughes Research Laboratories in Malibu, California. He managed to make a ruby laser function on 16 May 1960. Maiman sent a manuscript of his findings to the newly-established



Edward Lorenz.

magazine, *Physical Review Letters,* but its editor, Samuel Goudsmit, rejected it as "just another maser article." Maiman then turned to *Nature,* which published the results of his work on 6 August 1960 (Maiman 1960). Soon thereafter, Schawlow announced in *Physical Review Letters* that, along with five collaborators (Collins, Nelson, Schawlow, Bond, Garret and Kaiser 1960), he had gotten another laser to work. It was also a ruby laser, but considerably larger and more powerful than Maiman's. In light of all this, there is some question as to why it was Schawlow who received the Nobel Prize in 1981 (he shared it with Nicolaas Bloembergen and Kai Siegbahn), although, formally, it was for his and Bloembergen's contributions to laser spectroscopy.[36] Masers, and especially lasers (another "child" of quantum physics that makes quantum effects visible on a macroscopic scale), are instruments well known to the public, especially in certain applications (in detached retina operations, for example, which are carried out with lasers). But other uses of considerable scientific significance are not as well known. One of these is spectroscopy. The laser's high-energy monochromatic radiation makes it possible to precisely aim it at specific atomic levels; the results obtained offer considerable information on the properties of molecules, whose structure makes them much more difficult to study than atoms.

A non-linear world

The discoveries and developments discussed above are probably the most outstanding from, let us say, a fundamental perspective. But they do not include a group of advances that are opening new and surprising windows in science's understanding of nature. We are referring to non-linear phenomena; that is, those governed by laws involving equations with quadratic terms.[37]

Looking back at the history of physics, we can see that, until well into the twentieth century, most of the most basic theories were either essentially linear (Newton's theory of universal gravitation or Maxwell's electrodynamics, for example), or they could be used by non-linear systems, as occurs with Newtonian mechanics, but have been applied mainly to linear systems, even when it is absolutely clear that this implies a mere approximation of reality. The most straightforward example in this sense is the simple flat pendulum. Any high-school student, not to mention physics students, knows that the differential equation used to describe the movement of this type of pendulum is:

$$d^2\theta(t)/dt^2 + (g/l)\theta(t) = 0$$

where θ represents the angular movement of the pendulum, *l* his length, *g* the acceleration of gravity and *t*, time. Now, when we deduce (it is not a difficult problem) the equation that the motion of a simple flat pendulum should meet, it turns out that it is not the one shown above, but instead:

$$d^2\theta(t)/dt^2 + (g/l)\sin\theta(t) = 0$$

which is obviously not linear, since $\sin(\theta_1+\theta_2) \neq \sin\theta_1+\sin\theta_2$. In order to avoid this circumstance, which enormously complicates the problem's resolution, it is generally limited to small oscillations, that is, small angles, which make it possible to use Taylor's serial development of the sine function:

$$\sin\theta \approx \theta - \theta^3/6 + \ldots$$

keeping only the first term in order to obtain the first (linear) of the two equations shown above.

This very straightforward example shows us that so-called "classical physics," is not free of non-linear systems, but it tries to avoid them because of the mathematical difficulty they entail. In fact, there are no general systematic mathematical methods for dealing with non-linear equations. Of course many problems associated with non-linear systems (laws) have long been known, especially those from the field of hydrodynamics, the physics of fluids. Thus, for example, when water flows slowly through a tube, its movement (called *laminar*), is regular and predictable, but when the speed involved is greater, then the water's movement becomes *turbulent,* making whirlpools that follow irregular and apparently erratic trajectories that are typical characteristics of non-linear behavior. Aerodynamics is, of course, another

[36] Siegbahn received it for his contributions to the development of high-resolution electronic spectroscopy.

[37] Symbolically, it could be said that the expression of linearity is the equation, A + A = 2A, while in the world of non-linearity, the universe in which the meeting of two beings generates, *creates,* new properties, A + A ≠ 2A. In a rigorous sense, that is, a mathematical one, the essential difference between a linear system and a non-linear one is that, while two solutions of a linear system can be added to create a new solution to the initial system ("the superposition principle") that is not true in the case of non-linear systems.

example of non-linear domains, as everyone involved in aircraft design knows so well.[38]

The wealth of non-linear systems is extraordinary; especially the wealth and novelties they offer with respect to linear ones. From a mathematical perspective (which frequently correlates with real domains), non-linear equations/systems can describe transitions from regular to apparently arbitrary behavior; localized pulses that produce rapidly decaying perturbations in linear systems maintain their individuality in non-linear ones. That is, they lead to localized and highly coherent structures. This has obvious implications in the apparition and maintenance of structures related to life (from cells and multicellular organisms right up to, strange as it may sound, mental thoughts). One of the first known examples of this sort of behavior are the famous "solitons," solutions to non-linear equations in partial derivates called Korteweg-de Vries (or KdV equations), developed in 1895 as an approximate description of water waves moving through a narrow, shallow canal. But it was not until 1965 that Norman Zabusky and Martin Kruskal found a solution to this equation that represents one of the purest forms of coherent structures in motion (Zabusky and Kruskal 1965): the soliton, a solitary wave that moves with constant velocity. Far from being mathematical entelechies, solitons actually appear in nature: for example, in surface waves (that move essentially in the same direction) observed in the Andaman sea that separates the isles of Andaman and Nicobar in the Malaysian peninsula.

Chaos

An especially important case of non-linear systems is chaos systems. A system is characterized as chaotic when the solutions of equations that represent it are extremely sensitive to initial conditions. If those conditions change even slightly, the solution (the trajectory followed by the object described by the solution) will be radically modified, following a completely different path. This is the contrary of the non-chaotic systems that physics has offered us for centuries, in which small changes in the opening conditions do not substantially alter the solution. Extreme variability in the face of apparently insignificant changes in their starting points and conditions are what lead these systems to be called *chaotic.* But that does not mean that they are not subject to laws that can be expressed mathematically. We should emphasize that chaotic systems are described by laws codified as mathematical expressions, and these are actually similar to the ones that make up the universe of linear laws from Newton's dynamics.

Weather is one of the large-scale examples of chaotic systems; in fact, it was weather-research that revealed what chaos really is; small perturbations in the atmosphere can cause enormous climate changes. This was discovered by the United States theoretical meteorologist, Edward Norton Lorenz (1938-2008). In his weather research, he developed simple mathematical models and explored their properties with the help of computers. But, in 1960, he found that something strange occurred when he repeated previous calculations. Here is how he, himself, reconstructed the events and his reaction in the book, *The Essence of Chaos* (Lorenz 1995, 137-139), which he wrote years later:

> At one point, I decided to repeat some of the calculations in order to examine what was happening in greater detail. I stopped the computer, typed in a line of numbers that had come out of the printer a little earlier, and started it back up. I went to the lobby to have a cup of coffee and came back an hour later, during which time the computer had simulated about two months of weather. The numbers coming out of the printer had nothing to do with the previous ones. I immediately though one of the tubes had deteriorated, or that the computer had some other sort of breakdown, which was not infrequent, but before I called the technicians, I decided to find out where the problem was, knowing that that would speed up the repairs. Instead of a sudden interruption, I found that the new values repeated the previous ones at first, but soon began to differ by one or more units in the final decimal, then in the previous one, and then the one before that. In fact, the differences doubled in size more-or-less constantly every four days until any resemblance to the original figures disappeared at some point during the second month. That was enough for me to understand what was going on: the numbers I had typed into the computer were not exactly the original ones. They were rounded versions I had first given to the printer. The initial errors caused by rounding out the values were the cause: they constantly grew until they controlled the solution. Nowadays, we would call this chaos.

What Lorenz observed empirically with the help of his computer, is that there are systems that can exhibit unpredictable behavior (which does not mean "not subject to laws") in which small differences in a single variable have profound effects on the system's later history. Weather is such a chaotic system, which is why it is so hard to predict, so *unpredictable,* as we often put it. The article in which he presented his results (Lorenz 1963) is one of the great achievements of twentieth-century physics, although few non-meteorological scientists noticed it at the time. This was to change radically over the following decades. That change of attitude had much to do with a famous sentence that Lorenz included in a lecture he gave on December 1972 at a session of the annual meeting of the American Association for the advancement of Science: "a butterfly flapping its wings in Brazil can produce a tornado in Texas."[39]

It is becoming increasingly clear that chaotic phenomena are abundant in nature. We already see them at work in the fields of economics, aerodynamics, population biology (for example, in some "predator-prey"

[38] Of all the great theories of classical physics, the most intrinsically non-linear is the general theory of relativity (the field equations of this theory of gravitational interaction are non-linear).

[39] That lecture was not published in its time; it is included in Lorenz (1995, 185-188).

models), thermodynamics, chemistry and, of course, in the world of biomedicine (one example is certain heart problems). It seems that they can also show up in the apparently stable movements of the planets.

The consequences of the discovery of chaos—and, apparently, its ubiquity—for our view of the world are incalculable. The world is not how we thought it was, not only in the atomic domains described by quantum physics, but also in those ruled by the more "classic" Newtonian laws. They are Newtonian, of course, but unlike those used by the great Isaac Newton and all his followers, which were *linear*, these are *non-linear*. Nature is not linear, it is non-linear, but not all non-linear systems are chaotic, although the reverse is certainly true, for all chaotic systems are non-linear. Thus, the world is more complicated to explain and we cannot predict everything that is going to happen in the old Newtonian fashion. But why should nature be so "straightforward," anyway? What is marvelous is that we are able to discover such behavior and its underlying mathematical laws.

I could, and probably should have mentioned other developments that occurred or began in the second half of the twentieth century, including non-equilibrium thermodynamics, one of whose central elements are gradients or differences of magnitudes such as temperature or pressure. Their importance lies in the fact that those gradients are the true source of life, which has to struggle against nature's tendency to reduce gradients, that is, energy's tendency to dissipate according to the second law of thermodynamics (expressed by the much-used term, "entropy"). For living beings, thermodynamic equilibrium is equivalent to death, so understanding life necessarily requires understanding non-equilibrium thermodynamics, rather than just the equilibrium thermodynamics that predominated throughout most of the nineteenth and twentieth centuries. The complexity of life and other systems in nature is a natural result of the tendency to reduce gradients: wherever circumstances allow, cyclical organizations arise to dissipate entropy in the form of heat. It could even be argued—and this is a new, not especially Darwinian way of understanding evolution—that, inasmuch as access to gradients increases as perceptual capacities improve, then increasing intelligence is an evolutionary tendency that selectively favors prosperity by those who exploit dwindling resources without exhausting them. This branch of physics (and chemistry) experienced considerable growth during the second half of the twentieth century, making it a magnificent example of other advances in the field of physics that took place during that period, and possibly should have been addressed in the present text, even though they are "less fundamental" in some ways. But I have already written too much here, so it is time to stop.

Bibliography

Aczel, A. D. *Entrelazamiento*. Barcelona: Crítica, 2004 (original English edition from 2002).

Allen, J. and D. Misener. "Flow of liquid helium II". *Nature* 141 (1938): 75.

Alpher, R. A., R. C. Herman and G. Gamow. "Thermonuclear reactions in the expanding universe". *Physical Review Letters* 74 (1948): 1198-1199.

Aspect, A., J. Dalibard and G. Roger. "Experimental test of Bell's inequalities using time-varying analyzers". *Physical Review Letters* 49 (1982): 1804-1807.

Bardeen, J. and W. Brattain. "The transistor, a semiconductor triode". *Physical Review* 74 (1948) 230-231 (L).

—, "Physical principles involved in transistor action". *Physical Review* 75 (1949): 1208-1225.

—, L. N. Cooper and J. R. Schrieffer. "Microscopic theory of superconductivity". *Physical Review* 106 (1957): 162-164 (L).

Basov, N. G. and A. M. Prokhorov. "3-level gas oscillator". *Eksperim. i Teor. Fiz.* ***(JETP)*** 27 (1954): 431.

Bednorz, J. G. and K. A. Müller. "Possible high T_c superconductivity in the Ba-La-Cu-O system". *Zeitschrift für Physik B- Condensed Matter* 64 (1986): 189-193.

Bell, J. S. "On the Einstein-Podolsky-Rosen paradox". *Physics* 1 (1964): 195-200.

—, "On the problem of hidden variables in quantum mechanics". *Reviews of Modern Physics* 38 (1966): 447-452.

Bethe, H. "Energy production on stars". *Physical Review* 55 (1939a): 103.

—, "Energy production in stars", *Physical Review* 55 (1939b): 434-456.

—, and C. L. Critchfield. "The formation of deuterons by proton combination", *Physics Review* 54 (1938): 248-254.

Bondi, H. and Th. Gold. "The steady-state theory of the expanding universe". *Monthly Notices of the Royal Astronomical Society* 108 (1948): 252-270.

Born, M. "Zur Quantenmechanik der Stossvorgänge. (Vorläufige Mitteiling)". *Zeitschrift für Physik* 37 (1926): 863-867.

Bose, S. "Plancks Gesetz und Lichtquantenhypothese". *Zeitschrift für Physik* 26 (1924): 178-181.

Chandrasekhar, S. "The maximum mass of ideal white dwarfs". *Astrophysical Journal* 74 (1932): 81-82.

Chu, P. C. W. "Superconductivity above 90 K". *Proceedings of the National Academy of Sciences* 84 (1987): 4.681-4.682.

Clauser, J. F., M. A. Horne, A. Shimony and R. A. Holt. "Proposed experiment to test local hidden-variable theories". *Physical Review Letters* 23, (1969): 880-884.

Collins, R. J., D. F. Nelson, A. L. Schawlow, W. Bond, C. G. B. Garret and W. Kaiser. "Coherence, narrowing, directionality, and relaxation oscillations in the light emission from ruby". *Physical Review Letters* 5 (1960): 303-305.

Cooper, L. N. "Bound electron pairs in a degenerate Fermi gas". *Physical Review* 104 (1956): 1.189-1.190 (L).

Cornell, E. A. and C. E. Wiemann. "El condensado de Bose-Einstein". *Investigación y Ciencia*. Temas 21 (first trimester 2003): 82-87.

Einstein, A. "Zur Elektrodynamik bewegter Körper". *Annalen der Physik* 17 (1905a): 891-921.

—, "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt". *Annalen der Physik* 17 (1905b): 132-148.

—, "Die Feldgleichungen der Gravitation". *Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte* (1915): 844-847.

—, "Strahlungs-Emission und Absorption nach der Quantentheorie". *Deutsche Physikalische Gesellschaft. Verhandlungen* 18 (1916a): 318-323.

—, "Zur Quantentheorie der Strahlung". *Physikalische Gesellschaft Zürich. Mitteilungen* 18 (1916b): 47-62. Also published in *Physikalische Zeitschrift* 18 (1917): 121-128.

—, "Kosmologische Betrachtungen zur allegemeinen Relativitätstheorie", *Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte* (1917): 142-152.

—, "Quantentheorie des einatomigen idealen Gases". *Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte* (1924): 261-267.

—, "Quantentheorie des einatomigen idealen gases. 2. Abhandlung". *Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte* (1925): 3-14.

—, "Lens-like action of a star by the deviation of light in the gravitational field". *Science* 84 (1936): 506-507.

—, B. Podolsky and N. Rosen. "Can quantum. mechanical description of physical reality be considered complete?". *Physical Review* 47 (1935): 777-780.

Fermi, E. "Tentativo di una teoria dei raggi β". *Il Nuovo Cimento* 11 (1934a): 1-19.

—, "Versuch einer Theorie der β-Strahlen. I". *Zeitschrift für Physik* 88 (1934b): 161-177.

Feynman, R. P. "Space-time approach to quantum electrodynamics". *Physical Review* 76 (1949): 769-789.

—, and M. Gell-Mann. "Theory of the Fermi interaction". *Physical Review* 109 (1958): 193-198.

Fukuda, H., Y. Miyamoto and S. Tomonaga. "A self-consistent subtraction method in the quantum field theory. II". *Progress in Theoretical Physics* 4 (1939): 47-59.

Gell-Mann, M. "A schematic model of baryons and mesons". *Physic Letters* 8 (1964): 214-215.

—, *El quark y el jaguar*. Barcelona: Tusquets, 1995 (original English edition from 1994).

Georgi, H. and S. L. Glashow. "Unity of all elementary particle forces". *Physical Review Letters* (1974): 438.

Glashow, S. L. "Partial-symmetries of weak interactions". *Nuclear Physics* 22 (1960): 579-588.

Gold, Th. "Rotating neutron stars as the origin of the pulsating radio sources". *Nature* 218 (1968): 731-732.

Gordon, J., P. H., J. Zeiger and Ch. H. Townes. "Molecular microwave oscillator and new hyperfine structure in the microwave spectrum of NH_3". *Physical Review* 95 (1954): 282-284 (L).

Greene, B. *El universo elegante*. Barcelona: Crítica/Planeta, 2001 (original English edition from 1999).

Guth, A. H. "Inflationary universe: a possible solution to the horizon and flatness problem". *Physical Review* D 23 (1981): 347-356.

Hawking, S. W. "Occurrence of singularities in open universes". *Physical Review Letters* 15 (1965): 689

—, "Occurrence of singularities in cosmology". *Proceedings Royal Society* A 294 (1966a): 511-521.

—, "Occurrence of singularities in cosmology". *Proceedings Royal Society* A 295 (1966b): 490-493.

—, "Particle creation by black holes", in Isham, Penrose and Sciama, eds. (1975): 219-267.

—, and R. Penrose. "The singularities of gravitational collapse and cosmology". *Proceedings of the Royal Society of London* A 314 (1969): 529-548.

Heisenberg, W. "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen". *Zeitschrift für Physik* 33 (1925): 879-893.

—, "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik". *Zeitschrift für Physik* 43 (1927): 172-198.

Hewish, A., S. J. Bell, J. D. H. Pilkington, P. F. Scott and R. A. Collins. "Observation of a rapidly pulsating radio source". *Nature* 217 (1968): 709-713.

Hoddeson, L., L. Brown, M. Riordan and M. Dresden, eds. *The Rise of the Standard Model*. Cambridge: Cambridge University Press, 1997.

Hoyle, F. "A new model for the expanding universe". *Monthly Notices of the Royal Astronomical Society* 108 (1948): 372-382.

Hubble, E. "A relation between distance and radial velocity among extra-galactic nebulae". *Proceedings of the National Academy of Sciences of the U.S.A.* 15, (1929): 168-173.

—, and M. L. Humanson. "The velocity-distance relation among extra-galactic nebulae". *Astrophysical Journal* 74 (1931): 43-80.

Hulse, R. A. and J. H. Taylor. "Discovery of a pulsar in a binary system". *Astrophysical Journal* 195 (1975): L51-L53.

Isham, Ch. J., R. Penrose and D. W. Sciama, eds. *Quantum Gravity. An Oxford Symposium*. Oxford: Oxford University Press, 1975.

Kamerlingh Onnes, H. "Further experiments with liquid helium. C. On the change of electric resistance of pure metals at very low temperature. IV. The resistance of pure mercury at helium temperatures". *Communications from the Physical Laboratory at the University of Leiden #* 120a (1911): 3-15.

Kapitza, P. "Viscosity of liquid helium above the λ-point". *Nature* 141 (1938): 74.

Landau, L. "On the theory of stars". *Phys. Z. Sowjetunion* 1 (1932): 285-287.

Lemaître, G. "Un univers homogène de masse constante et de rayon croissant, rendant compte de la vitesse radiale des nébuleuses extra-galactiques". *Annales de la Société Scientifique de Bruxelles* 47 (1927): 49-59.

Linde, A. D. "A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problem". *Physics Letters* 108 D (1982): 383-393.

London, F. "The λ-point phenomenon of liquid helium and the Bose-Einstein degeneracy". *Nature* 141 (1938): 643-644.

Lorenz, E. N. "Deterministic non-periodic flows". *Journal of Atmospheric Science* 20 (1963): 130-141.

—, *La esencia del caos.* Madrid: Debate, 1995 (original English edition from 1993).

Maiman, Th. H. "Stimulated optical radiation in ruby". *Nature* 187 (1960): 493-494.

Mather, J. C. et al. "A preliminary measurement of the cosmic microwave background spectrum by the Cosmic Background Explorer (COBE) satellite". *Astrophysical Journal* 354 (1990): L-37-L40.

Mayor, M. and D. Queloz. "A Jupiter-mass companion to a solar-type star". *Nature* 378 (1995): 355-359.

Müller, K. A. and J. G. Bednorz. "The discovery of a class of high-temperature superconductivity". *Science* 237 (1987): 1133-1139.

Oppenheimer, J. R. and H. Snyder. "On continued gravitational contraction". *Physical Review* 56 (1939): 455-459.

—, and G. Volkov. "On massive neutron cores". *Physical Review* 55 (1939): 374-381.

Penrose, R. "Gravitational collapse and space-time singularities". *Physical Review Letters* 14 (1965): 57-59.

Penzias, A. A. and R. W. Wilson. "A measurement of excess antenna temperature at 4080 Mc/s". *Astrophysical Journal* 142 (1965): 414-419.

Perlmutter, S. et al. "Discovery of a supernova explosion at half the age of the universe and its cosmological implications". *Nature* 391 (1998): 51-54.

Planck, M. "Zur Theorie des Gesetzes der Energieverteilung im Normalspektrum". *Verhandlungen der Deutschen Physikalischen Gesellschaft* 2 (1900): 237-243.

Ryle, M. "Radio stars and their cosmological significance". *The Observatory* 75 (1955): 137-147.

Salam, A. "Weak and electromagnetic interactions", in Svartholm, ed. (1968).

Schawlow, A. and Ch. H. Townes. "Infrared and optical masers". *Physical Review* 112 (1958): 324-327.

—, "A medium in which a condition of population inversion exists". U.S. Patent 2.929.922, 22 March 1960.

Schödinger, E. "Quantisierung als Eigenwertproblem. (Erste Mitteilung)". *Annalen der Physik* 79 (1926): 361-376.

Schmidt, B. et al. "High-Z supernova search: Measuring cosmic deceleration and global curvature of the universe using type Ia supernova". *Astrophysical Journal* 507 (1998): 46-63.

Schweber, S. "A historical perspective on the rise of the standard model". In Hoddeson, Brown, Riordan and Dresden, eds. (1997): 645-684.

Schwinger, J. S. "On radiative corrections to electron scattering". *Physical Review* 75 (1949): 898-899 (L).

Shockley, W. "Density of surface states on silicon deduced from contact potential measurements". *Physical Review* 72 (1947): 345.

—, "Modulation of conductance of thin films of semiconductors by surface changes". *Physical Review* 74 (1948): 232-233.

Smolin, L. *Las dudas de la física en el siglo XXI*. Barcelona: Crítica, 2007 (original English edition from 2006).

Smoot, G. et al. "Structure of the COBE differential microwave radiometer first year maps". *Astrophysical Journal* 396 (1992): L1-L5.

—, and K. Davison. *Arrugas en el tiempo*. Barcelona: Círculo de Lectores, 1994 (original English edition from 1993).

Sudarshan, E., C. G. and R. E. Marshak. "The nature of the four-fermion interaction". *Padua Conference on Mesons and Recently Discovered Particles*. Padua, 1957: V14-24.

—, "Chirality invariance and the universal Fermi interaction". *Physical Review* 109 (1958): 1.860-1.862.

Svartholm, N., ed. *Elementary Particle Theory: Relativistic Groups and Analyticity*. Stockholm: Almqvist and Wilksell, 1968.

Taylor, J. H, L. A. Fowler and P. M. McCulloch. "Measurements of general relativistic effects in the binary pulsar PSR1913+16". *Nature* 277 (1979): 437-440.

Tisza, L. "Transport phenomena in helium II". *Nature* 141 (1938): 913.

Wagoner, R. V., W. A. Fowler and F. Hoyle. "On the synthesis of elements at very high temperatures". *Astrophysical Journal* 148 (1967): 3-49.

Walsh, D., R. F. Carswell and R. J. Weyman. "0957+561 {A}, {B}: twin quasistellar objects or gravitational lens?". *Nature* 279 (1979): 381.

Weinberg, S. "A model of leptons". *Physics Review Letters* 19 (1967): 1.264-1.266.

—, *The First Three Minutes: A Modern View of the Origin of the Universe*. New York: Basic Books, 1979.

Weizsäcker, C. F. von. "Über elementumwandlungen im inner der sterne, II". *Physikalische Zeitschrift* 39 (1938): 633-646.

Wheeler, J. A. and K. Ford. *Geons, Black Holes and Quantum Foam*. New York: Norton, 1998.

Witten, E. "String theory dynamics in various dimensions". *Nuclear Physics B* 443 (1995): 85-126.

Wolszczan, A. and D. Frail. "A planetary system around the millisecond pulsar PSR1257+12". *Nature* 355 (1992): 145-147.

Yoshimura, M. "Unified gauge theories and the baryon number of the universe". *Physical Review Letters* 41 (1978): 281-284.

Zabusky, N. J. and M. D. Kruskal. "Interaction of "solitons" in a collisionless plasma and the recurrence of initial states". *Physical Review Letters* 15 (1965): 240-243.

Zweig, G. "An SU(3) model for strong interaction symmetry and its breaking". *CERN Report* 8181/Th 401 (January 1964) and 8.

the art of the invisible: achievements, social benefits, and challenges of nanotechnology

SANDIP TIWARI ROBERT McGINN

The history of science and engineering as an important social force is relatively short. Most would date it to the Copernican revolution of the sixteenth century, i.e. for less than a quarter percent of the time we have existed on this planet. With the advent of the scientific process—using abstract agnostic tools of mathematics, questioning, postulating, theorizing, predicting, verifying, believing enough in theories to go ahead but doubting enough to notice errors and faults—came the modern approach to learning and invention. Overcoming dogmas, even in the face of contradicting observations, has always been a challenge to society and always will be; the comfort of "business as usual" can't be overstated. This holds true in scientific endeavor too. But the physical and life sciences, with engineering and medicine as their professional areas of practice, are among the few undertakings where revolutions can happen relatively easily. Einstein's theory of relativity—the "absoluteness" of the speed of light and gravity as a deformation in space-time; quantum mechanics as an entirely new mechanics to describe reality that is based on probabilistic approaches—indeed the philosophical understanding of reality as a result of observation; Gödel's theorem of the limits of provability within any axiomatic system; the genomic decoding of the basis of life and the understanding of metabolism, replication, and reproduction, these are ideas that were rapidly adopted in the technical community as they stood the test of the scientific approach.

The scientific pursuit of truths and the drive to apply truths know no national boundaries and adapt to contemporary conditions. Among the progenitors of the dawn of scientific civilization, Copernicus was an ecclesiastic and from Poland; Bruno, who paid with his life for standing up for his beliefs against the dogma, was primarily a theologian and from Italy; Tycho de Brahe, a court mathematician from Denmark; and Johannes Kepler of Germany and Galileo Galilei of Italy teachers. Not all pioneers of science were teachers, even though universities as institutions of learning[1] had existed for a long time. In the last century, Albert Einstein started as a patent clerk and Neils Bohr's contributions came from his nomadic style at his center at a state-supported home institution, not unlike Copernicus and Kepler. In current times, as the power and economic impact of science and engineering have grown dramatically, numerous research institutions have come into existence, started by scientists themselves,[2] and were either self-funded or funded by philanthropists and others who are a kind of modern royalty: venture capitalists and small company founders who have gained fortunes through the applications of science

[1] In the East, Nalanda University spanned the 5th to the 12th centuries AD. The oldest continuously operating university, the University of Al-Karaouine, Morocco, was founded in 859 AD, and the oldest Western university, the University of Bologna, in 1088 AD. The early universities arose from religious institutions and gained increasing independence as the power of the religious hierarchy declined.

[2] J. Craig Venter, who was an important participant in sequencing of human genes through the Celera Genomics company, now heads the J. Craig Venter Institute, a self-funded research entity, whose most recent success has been a significant step towards building an artificial cell. Leroy Hood, who was among the early pioneers in tools for molecular diagnostics, heads the Institute for Systems Biology, an independent institution. These approaches are not unlike those of Thomas Edison, Graham Bell, or Nikola Tesla at the turn of the nineteenth century.

and engineering. Universities, as in the past, play a part, but are not the sole institutional agents of progress. State-funded laboratories, independent laboratories, and industrial laboratories, particularly for biological sciences, are all involved in the discovery and applications enterprise.

A scientific revolution originates with unique individuals of incredible will and inner strength, people who create an immense centripetal force in the form of a central simple vision as a universal organizing principle.[3] Scientific progress, a period of consolidation, happens because of individuals who pursue many ends centrifugally, employing a variety of tricks to take advantage of the connections centered on the organizing principle in a world full of complexity. Scientific and engineering progress relies both on the central discovery and the ensuing complex assembly. Mendeleev's creation of the periodic table before atomic particles and atoms were known or observed, Darwin's evolutionary principle formulated without any molecular, genetic or organismic knowledge, and Heisenberg's creation of quantum mechanics overthrowing Newtonian determinism are all instances of overturning dogmas and creating new principles.

It is humbling to realize that chemistry, biology, and physics, as we know them and use them today in engineering and medicine, by and large, didn't exist just a century and a half ago. From the discovery and understanding of chemical elements quickly arose our ability to make ammonia, and from it agriculture-enhancing fertilizers that make possible existence of nearly seven billion humans on earth. Genetic interactions and evolutionary understanding of mutations are central approaches in fighting disease and healthier and longer life. Computing and communications, which depend on electronics, draw the principles of operation of their hardware from quantum mechanics and information theory. We are enormously fortunate to live in an age of discovery and the adventure of applying those great discoveries.

Centrifugal advances also depend on the availability of tools—instruments of observation and creation. The smaller the tool, greater the likelihood of it being personalized, individualized and humanized, i.e. made friendly for individual to use. Because of this quality, tools are used by many, thus stoking the creativity of a larger collection of scientists and engineers, in turn affecting a larger group of society. The water wheel evolved into the steam engine, later into the electric engine and the combustion engine. Each one came in many forms. The combustion engine drives the airplane, the car, and the scooter in various incarnations. The electric engine runs the train, the air-conditioner, and even the hard disk drive of the laptop computer. We know now that there exists an engine of molecular life—the ATP engine called ATP synthase. It converts chemical energy to mechanical motion within our bodies. Who knows what doors this discovery and its synthetic laboratory creations will open? But, the constant theme in this miniaturization process has been to find applications that are useful to us as humans. Hospital operation procedures have changed dramatically due to endoscopic tools; in most cases, hospital stays are now eliminated. The mobile phone and other communication instruments are everywhere, even in the poorest regions of the world, arguably bringing the greatest benefit there through easier and open information exchange. Personal software for writing, drawing, and visualizing abound on our small computers. In all of these, miniaturization and personalization have had a spectacular impact.

Technological progress of course also has its dark side, depending on the innovations themselves and on the ways they are diffused and used. Fertilizers, computers, and combustion engines all consume enormous amounts of energy,[4] are sources of pollution, and have imbalanced our world. The Industrial Revolution in Europe drastically reduced the average human life span. Much of the energy consumed in the world today took billions of years to accumulate on our planet, making possible seven billion humans instead of perhaps a billion, but in turn affecting global climate. Personalized tools have a major impact through a multiplicative effect. Cars are a good example, but so is the cell phone. Each new creation, and the new ways in which society as a whole and its individuals interact, creates a new divide between the haves and have-nots, between those who adapt and those who don't, those who learn to take advantage of the tools economically and socially and those who don't. So, while average wellbeing may rise, disparities also usually rise. Because technology often eases manual tasks, those on the bottom rung of the economic ladder potentially suffer the most from new technologies.

It is in this societal context that we now turn to the promise and challenges of a burgeoning new area of technical specialization: nanoscale science, engineering and technology—often shortened to "nanotechnology." Fundamentally, nanotechnology is a phenomenon of a size scale—a dimension. Like biology that encompasses a very large breadth of life science areas, nanotechnology has an impact encompassing the breadth of science, engineering, and technology that the nanoscale dimension affects. Perhaps some time in the future we may choose to call it "nanology"

[3] To quote Arthur Koestler, "The more original a discovery, the more obvious it seems afterwards."

[4] The fertilizer and information industries are each claimed to use nearly 10% of the world consumption of energy, and the combustion engine even more.

to reflect this breadth, rather than just referring to nanotechnology, nanoscience, and nanoengineering.

If we take any bulk material that we can see with our naked eyes—whether the material be hard or soft, inorganic or organic—and make it smaller, it still has the same properties. A large piece of diamond, iron, plastic, or rubber has the same properties as a small piece of diamond, iron, plastic, or rubber, and we use these materials and their reproducibility to great advantage where these properties are effective. Bridges can be large and small—carry a single vehicle lane across a stream or a massive train across an ocean channel. Plastic is used in car panels and in little watches. On the other hand, if we go to the extreme end of reducing a material's size, i.e. their atomic or molecular level, the smallest size at which we could identify them, their properties will be entirely different. An atom or a molecule has properties that arise from the quantum mechanical interactions that lead to their existence as stable units. Carbon, an atom, forms diamond, but it also forms graphite, and is also the major constituent of soot, the result of inefficient combustion. All of these bulk assemblies of carbon have different properties.

So, in going from the atomic scale to the bulk scale a large change takes place in the properties. This happens in the nanometer size region. The properties of materials and their physical and chemical interactions arise from the forces of nature—the atomic or molecular bond is a result of quantum-mechanics and electromagnetic forces. Through interactions of and between atoms, molecules, electrons (the carriers of charge), and photons (the carriers of light), physical, chemical, and biological processes undergo a dramatic transformation in properties at the nanoscale dimension. This dimension range bridges the atom and molecule at one end and bulk materials at the other. The reason for this is that the forces at the center of these interactions, the forces that result in the characteristic properties, are fundamentally nanoscale in nature.

And the property changes at the nanoscale are not simply small. They can be dramatically different—new properties emerge, ones to which we did not previously have access at either the macro or micro scales.

Quantum mechanical tunneling is one phenomenon that has been employed successfully in the past decade: in semiconductor memories that don't lose their data and have no moving parts, those used in the camera, in the cell phone, and in the thumb drive. In these devices, electrons are simply made to tunnel through an insulating region at low voltages. It happens because of the wave nature of the electron and the ability of the wave to penetrate small distances—nanoscale distances—in an insulator. Such fundamental properties as melting temperature, magnetization, charge capacity, etc., can all be altered without changing the chemical composition of the material because of this wavelike property and the interactions at the nanoscale. Because of this wave nature, interactions between electrons and light can also change at the nanoscale—a property that medieval glass blowers utilized dramatically in stained glass. Stained glass often uses nanoscale particles of gold or silver. These particles provide red, blue, green, brown, or other colors by enhancing the scattering of that color depending on the particle's size. The electrons in gold or silver nanoparticles interact with photons of light creating the color. Scientists describe this collective interaction between electron plasma and photons through a particle that they call a "plasmon." The glass blowers had unknowingly developed the technology for precipitating them controllably in that size. The light that is carried in optical fibers that makes fast data communication possible is generated using lasers that are very efficient light sources through enhanced electron-photon interactions arising at the nanoscale in artificially created quantum wells. Even the optical fiber uses light confinement at the nanoscale to move it with limited loss over long distances.

Chemical reactions result from interactions between atoms and molecules in their neutral, excited, or charged states. The reacting species need to come close together and have energetically favorable pathways for the reactions to be effective. Catalysis is central to providing this effectiveness: a catalyst, while chemically unchanged, provides a low-energy path to increasing reaction rates. It does this by mediating via a surface where molecules come together, attach, and react with each other in energetically favorable conditions, leaving the catalyst undisturbed at the end of the reaction. As one approaches smaller dimensions, the surface area to volume ratio increases—a nanoscale effect. A significant non-linear enhancement in this simple property makes catalysis enormously effective. The Haber-Bösch process for making ammonia, a key ingredient for making fertilizers, productively uses catalysis in a number of steps. Hydrogen is derived from methane in natural gas using nickel oxide. Ammonia is then formed from this hydrogen and nitrogen using iron, derived from magnetite, with an ultimate conversion efficiency of 98%, i.e. nearly perfectly.

Magnetite, a form of iron oxide, is a material whose nanoscale properties have been utilized in nature for orders of magnitude longer than the glass blowers. Magnetite is magnetic. As a collection of nanoscale crystals—arranged as chains, thereby allowing the collection to become a highly sensitive magnet—it endows organisms with the property of magnetotaxis, the ability to discriminate the magnetic field lines of the Earth. So, *Magneteospirillum magnetotacticum*, a bacterium found in ponds and first isolated in 1975, along with many others, is magnetotactic because at the small scale, in the form of a collection, deviations from Earth's field lines can be discriminated by the primitive organism. Many animals use this magnetic information for navigation, including pigeons, loggerhead turtles, and spiny lobsters. In the evolutionary process, nature evolved ways by which inorganic nanocrystals could be formed in largely organic systems, something we are still learning to do controllably in the laboratory. Another interesting nanoscale example from nature is the iridescent color of some butterflies and peacock feathers. These are nanoscale optical interference effects from the three-dimensional structures that nature builds, and that, in the laboratory, we have only recently taken the first steps in recreating. Biological phenomena tend to be immensely complex, resulting as they do from a combination of the randomness of events and a large number of interactions that happen between larger numbers of entities, under the influence of local forces. The phenomena are sensitive to initial conditions, small perturbations, have a large number of interacting components, and often a large number of pathways by which the system can evolve. If a human has insufficient energy input, i.e. hasn't eaten enough, the body knows how to slow down the metabolism. Unlike much of what we do in physics, chemistry, and engineering, this is immensely more complex, involving a variety of interactions at various scales. That simple and complex organisms have developed approaches to making nanoscale single crystal magnetic domains to achieve these properties is a tribute to nature's resourcefulness and biology's cleverness—characteristics that the human species discovers regularly.

The last few decades set the stage for the development of condensed matter science and engineering where small, personalized tools became pervasive, and where the ability to control and observe at the nanoscale became available to a large community. These tools allow us to assemble, manipulate, control, probe, image, and look at a myriad of properties at the nanoscale. Of the tools, the scanning tunneling microscope and the atomic force microprobe have garnered the most press. But, just as significant have been many of the fabrication tools that let us define, pattern, and connect at the nanoscale dimension: new techniques for visualization, tools that allow us to self-assemble monolayers on surfaces, tools that let us synthesize, and in general tools that let us do this reproducibly, cheaply, and quickly. We can now synthesize atom by atom, and we can also sculpt to get down to near the atomic level. We can probe phenomena that exist at this scale through a large toolset that gives us a variety of views. And because the properties change dramatically when one gets down to the smallest units, we can leverage those properties by utilizing assembling and sculpting techniques. This in turn has made it possible for a large community to reach down into the nanoscale world. The nanoscale is a dimension, not a discipline, and the properties at the nanoscale show up in and are connected to all the disciplines. The profound impact of this, through the open large-scale participation of the community and the breadth of disciplines, has been that a large new area of interesting and exciting work has grown at the interfaces. Engineering, physical, and life-sciences have commingled like never before. And this has led to immense progress and utility that could not have been foreseen even a decade ago.

A few examples of this breadth, at the heart of human existence, will illustrate this point. Let us look at some of the challenges facing the world. Major ones revolve around being sustainable—a large complex community of humans and millions of other species living sustainably, i.e. in equilibrium with each other and with the wider natural world. Energy, better health, equity and alleviation of poverty, education and conservation all immediately come to mind as unifying themes for sustainability. Some sustainability-related questions immediately arise: can we lower energy consumption—in transportation, lighting, food production, and other facets of living by recreating our environment (heating, cooling, and aesthetics) and communications (in information exchange, in computing, and all the mobile instruments)? Can we help with water problems by producing clean water, removing heavy metal impurities such as arsenic and reducing water use? Can we improve agriculture productivity by producing plants for healthier diets that are more disease resistant and that consume less energy and water? Can we provide more efficient carbon sequestration through physical and biological approaches? Can we improve management of forestry resources by using less paper and introducing better paper production techniques? Can we improve on health care by introducing cheaper and earlier diagnosis, detect contamination, cure diseases, improve on treatment or slow degenerative diseases, and attack the most pernicious of the diseases—malaria and cancer? Nanotechnology holds promise for addressing all of these sustainability-related issues.

Strength of materials and the properties of surfaces of materials are in use all around us. Polymers, whose widespread industrial-scale synthesis started in mid-twentieth century, became ubiquitous by the turn of the century. Some would argue that plastics were the backbone of China's industrial revolution and a key to the transformation of everyday life, from children's toys to the widespread use in home and office and in ubiquitous packaging. Plastics and polymers achieve their properties through surface interactions of chains of hydrocarbons. Both of these are affected by new nanotechnology inventions. Carbon nanotubes, based as they are on a strong carbon bond—a different configuration than that of diamond—provide strong intrinsic and surface interaction properties; they can withstand stronger forces than steel of similar dimensions.[5] Pull them into strands similar to the way polymers are employed, and one gets materials of great strength. Carbon nanotubes are now introduced into plastics to make them more resilient, for example in sports equipment such as tennis rackets and golf clubs. Composites, such as concrete, fiberglass, and Kevlar are combined materials that achieve strength through the surface interactions. Concrete can be made much lighter and still maintain its strength through use of cenospheres, the hollow alumina and silica structures akin to carbon buckyballs, found in the ash of coal power-plants. The strength of nanoscale material and the strong interface allows these composites to be stronger than was possible before.

The surface is critical to this property.

We mentioned catalysis, and its centrality to the ammonia production process, as one of the major breakthroughs at the turn of last century. Today, zeolites play a similar role. These are microporous solids that are efficient catalysts based on oxides of aluminum and silicon. Millions of tons of them help fracture petroleum into gasoline and numerous other hydrocarbons, making the impact of oil less environmentally destructive.

Advances in nanotechnology seem likely to lead to major gains in energy production, energy consumption, communication, and health promotion. Let us consider some notable developments in these areas.

Fuel cells, batteries, photo-electro energy and electro-photo energy conversion are examples where efficiency improvements connected to energy are happening rapidly through new materials, thin membranes, and efficient conversion processes. Light sources made out of semiconductors are highly efficient, factors of ten better than the incandescent light bulb, and are reliable and longer lasting. We see them today in traffic lights, but we will see them increasingly in general lighting as issues of cost and of satisfying human color spectrum preferences are resolved. Light sources are also being created from organic materials, though in this case, the challenge of achieving reliability is considerably higher. Photovoltaic generation is also benefiting from efficiencies realizable at the nanoscale. A new type of solar cell that is starting to make the transition from laboratory to manufacturing is the Grätzel cell. These cells use nanocrystalline titania, dies, and organic materials for electron transport to achieve a few percent of efficiency in energy conversion. Titania is a material found in paint, sandpaper and many other places where its strength is utilized. It also absorbs photons efficiently and hence is also employed in suntan lotions. The new photovoltaic structures use low energy processes in fabrication, unlike the more popular current silicon photovoltaics, making the cost and energy used in fabricating them small. The enhanced surface interactions can be used to reduce contamination. In heavily populated regions such as Asia's Gangetic plane, with the dropping of the water table many deeper wells now being used for hygienic drinking water are naturally contaminated by large concentrations of arsenic. Improved efficiency of electrochemical processes on the surface allows the arsenic to be efficiently scavenged with iron oxide nanoparticles.

Electronics, computing, and communications have benefited tremendously from the properties of the nanoscale, a scale where wave electron and the material interact in many ways to produce interesting properties. Consider, for example, the case of data storage. Every day humanity is creating more data than the total amount of data that was stored just twenty years ago. Non-volatile semiconductor storage is utilized in our cameras, phones, miniature music players, and for storing and exchanging information. These work because tunneling, a quantum mechanical phenomenon, takes place at the nanoscale. The large amounts of data that Google searches and that enterprises store away have become possible because magnetic disk drives store more in a smaller area, i.e. are more compact and also cost less. This becomes possible because the ability to sensitively read and write has improved by taking advantage of the electron spin and field interactions that occur at nanoscale. Our fast communications infrastructure is dependent on optical transmission. The small

[5] There is currently a communicable disease infecting the science community: over-exuberance in claims that border on incredulity. This disease that has always been around is particularly pernicious because the breadth of the disciplines makes it difficult for many to see through the wild claims. Perhaps some of this is a societal and ethical issue as much pressure is put on scientists to justify their research. There is also a school of thought that young people can be inspired to pursue science and engineering through excitement that mostly relies on over-exuberance—an approach that has Sbecome easier in these times of the Internet, short attention span, and the ease in creation of wild visual imagery through personalized software. The application of carbon nanotubes for space elevators is one such myth (see "The space elevator: going down?" in *Nature Online* published May 22, 2006, and available at http://www.nature.com/news/2006/060522/full/news060522-1.html). Similar claims abound regarding the use of molecules and other atomic-scale approaches in electronics.

laser diodes and amplifiers and optical fibers employ confinement of carriers and photons in small dimensions for large-scale improvement in efficiencies in signal generation and transmission. Smaller devices also consume less power, so energy consumption per device has also decreased over time. However, the personalization of small instruments (e.g. computers) has also meant that more people are using them. Hence absolute power numbers have not decreased.

This precision sensing and control applied widely in electronics has also been a major determinant of how nanotechnology is being applied in biosciences. One of the fundamental challenges in biosciences has been the detailed understanding of phenomena under the specific chemical and physical conditions that exist in real environments. When polymerase chain reaction (PCR) was invented, it provided a technique to amplify a piece of DNA of interest and thus have more copies at one's disposal for study and analysis. In a small tool, PCR made it possible to generate millions of copies of a desired strand of DNA, and use them for genetic manipulation. Similar gains were made by use of microarrays, monoclonal techniques, and use of fluorescent proteins. Much biological experimentation, however, continues to depend on statistical analysis of data where a large collection of such interactions is taking place, and one extracts from it possible models to describe the specificity. Physical sciences techniques tend to strip away most of the extraneous phenomena and simplify a system so that the phenomena and properties of interest can be studied rigorously. With the advent of many nanoscale techniques, techniques that get down the smallest scale, it becomes possible to start unraveling the secrets without having to resort to statistical analysis and we can now study all the possibilities comprehensively.

Doing so, however, requires ultra-sensitive sensors. Size control allows one to make a wide variety of ultrasensitive sensors. A fluorescent molecule can be replaced by a more robust optically active nanoparticle tuned to specific wavelength for response, and tied to a molecule whose chemistry is being studied. One can use the plasmonic (electron plasma—electromagnetic) interactions to achieve localization of heating through the local coupling of energy at nanoscale dimensions. Cantilevers can be reduced down in dimension to a point where single-atom weight sensitivity can be achieved through observation of frequency shifts. Nanotools can be made to isolate, control and grab, and build scaffolds to organize cells, grow cells, pattern cells, and probe cells in two-dimensional and three-dimensional assemblies. It is possible to use optical tweezers to grab nanoparticles, move them around, and if desired study the various possibilities of reactions with molecules that are tethered to them. So one can imagine putting nanoparticles and other machinery to observe and interact *inside* cells and in tissues and do real-time sensing and imaging and unravel the complex workings inside the cell itself. One can work with these tools under realistic conditions of interest because of the large improvements in sensitivity, imaging, and control that nanoscale has provided.

Scientists tend to overestimate what can be done over a short time horizon—about ten years—and to underestimate what may be possible over a long time horizon—fifty years. What is very interesting in nanotechnology is that, because of its foundation in the important nano-length scale, its reach across disciplines is extensive. Never in the past have researchers generated knowledge so widely applicable across technical disciplines. The last decade has been a good start. But, as the tools and understanding develop, many new uses will be opened up that will be made possible by the knowledge at the boundaries of disciplines. Progress should continue to accelerate in the physical science and engineering space where photovoltaics, lighting, energy-efficient computing, information storage and retrieval, and communications should all continue their forward march. It can reasonably be argued that chemistry and materials science have focused on nanoscale phenomena since their inception; after all catalysis or synthesis of molecules, preparation of composites, and hard coatings have been around forever and draw on nanoscale interactions. What is new is that sensitive tools give us the ability to understand these phenomena better. New techniques for synthesis—of membranes, nanocrystals, and new material forms—should help improve technology in major areas of societal needs, such as fuel cells, energy storage, and contamination removal. The use and development of nanotechnology tools are very much in their infancy for life sciences. For use by the life scientists, the tools need to become more user-friendly, a systems-design task, but one that could enable cheap and quick decoding of the genetics of a complex organism, and diagnosing and delivery of drugs locally through nanoscale encapsulated systems, so considerably advancing preventive medicine.

Before closing, we return to the discussion of the societal consciousness of science and engineering, and specifically nanotechnology. The social problems and issues we encounter are no different, arising as they often do from humans and institutions attempting to "succeed." In the life sciences, there has been a greater societal awareness because of, among other reasons, the influence of pharmaceuticals in contemporary life and because of the relatively easy way havoc happens: witness anthrax, Vioxx, smoking, and thalidomide, each contributing growing social awareness. Practitioners working in the physical science and engineering worlds also need to develop approaches so that the research and development process remains ethical and holds the greater good of society paramount.

Because it has this vast reach, particularly in health and environment, Nanoscale research needs to be conducted in accordance with sound ethical values and principles by means of ethically responsible practices.

Here are a few potential landmines.

Inexpensive information connectivity, together with vast informational storage, and the capability and inclination of individuals, groups, and states to snoop, is a potential nightmare that has intensified over the past several years in the Western and the Eastern world. Nanotechnology enhances this potent capability. How should society and research work out this dilemma?

Humanity is the first evolutionary creation capable of changing the path of the survival of the fittest. What is the relationship between humans and nature? Should we acknowledge and defer to the primacy of nature? Is it ethically permissible or responsible to modify, even to entirely reconstruct natural organisms? When we replace or augment living parts, where is the boundary between the human and the machine? The time is likely not far off when neural sensors will uncover the workings of human emotions, personality, and perhaps even consciousness. In courts in India, functional magnetic resonance imaging is already being accepted as evidence of truthfulness and lying. Using neural actuators, in laboratory experiments monkeys have been electronically guided to peel and eat bananas. The Joy-Dyson[6] debates centered around the primal fear of such potent new technology arriving prior to societal readiness to manage it safely and equitably.

Should work in these directions connected to nanotechnology be halted or hamstrung because of controversial ethical issues or scenarios of possible disaster, as some have suggested? It is our view that the issues should be clearly identified and carefully considered. This should be done concurrently with continuing research and development imbued with responsible practices. Equally important is providing opportunities for practicing and future workers in the area, who are currently students, to look reflectively at their work in the context of the society in which it unfolds and which it shapes.

The importance of safety in handling and using nanomaterials, given the possibilities of health and safety hazard due to the enhanced nanoscale reactive properties, comes through in a recent survey[7] of nanotechnology researchers. Safe practices are also a matter of laboratory and community cultures. Considerations of time, money, status, and competition can press researchers and managers to cut corners. Historically, in most areas, governments have usually done the minimum amount necessary until pressured by those who have been or may be affected. Regulation has been a trailing edge in the safe operation of coal mines, textile factories, tire production plants, asbestos, glycol and other chemicals in the semiconductor industry, and lead in paints and gasoline. We are still debating the possible role in the increased incidence of brain cancer in cell phone users due to electromagnetic interactions nearly a decade after their widespread adoption in the developed world. Many in authority still do not recognize the role of humans and green house emissions in global warming. While nanotechnology is likely to play an important role in pollution prevention, e.g. by facilitating the removal of arsenic removal and cleaning of water, nanomaterials can also potentially introduce pollution. Silver is used as an antibacterial agent in the health industry, is employed in band-aids, and is being increasingly used as a microbe-killing agent. How much of this material is entering the water system as a result of washing?[8] Given the large amounts of money being invested in development, pressures for quick environmental regulatory approval without sufficient scientific check will be intense. While these risks are the result of shortcomings of societal procedures and processes, not in nanotechnology *per se*, nanotechnology researchers should bear them in mind.

For it is a fundamental ethical responsibility of scientists and engineers to attempt to prevent harm while carrying out their professional endeavors. Beyond promoting laboratory safety, preserving data integrity, recognizing contributions through due credit, and respecting intellectual property rights, does the researcher have any responsibility vis-à-vis the social consequences of research? The atomic bombing of Hiroshima and Nagasaki during the World War II set in motion a long period of introspection and public involvement from scientists in the societal debate;[9] how does one control a genie let out of the bottle? In the traditional view, society at large, not the individual researcher, is ethically responsible for what is done with the generated knowledge. However, the individual knowledgeable researcher also bears

[6] See F. Dyson, "The Future Needs Us!" in the *New York Review of Books*, vol. 50, no. 2, February 13, 2003. Emergent behavior, i.e. unpredictable behavior, appears in complex systems, i.e. those with a large number of interacting elements. Crowd behavior is an example of this. It is a very appropriate subject for thoughtful discussion and debate for man-made creations that can take a life of their own. However, one example of this that drew inspiration from nanotechnology and attracted a lot of popular attention—Michael Crichton's book *"Prey"*—is founded on faulty science. Particularly powerful is the description of swarms of nanorobots that take over bodies and the environment, that can fly and swim rapidly like insects and other similar, larger living objects. This is not possible because the viscous drag on the increased surface area slows nanoscale objects. It is like humans trying to swim in molasses.

[7] R. McGinn, "Ethics and Nanotechnology: Views of Nanotechnology Researchers." *Nanoethics*, vol. II, no. 2, 2008.

[8] See http://pubs.acs.org/cen/news/86/i15/8615news1.html. The impact of nitrogen and phosphorus as runoffs from large-scale use of fertilizers can be seen in most of the Western world. In the East, the poorer parts suffer from the disappearance of the water table (and backfilling by salt water near coastlines), and deeper wells that reach into arsenic contaminated tables, such as in West Bengal in India and in Bangladesh. This massive water depletion happened as the two-hundred-year-old invention of the diesel and electric motor became a personalized tool in the third world.

[9] Science and scientists rarely serve as inspiration for art. "Doctor Atomic," a popular opera that premiered in 2005, drew thoughts from the Manhattan project, in which the Atomic bomb first became reality. The nation's leading scientists of the day and Robert Oppenheimer, their leader, debated the bomb, even while they frantically worked on the weapon that Oppenheimer, the central character of "Doctor Atomic," quoting Bhagwat Geeta, later

termed “I am become Death, the shatterer of worlds.” It is interesting to note that the firebombings of WWII killed many more innocent people. In later overt and covert wars, Agent Orange, depleted uranium, cluster bombs, and land mines have left a large thumbprint in time without similar societal reaction, probably because such perniciousness is spread over a longer time.

some responsibility. Researchers cannot always plead ignorance of the risks posed by the powerful ‘engines’ they create. Contemporary researchers develop and facilitate the diffusion of their creations in societies whose character they know. While it is not always foreseeable that particular fruits of research will be turned into ethically troubling applications, at times this is the case (e.g. when there are substantial military or economic advantages to be gained even if ethically problematic effects are foreseeable as a byproduct). Hence, if researchers have reason to believe that their work, or work in their field, will be applied such as to create a risk of significant harm to humans, they have an ethical responsibility to alert the appropriate authorities or the public about the potential danger.

These preceding examples and brief discussion of responsibility point to the difficulties that arise when rapid scientific advances with major societal implications unfold rapidly and the society has to find the balance between fostering productive research and development activity and sustaining an effective regulatory and safety framework. One response to this challenge in recent years is the increased societal pressure for the contemporary researcher to acquire a hybrid competence: technical virtuosity wed to a sensitive ethical compass. In the words of Samuel Johnson, “integrity without knowledge is weak and useless, and knowledge without integrity dangerous and dreadful.”

For practicing scientists and engineers, one of the pleasures of their discipline is that great science leaves us unalone—content and emotionally happy to have found a piece of truth that one can call one’s own. Creative engineering gives the pleasure of coupling scientific discovery to the joy of having brought into the world a creation for the common good. At their best, these enterprises embody the ideals of a civilized life: the search for truth and the practice of good citizenship. Nanotechnology is in this classical tradition; it is here, it is growing vigorously, and, with prudent stewardship, will move human society forward in innumerable welcome ways.

the trajectory of digital computing

PAUL E. CERUZZI

No word has been overused when discussing computers as much as the word “revolution.” If one is to believe the daily press and television accounts, each new model of a chip, each new piece of software, each new advance in social networking, each new model of a portable phone or other portable device, will bring about revolutionary changes in our lives. A few weeks week later the subject of those reports is strangely forgotten, having been replaced by some new development, which we are assured, this time, is the real turning point.

Yet there is no question that the effect of computing technology on the daily lives of ordinary people has been revolutionary. A simple measure of the computing abilities of these machines, as measured by metrics such as the amount of data it can store and retrieve from its internal memory, reveals a rate of advance not matched by any other technologies, ancient or modern. One need not resort to the specialized vocabulary of the computer engineer or programmer: the sheer numbers of computers and digital devices installed in homes and offices or carried by consumers worldwide shows a similar rate of growth, and it is not slowing down. An even more significant metric looks at what these machines do. Modern commercial air travel, tax collection, medical administration and research, military planning and operations—these and a host of other activities bear an indelible stamp of computer support, without which they would either look quite different or not be performed at all.

An attempt to chronicle the history of computing in the past few decades faces the difficulty of writing amidst this rapid evolution. A genuine history of computing must acknowledge its historical roots at the foundations of civilization—which has been defined in part by the ability of people to manipulate and store symbolic information. But a history must also chronicle the rapid advances in computing and its rapid spread into society since 1945. That is not easy to do while maintaining a historical perspective. This essay identifies and briefly describes the essential persons, machines, institutions, and concepts that make up the computer revolution as it is known today. It begins with the abacus—first not only alphabetically but also chronologically one of the first computing instruments to appear. It carries events up to the twenty-first century, when networking of personal computing machines became commonplace, and when computing power spread to portable and embedded miniature devices.

Digital devices continue to evolve as rapidly as ever. But the personal computer in some ways has reached a plateau. The physical features of these machines have stabilized: a keyboard (descended from the venerable typewriter of the 1890s); a rectangular box containing the electronic circuits and disk storage; above that a display terminal (descended from the venerable television screen of the late 1940s). The electronic circuits inside, though more capable every year, have also stabilized: for the last 35 years they have consisted of integrated circuits made of silicon and encased in black plastic packages, mounted on plastic boards. Portable or "laptop" computers collapse this configuration but are essentially the same. Engineers and customers alike agree that this physical design has many drawbacks—consider for example the injuries to the muscles of the hand caused by over-use of a keyboard designed a century ago. But the many attempts to place the equivalent power, versatility, and ease of use onto other platforms, especially the portable phone, have not yet succeeded.

The programs that these computers run—the "software"—are still evolving rapidly. The things that these computers are connected to—the libraries of data and world-wide communications networks—are also evolving rapidly. It is not possible to anticipate where all that will lead. In the intervening time between the writing and publication of this essay, it is possible that the nature of computing will be transformed so much as to render parts of this study obsolete. Silicon Valley engineers talk of events happening in "Internet time": about six years faster than they happen elsewhere. Even after stripping away some of the advertising hyperbole, that observation seems to be true.

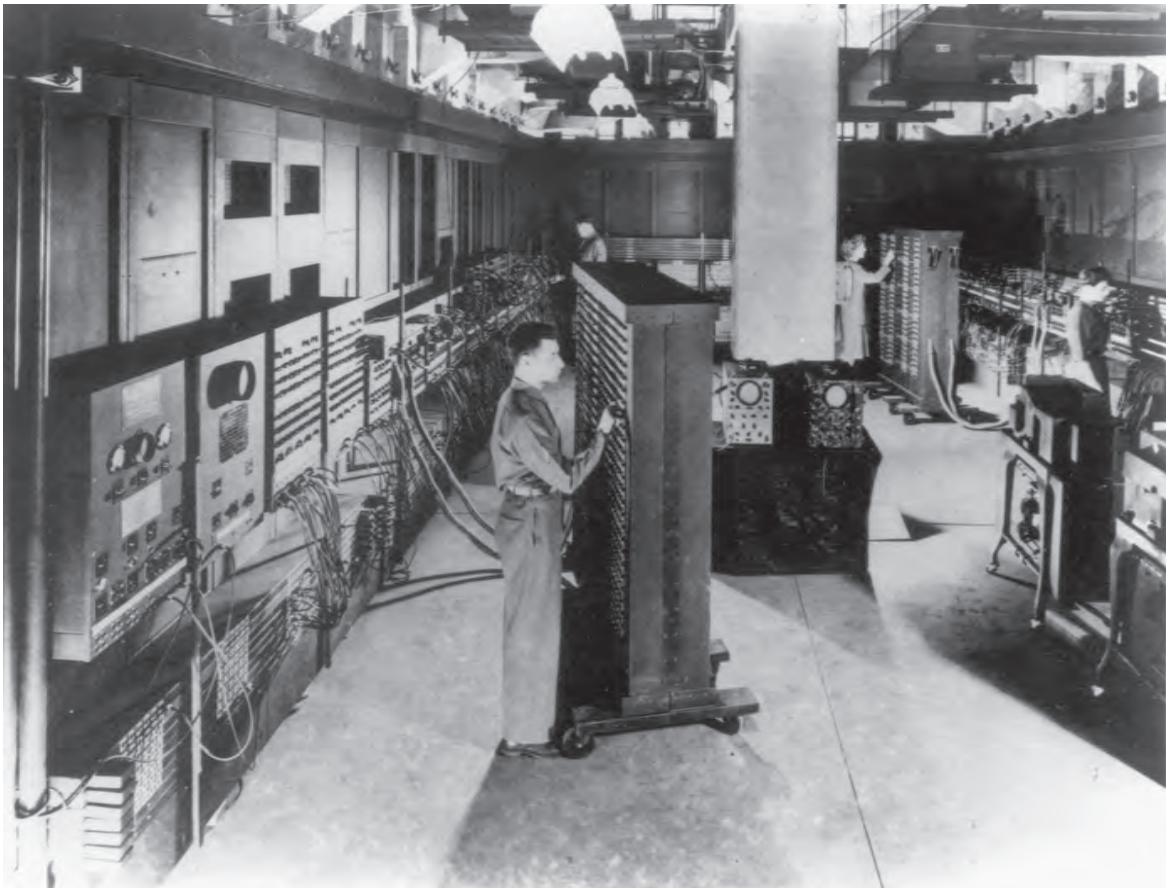
There are at least four places where one could argue the story of computing begins. The first is the obvious choice: in antiquity, where nascent civilizations developed aids to counting and figuring such as pebbles (Latin *calculi*, from which comes the modern term "calculate"), counting boards, and the abacus—all of which have survived into the twentieth century (Aspray 1990).

But these devices were not computers as we normally think of that term. To the citizen of the modern age, computing machinery implies a device or assembly of devices that takes over the drudgery of calculation and its sister activity, the storage and retrieval of data. Thus the second place to start the story: the 1890s, when Herman Hollerith developed the punched card and a system of machines that summed, counted, and sorted data coded into those cards for the US Census. The Hollerith system came along at a critical time in history: when power machinery, symbolized by the steam engine and by steam or water-powered factories, had transformed production. That linking of energy to production created a demand to control it—not only physical control but also the management of the data that industrialization brought with it. Hollerith's tabulator (and the company he founded, which formed the basis for the IBM Corporation) was but one of many such responses: others included electric accounting machines, cash registers, mechanical adding machines, automatic switching and control mechanisms for railroads, telephone and telegraph exchanges, and information systems for international commodity and stock exchanges.

But, the modern reader protests, that does not sound like the right place to start either. The real revolution in computing seems to have something to do with electronics—if not the silicon chips that are ubiquitous today, then at least with their immediate ancestors the transistor and vacuum tube. By that measure the computer age began in February, 1946, when the US Army publicly unveiled the "ENIAC"—"Electronic Numerical Integrator and Computer," at a ceremony at the Moore School of Electrical Engineering in Philadelphia. With its 18,000 vacuum tubes, the ENIAC was touted as being able to calculate the trajectory of a shell fired from a cannon faster than the shell itself traveled. That was a well-chosen example, as such calculations were the reason the Army spent over a half-million dollars (equivalent to several million in current dollars) for an admittedly risky and unproven technique.

Another early machine that calculated with vacuum tubes was the British "Colossus," of which several copies were built and installed at Bletchley Park in England during World War II, and used with great success to break German codes. These machines did not perform ordinary arithmetic as the ENIAC did, but they did carry out logical operations at high speeds, and at least some of them were in operation several years before the ENIAC's dedication. Both the ENIAC and Colossus were preceded by an experimental device built at Iowa State University by a physics professor named John V. Atanasoff, assisted by Clifford Berry. This machine, too, calculated with vacuum tubes, but although its major components were shown to work by 1942, it was never able to achieve operational status (Burks and Burks 1988).

Once again, the reader objects: is it not critical that this technology not simply exists but also is prevalent on the desks and in the homes of ordinary people? After all, not many people—perhaps a few dozen at most—ever had a chance to use the ENIAC and exploit its extraordinary powers. The same was true of the Colossus computers, which were dismantled after the War ended. By that measure the "real" beginning of



The ENIAC, 1945, at the University of Pennsylvania. Smithsonian Institution.

the computer revolution would not be in 1946 but in 1977, when two young men, Steve Jobs and Steve Wozniak, from an area now known as Silicon Valley, unveiled a computer called the "Apple II" to the world. The Apple II (as well as its immediate predecessor the "Altair" and its successor the IBM PC) brought computing out of a specialized niche of big businesses or the military and into the rest of the world.

One may continue this argument indefinitely. Young people today consider the beginning of the computer revolution even more recently, i.e., when the Internet first allowed computers in one location to exchange data with computers elsewhere. The most famous of these networks was built by the United States Defense Department's Advanced Research Projects Agency (ARPA), which had a network (ARPANET) underway beginning in 1969. But there were others, too, which linked personal and mini-computers. When these merged in the 1980s, the modern Internet was born (Abbate 1999).

Actually there are many places to begin this story. As this is being written, computing is going through yet a new transformation, namely the merging of the personal computer and portable communications devices. As before, it is accompanied by the descriptions in the popular press of its "revolutionary" impact. Obviously the telephone has a long an interesting history, but somehow that story does not seem to be relevant here. Only one thing is certain: we have not seen the last of this phenomenon. There will be more such developments in the future, all unpredictable, all touted as the "ultimate" flowering of the computer revolution, all relegating the events of previous revolutions to obscurity.

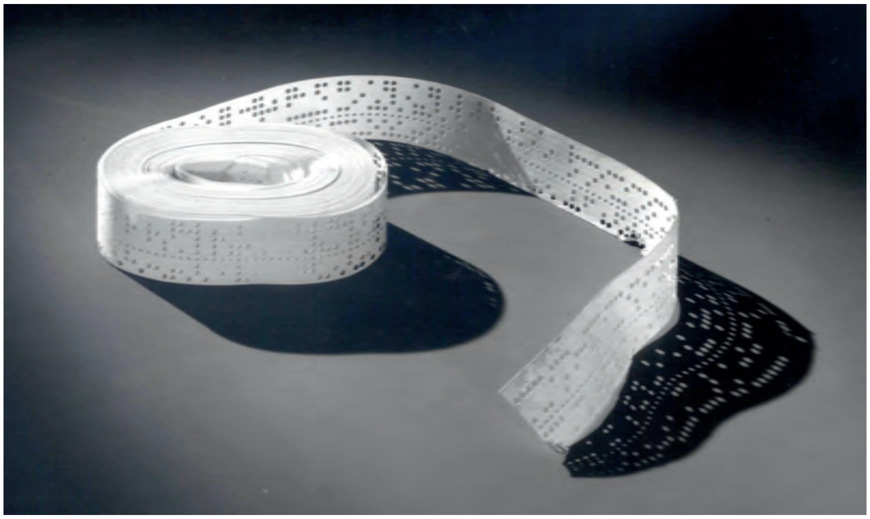
This narrative begins in the 1940s. The transition from mechanical to electronic computing was indeed significant, and that transition laid a foundation for the phenomena such as personal computing that followed. More than that happened in those years: it was during the 1940s when the concept of "programming" (later extended to the concept of "software") emerged as an activity separate from the design of computing machinery, yet critically important to that machinery's use in doing what it was built to do. Finally, it was during this time, as a result of experience with the first experimental but operational large computing machines, that a basic functional design of computing machines emerged—

an "architecture," to use a later term. That has persisted through successive waves of technological advances to the present day.

Therefore, in spite of all the qualifications one must put on it to make it acceptable to academic historians, one may argue that the ENIAC was the pivot of the computer revolution (Stern 1981). That machine, conceived and built at the University of Pennsylvania during the World War II, inaugurated the "computer age." As long as one understands that any selection is somewhat arbitrary and as long as one gives proper credit to earlier developments, including the work of Babbage and Hollerith, as well as the invention of the adding machine, cash register, and other similar devices, no harm is done.

Introduction

An ability to count and to represent quantities in some kind of symbolic notation was common to nearly all cultures, however "primitive" they may have appeared to modern scholars. Physical evidence of that ability is much more difficult to obtain, unless a durable medium such as clay tablets was used. We know that the concept of representing and manipulating quantitative information symbolically by pebbles, beads, knots on a string, or the like arose independently throughout the ancient world. For example, Spanish explorers to the New World found the Inca Indians using a sophisticated system of knotted strings called *quipu*, while similar systems of knotted strings are mentioned in the Bible, and at least one—the rosary—survives to the present day. A highly abstracted version of representation by beads evolved into the abacus, of which at least three different forms survive in modern China, Japan, and Russia. In the hands of a skilled operator an abacus is a powerful, compact, and versatile calculating tool. Other related aids to calculating were also in use in Western countries by the Middle Ages. These included counting boards with grids or patterns laid on them to facilitate addition (from this comes the modern phrase "over the counter" trading), and tokens used on these boards (these survive as gambling chips used in casinos).



BASIC paper tape, 1975. Almost as soon as he heard of the Altair, Bill Gates dropped out of college, and with his high school friend Paul Allen, moved to Albuquerque. They produced this version of the BASIC programming language for the Altair, which was critical in making that computer a practical device. Smithsonian Institution.

It is important to recognize that these devices were used only by those whose position in government, the Church, or business required it. With that qualification one could say these were in "common" use, but not in the sense of being ubiquitous. This qualification applies to all computing machines. The adoption of such machines depends on how costly they are, of course, but also crucially on whether they meet the needs of people. As Western society industrialized and became more complex those needs increased, but it is worth noting that even in spite of the steep drop in prices for computers and for Internet access, they have not achieved total penetration into the consumer market and probably never will.

Before moving on to calculating machinery it is worth noting one other aid to calculation that was in wide use and that survives in a vestigial form into the modern age. That is the printed table, which listed values of a mathematical function, for example. These can be traced back as far as the ancient Greeks, and they were extensively used by astronomers for their own use and, more importantly, for use by sailors on the open seas. Statistical tables, such as mortality rates for example, were developed for the insurance industry. Pocket calculators and "spreadsheet" computer programs allow one to compute these values on the spot, but tables still have their place. There are still a few places where one finds such tables in use. The continued use of tables shows their intimate connection with one of the fundamental uses of modern electronic computers (Kidwell and Ceruzzi 1994).

Most of the above devices worked in tandem with the Hindu-Arabic system of notation, in which a symbol's value depends not just on the symbol itself (e.g., 1, 2, 3...) but also on its place (with the all-important zero used as a place holder). This notation was vastly superior to additive notations like Roman numerals, and its adoption by Europeans in the late Middle Ages was a significant milestone on the road to modern calculation. When performing addition, if the sum of digits on one column was greater than nine, one had to "carry" a digit to the next column to the left. Mechanizing this process was a significant step from the aids to calculation mentioned above to automatic calculation. A sketch and a fragmentary description contained in a letter to Johannes Kepler indicate that Professor Wilhelm Schickard of the German town of Tuebingen built such a device in the early 1600s. No pieces of it are known to survive.

In 1642 the French philosopher and mathematician Blaise Pascal invented an adding machine that has the honor of being the oldest known to have survived. Digits were entered into the calculator by turning a set of wheels, one for each column. As the wheels passed through the value of “9,” a tooth on a gear advanced the adjacent wheel by one unit. Pascal took care to ensure that the extreme case, of adding a “1” to a sequence of “9s,” would not jam the mechanism. Pascal’s machine inspired a few others to build similar devices, but none was a commercial success. The reasons for that have become familiar: on the one hand it was somewhat fragile and delicate and therefore expensive, on the other hand the world in which Pascal lived was not one that perceived such machines to be a necessity of life.

About thirty years later the German philosopher and mathematician Gottfried Wilhelm Leibniz, satirized in Voltaire’s *Candide* and famous as the cocreator of the Calculus, learned of Pascal’s invention and attempted to construct a calculator independently. He succeeded in building a machine that not only could add but also multiply, using a gear that engages a variable number of teeth depending on where the operator had set a dial. His calculator did not work well, but the “stepped-drum” became the basis for nearly all multiplying calculators until the late nineteenth century. One modern descendant, the Curta, was small enough to fit in a pocket and was produced and sold into the 1970s.

The onset of a more mercantile society with a growing middle class made conditions more favorable for commercial success. Around 1820, Charles Xavier Thomas, a pioneer in establishing the insurance industry in France, built and marketed his “Arithmometer,” which used the Leibniz stepped drum to perform multiplication. Sales were poor at first, but it became quite popular after 1870, selling about one hundred a year. By then industrialization was in full swing, and Thomas’s machine was joined by a number of rivals to meet the demand (Eames and Eames 1990).

These demands were met on both sides of the Atlantic. Two “adding machines” developed in the United States were especially significant. Neither was capable of multiplication, but ability to do rapid addition, their ease of use, modest (though not low) cost, and rugged construction more than compensated for that deficiency. In the mid-1880s Dorr E. Felt designed and patented an adding machine that was operated by pressing a set of number keys, one bank of digits for each place in a number. What was more, the force of pressing the keys also powered the mechanism, so the operator did not have to pause and turn a crank, pull a lever, or do anything else. In the hands of a skilled operator, who neither took her fingers away from nor even looked at the keyboard, the Felt “Comptometer” could add extremely quickly and accurately. Selling for around US$125, Comptometers soon became a standard feature in the American office of the new century. At around the same time, William Seward Burroughs developed an adding machine that printed results on a strip of paper, instead of displaying the sum in a window. His invention was the beginning of the Burroughs Adding Machine Company, which made a successful transition to electronic computers in the 1950s and after a merger with Sperry 1980s has been known as the Unisys Corporation.

In Europe calculating machines also became a standard office product, although they took a different tack. The Swedish engineer W. Odhner invented a compact and rugged machine that could multiply as well as add, using a different sort of gear from Leibnitz’s (numbers were set by levers rather than by pressing keys). That led to a successful product marketed under the Odhner, Brunsviga, and other names.

No discussion of computing machinery is complete without mention of Charles Babbage, the Englishman who many credit as the one who first proposed building an automatic, programmable computer—the famous “Analytical Engine.” He came to these ideas after designing and partially completing a more modest “Difference Engine,” which itself represented a great advance in the state of calculating technology of the day. Details of Babbage’s work will be given later, but he did in fact propose, beginning in the 1830s, a machine that had all the basic functional components of a modern computer: an arithmetic unit he called the “Mill,” a memory device he called the “Store,” a means of programming the machine by punched cards, and a means of either printing the results or punching answers onto new sets of cards. It was to have been built of metal and powered by a steam engine. Babbage spent many years attempting to bring this concept to fruition, but at his death in 1871 only fragments had been built.

How different the world might have looked had he completed his machine makes for entertaining speculation. Would we have had an Information Age powered by steam? But once again, as with Pascal and Leibniz, one must keep in mind that the world was not necessarily waiting for a computer to be invented. To have made a real impact, Babbage would not only have had to surmount the technical obstacles that dogged his Analytical Engine, he would also have had to exert considerable powers of salesmanship to convince people that his invention was of much use. Evidence for that view comes from the fact that the Swedes Georg and his son Edvard Scheutz completed a working Difference Engine in 1853, which is regarded as the world's first successful printing calculator ever sold (Merzbach 1977). One of the machines was sold to the Dudley Observatory in Albany, New York, but the Scheutz Engine had little impact on science or commerce. The Information Age had to wait.

By the end of the nineteenth century the state of the art of calculating had stabilized. In the commercial world the simple Comptometer or Odhner had taken its place alongside other office equipment of similar scope, like the typewriter or telegraph ticker. In the world of science—still a small world in those years—there was some interest but not enough to support the construction of more than an occasional, special-purpose machine now and then. Those sciences that required reckoning, such as astronomy, made do with printed tables and with human "computers" (that was their job title) who worked with pencil, paper, books of tables, and perhaps an adding machine. A similar situation prevailed in the engineering professions: books of tables, supplemented by an occasional special-purpose machine designed to solve a special problem (e.g., the Tide Predictor, the Bush Differential Analyzer). After about 1900, the individual engineer might also rely on simple analog devices like the planimeter and above all the slide rule: an instrument of limited accuracy but versatile and sufficient for most of an engineer's needs.

Herman Hollerith's system of punched cards began as such a special-purpose system. In 1889 he responded to a call from the Superintendent of the US Census, who was finding it increasingly difficult to produce census reports in a timely fashion. The punched card and its accompanying method of coding data by patterns of holes on that card, and of sorting and counting totals and subtotals, fit the Bureau's needs well. What happened next was due as much to Hollerith's initiative as anything else. Having invented this system he was impatient with having a sole customer that used it only once a decade, and so embarked on a campaign to convince others of its utility. He founded a company, which in 1911 merged with two others to form the Computing-Tabulating-Recording Corporation. In 1924, upon the accession of Thomas Watson to the leadership position of C-T-R, the name was changed to International Business Machines. Watson was a salesman who understood that these devices had to meet customer's needs in order to thrive. Meanwhile the Census Bureau, not wishing to rely excessively on one supplier, fostered the growth of a rival, Remington Rand, which became IBM's chief rival in such equipment for the next half-century.

The ascendancy of punched card equipment looks in hindsight to have been foreordained by fate: its ability to sort, collate, and tabulate large amounts of data dovetailed perfectly with the growing demands for sales, marketing, and manufacturing data coming from a booming industrial economy. Fate of course was there, but one must credit Hollerith for his vision and Watson for his tireless promotion of the technology. When the US economy faltered in the 1930s, IBM machines remained as popular as ever: satisfying American and foreign government agencies' appetites for statistical data. Watson, the quintessential salesman, furthermore promoted and generously funded ways of applying his company's products to education and science. In return, some scientists found that IBM equipment, with minor modifications, could be put to use solving scientific problems. For astronomers like L. J. Comrie, punched card equipment became in effect a practical realization of Babbage's failed dream. Other scientists, including the above-mentioned Atanasoff, were beginning to propose special-purpose calculators that could execute a sequence of operations, as the never-completed Babbage Analytical Engine was to do. These scientists did so against a background of IBM tabulators and mechanical calculators that came close to meeting the scientists' needs without the trouble of developing a new type of machine (Eckert 1940).

Looking back on that era one sees a remarkable congruence between the designs for these programmable calculators and that of the never-completed Analytical engine. But only Howard Aiken, a professor at Harvard University, knew of Charles Babbage beforehand, and even Aiken did not adopt Babbage's design for his own computer at Harvard. Babbage was not entirely unknown in the 1930s, but most historical accounts of him described his work as a failure, his Engines as follies. That was hardly a story to inspire a younger generation of inventors. Those who succeeded where Babbage had failed, however, all shared his passion and single-minded dedication to realize in gears and wire the concept of automatic computing. They also had a good measure of Thomas Watson's salesmanship in them.

First among these equals was Konrad Zuse, who while still an engineering student in Berlin in the mid-

1930s sketched out an automatic machine because, he said, he was "too lazy" to do the calculations necessary for his studies. Laziness as well as necessity is a parent of invention. As the Nazis plunged the world into war, Zuse worked by day at an aircraft plant in Berlin; at night he built experimental machines in his parents' apartment. His "Z3" was running in December 1941; it used surplus telephone relays for calculation and storage, discarded movie film punched with holes for programming (Ceruzzi 1983).

In 1937 Howard Aiken, while working on a thesis in physics at Harvard, proposed building what eventually became known as the "Automatic Sequence Controlled Calculator." His choice of words was deliberate and reflected his understanding that the punched card machine's inability to perform sequences of operations limited its use for science. Aiken enlisted the help of IBM, which built the machine and moved it to Harvard. There, in the midst of World War II, in 1944, it was publicly dedicated. The ASCC thus has the distinction of being the first to bring the notion of automatic calculation to the public's consciousness. (German spies also brought this news to Zuse, but by 1944 Zuse was well along with the construction of a machine the equal of Aiken's.) The ASCC, or Harvard Mark I as it is usually called, used modified IBM equipment in its registers, but it could be programmed by a paper tape.

In 1937 George Stibitz, a research mathematician at Bell Telephone Laboratories in New York, built a primitive circuit that added number together using binary arithmetic—a number system highly unfriendly to human beings but well-suited to electrical devices. Two years later he was able to persuade his employer to build a sophisticated calculator out of relays that worked with so-called "complex" numbers, which arose frequently in the analysis of telephone circuits. The Complex Number Computer was not programmable, but during the World War II it led to other models built at Bell Labs that were. These culminated in several large, general-purpose relay computers. They had the ability not only to execute any sequence of arithmetic operations but also to modify their course of action based on the results of a previous calculation. This latter feature, along with electronic speeds (discussed next) is usually considered to be a crucial distinction between what we know today as "computers" and their less-capable ancestors the "calculators." (In 1943 Stibitz was the first to use the word "digital" to describe machines that calculate with discrete numbers.)

Rounding out this survey of machines was the Differential Analyzer, built by MIT Professor Vannevar Bush in the mid-1930s. This machine did not calculate "digitally" to use the modern phrase, but worked on a principle similar to the "analog" watt-hour meter found at a typical home. In others respects the Bush Analyzer was similar to the other machines discussed above. Like the other pioneers, Bush had a specific problem to solve: analyzing networks of alternating current power generators and transmission lines. The Differential Analyzer was a complex assembly of calculating units that could be reconfigured to solve a range of problems. The demands of the World War II led to a number of these machines being built and applied to other, more urgent problems. One, installed at the Moore School of Electrical Engineering in Philadelphia, was an inspiration for the ENIAC.

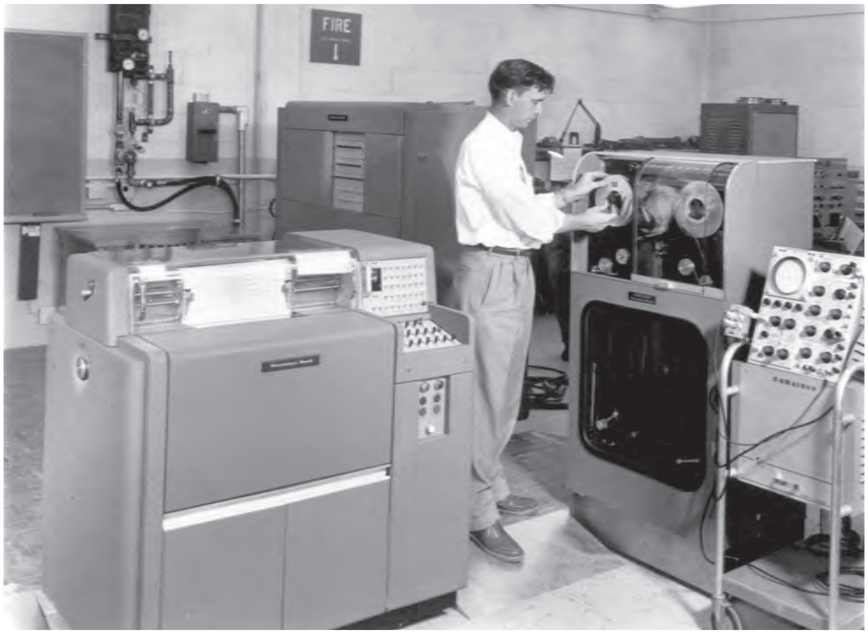
All of these machines used either mechanical gears, wheels, levers or relays for their computing elements. Relays are electrical devices, but they switch currents mechanically, and so their speed of operation is fundamentally of the same order as pure mechanical devices. It was recognized as early as 1919 that one could design a circuit out of vacuum tubes that could switch much faster, the switching being done inside the tube by a stream of electrons with negligible mass. Tubes were prone to burning out, operating them required a lot of power, which in turn had to be removed as excess heat. There was little incentive to build calculating machines out of tubes unless their advantage in speed overcame those drawbacks.

In the mid-1930s John V. Atanasoff, a physics Professor at Iowa State University, recognized the advantages of tube circuits for the solution of systems of linear equations. This type of problem is found in nearly every branch of physics, and its solution requires carrying out large numbers of ordinary arithmetic operations plus the storage of intermediate results. With a modest university grant Atanasoff began building circuits in 1939 and by 1942 had a prototype that worked except for intermittent failures in its intermediate storage unit. At that point Atanasoff moved to Washington, D.C. to work on other wartime projects. He never finished his computer. At the same time in Germany, a colleague of Zuse's named Helmut Schreyer developed tube circuits that he proposed as a substitute for the relays Zuse was then using. His proposal formed the basis of his doctoral dissertation, but aside from a few breadboard models little progress was made.

The first major, successful application of vacuum tubes to computing came in England, where a team of codebreakers, in ultra secrecy, developed a machine to assist with the decoding of intercepted German military radio traffic. Here was a clear case where electronic speeds were needed: not only were there many combinations of "keys" to consider, but the military value of an intercepted military message diminishes rapidly with time, often becoming utterly worthless in a few days. The first so-called "Colossus" was completed by 1943 (about the time the ENIAC was begun), and by war's end there were ten in operation. Details of the Colossus remain secret, even after 65 years. But it has been revealed that although these machines did not perform arithmetic as a calculator did, they could and did perform logical operations on symbolic information, which is the heart of any electronic processing circuit today.

The ENIAC, built at the University of Pennsylvania and unveiled to the public in February 1946, belongs more to the tradition of the machines just described than to the general purpose electronic computers that followed. It was conceived, proposed, and built to solve a specific problem—the calculation of firing tables for the Army. Its architecture reflected what was required for that problem, and it was an architecture that no subsequent computers imitated. Only one was built. And though the end of the war reduced the urgency to compute firing tables, military work dominated the ENIAC's schedule throughout its long lifetime (it was shut down in 1955). In the 1940s computing was advancing on a number of fronts. The examples mentioned above were the most prominent, but behind them were a host of other smaller yet also significant projects.

The metaphor of linear progress (i.e., using the term "milestone") is inappropriate. Advances in computing in the 1940s were more like an army advancing across broken terrain. The ENIAC, by virtue of its dramatic increase in arithmetic speeds, pushed the "calculating" function of computing machines way ahead of the other functions of computers, such as the storage of data or the output of results. These now had to scurry to catch up. Of those other functions, none appeared as a greater hindrance than the one of supplying the processor with instructions. John Mauchly said it succinctly: "Calculations can be performed at high speed only if instructions are supplied at high speed." So while it was being built, the ENIAC revealed to its creators the need for internal, electronic storage of instructions. Every machine has "software": a set of procedures by which it is properly used. Before electronics, the speeds of machinery were commensurate with human beings. Only with the electronic computer is there this bifurcation, and that is the truly "revolutionary" nature of the digital age. The ENIAC, by virtue of its high arithmetic speeds, brought programming to the fore. (It is no coincidence that the term "to program" a computer came from the ENIAC team.)

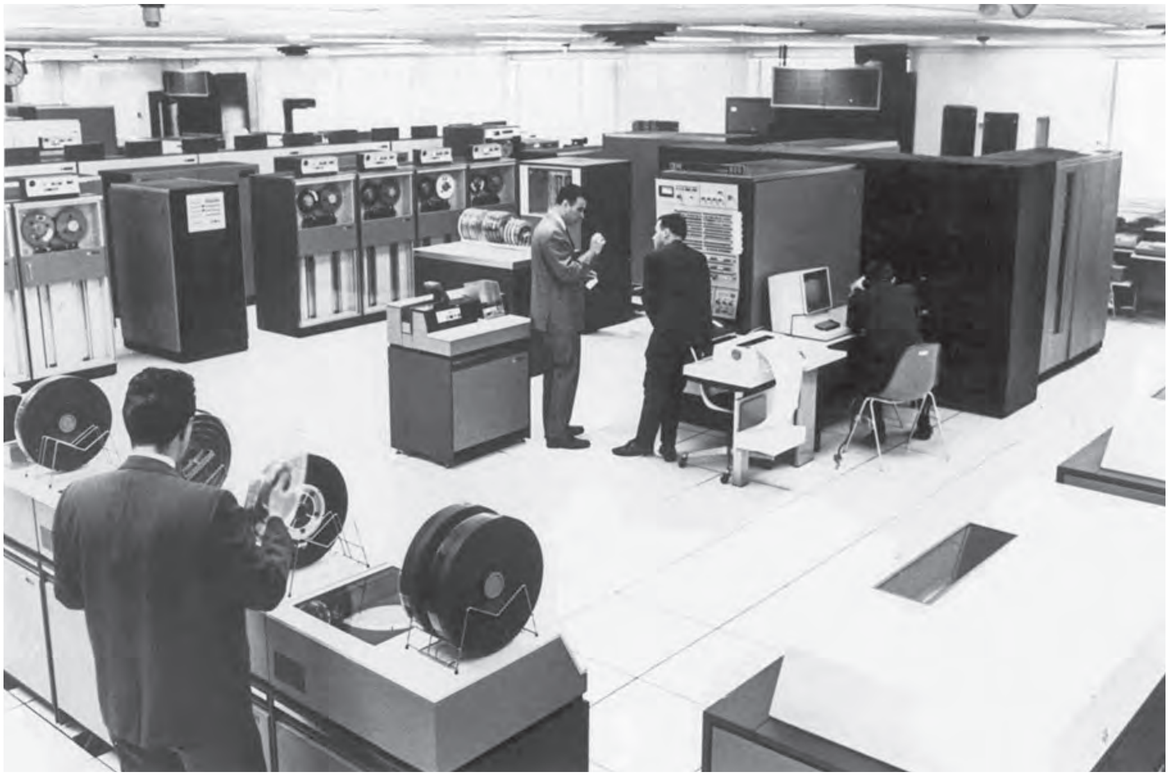


UNIVAC I computer, at the Lawrence Livermore Laboratory, California, ca. 1952. Smithsonian Institution.

The ENIAC is thus in the ironic position of being a pivot of history because of its shortcomings as well as its capabilities. It was not programmed but laboriously "set up" by plugging wires, in effect rewiring the machine for each new job. That meant that a problem that took minutes to solve might require several days to set up. By contrast, the ENIAC's electromechanical cousins, like the Harvard Mark I, might be programmed in a few hours but take days to run through the equations.

Even as the ENIAC was taking shape in the early 1940s its designers were thinking about what the machine's successor would look like. The ENIAC team was in hindsight perfectly suited to the task: it included people with skills in electrical engineering, mathematics, and logic. Out of their discussions came a notion of designing a computer with a dedicated memory unit, one that stored data but did not necessarily perform arithmetic or other operations on its contents. Instructions as well as data would be stored in this device, each capable of being retrieved or stored at high speeds. That requirement followed from the practical need for speed, as Mauchly stated above, as well as the engineering desire to have the memory unit kept simple without the extra complication of partitioning it and allocating space for one or the other type of data.

From that simple notion came much of the power of computing that followed. It has since become associated with John von Neumann, who joined the ENIAC team and who in 1945 wrote a report about



IBM System 360 Mainframe Installation, ca. 1965. The System 360 was one of the most popular and influential mainframe computers, and formed the basis for IBM's main line of business into the 1990s. Smithsonian Institution.

the ENIAC's successor, the EDVAC, in which the notion is explained. But clearly it was a collaborative effort, with the ENIAC then under construction as a backdrop.

All the advantages of this design would be for naught if one could not find a reliable, cheap, and fast memory device of sufficient capacity. Eckert favored using tubes of mercury that circulated acoustic pulses; von Neumann hoped for a special vacuum tube. The first true stored-program computers to operate used either the mercury tubes or a modified television tube that stored data as spots of electrical charge (Randell 1975). These methods offered high speed but were limited in capacity and were expensive. Many other designers opted to use a much slower, but more reliable, revolving magnetic drum. Project Whirlwind, at MIT, broke through this barrier when in the early 1950s its team developed a way of storing data on tiny magnetized "cores"—doughnut shaped pieces of magnetic material (Redmond and Smith 1980).

Generations: 1950–1970

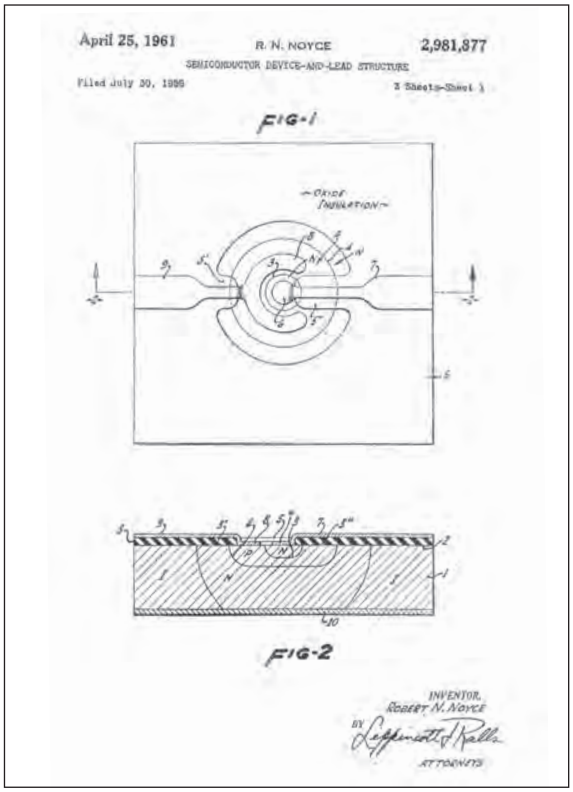
Eckert and Mauchly are remembered for more than their contributions to computer design. It was they, almost alone in the early years, who sought commercial applications of their invention, rather than confining it to scientific, military, or very large industrial uses. The British were the first to develop a computer for commercial use: the LEO, a commercial version of the EDSAC computer, built for the catering company J. Lyons & Company, Ltd. And it was in use by 1951. But like Babbage's inventions of the previous century, the British were unable to follow through on their remarkable innovation (Bird 1994). In the United States, Eckert and Mauchly faced similar skepticism when they proposed building computers for commercial use, but they were eventually able to succeed although losing their independence in the process. Given the engineering difficulties of getting this equipment to operate reliably, the skepticism was justified. Nevertheless, by the mid-1950s Eckert and Mauchly were able to offer a large commercial computer called the UNIVAC, and it was well received by the approximately twenty customers who acquired one.

Other companies, large and small, entered the computer business in the 1950s, but by the end of the decade IBM had taken a commanding lead. That was due mainly to its superior sales force, which ensured that customers were getting useful results out of their expensive investment in electronic equipment. IBM offered a separate line of electronic computers for business and scientific customers, as well as a successful line of smaller, inexpensive computers, like the 1401. By 1960 the transistor, invented in the

1940s, was reliable enough to replace the fragile vacuum tubes of an earlier day. Computer memory now consisted of a hierarchy of magnetic cores, then slower drums or disks, and finally high-capacity magnetic tape. Entering data and programs into these "mainframes" was still a matter of punching cards, thus ensuring continuity with the Hollerith equipment that was IBM's foundation.

In 1964 IBM unified its product line with its "System/360," which not only covered the full circle of science and business applications (hence the name), but which also was offered as a family of ever-larger computers each promised to run the software developed for those below it. This was a dramatic step that transformed the industry again, as the UNIVAC had a decade earlier. It was recognition that "software," which began as almost an afterthought in the crush of hardware design, was increasingly the driving engine of advances in computing.

Following IBM in the commercial market were the "Seven Dwarfs": Burroughs, UNIVAC, National Cash Register, Honeywell, General Electric, Control Data Corporation, and RCA. England, where the first practical stored-program computers operated in the late 1940s, also developed commercial products, as did France. Konrad Zuse, whose "Z3" operated in 1941, also founded a company—perhaps the world's first devoted to making and selling computers. But with only minor exceptions, European sales never approached those of US firms. The Soviets, although competitive with the US in space exploration, could not do the same in computers. They had to content themselves with making copies of the IBM System/360, which at least gave them the advantage of all the software developed by others. Why the USSR lagged so far behind is a mystery, given its technical and especially mathematical excellence. Perhaps Soviet planners saw the computer as a double-edged sword, one that could facilitate State planning but also made possible decentralized sharing of information. Certainly the absence of a vigorous free-market economy, which drove the technical advances at UNIVAC and IBM, was a factor. In any event, free-market forces in the US were augmented by large amounts of money supplied by the Defense Department, which supported computing for so-called "command-and-control" operations as well as for logistics and on-board missile guidance and navigation.

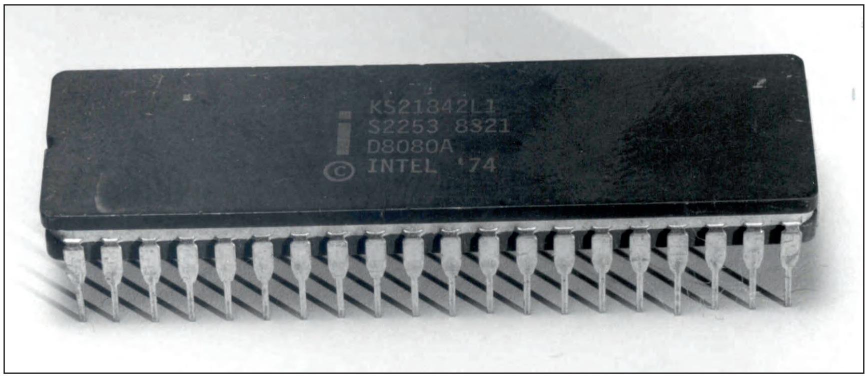


Robert Noyce, Patent for the Integrated Circuit ("chip"), 1961. Noyce, who at the time worked at Fairchild Semiconductor, and Jack Kilby, who worked at Texas Instruments, are usually credited as co-inventors of the integrated Circuit. United States Patent and trademark Office.

The minicomputer and the chip

If computing technology had stood still in the mid-1960s, one would still speak of a "computer revolution," so great would its impact on society have been. But technology did not stand still; it progressed at ever-greater rates. It took ten years for the transistor to come out of the laboratory and into practical commercial use in computers. That had an effect on the large mainframe systems already mentioned, but the transistor had an even bigger effect on smaller systems. Beginning around 1965, several new products appeared that offered high processing speeds, ruggedness, small size, and a low price that opened entirely new markets. The "PDP-8," announced that year by a new company called Digital Equipment Corporation, inaugurated this class of "minicomputers." A concentration of minicomputer firms emerged in the Boston suburbs. Both in people and in technology, the minicomputer industry was a direct descendant of the Defense Department funded Project Whirlwind at MIT (Ceruzzi 1998).

As computer designers began using transistors, they had to confront another technical problem, which in earlier years had been masked by the fragility of vacuum tubes. That was the difficulty of assembling, wiring, and testing circuits with thousands of discrete components: transistors, resistors, and capacitors. Among the many proposed solutions to this interconnection problem were those from Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Semiconductor, who each



Intel 8080 Microprocessor, 1974. The Intel 8080 was used in the first personal computers. It was not the first microprocessor, but it was the first to have on a sinlge chip the power of a practical computer. Smithsonian Institution.

filed for patents in 1959. Their invention came to be known as the "integrated circuit." Drawing on the base of knowledge built up on silicon transistors, these two companies were able to bring this invention into commercial use quickly: by the end of the 1960s the silicon chip had become the principal device in computer processors and was beginning to replace memory cores as well.

Besides co-inventing the integrated circuit, Noyce did something else that would shape the direction of computing. In 1968 he left Fairchild and co-founded a new company, called Intel, devoted to making memory chips as a replacement for magnetic cores. The Santa Clara Valley, on the peninsula south of San Francisco, was already a center for microelectronics. But Noyce's founding of Intel raised that activity to a feverish pitch. In 1971 a journalist dubbed the region "Silicon Valley": a name that implies not just the computer engineering that goes on there but also the free-wheeling, entrepreneurial culture that drives it (Ceruzzi 1998).

By the mid-1970s IBM's dominance of computing worldwide was under assault from three directions. From Silicon Valley and the Boston suburbs came waves of small but increasingly capable systems. From the US Justice Department came an antitrust suit, filed in 1969, charging IBM with unfairly dominating the industry. From computer scientists doing software research came the notion of interactive use of computers by a procedure known as "time sharing," which gave a number of users the illusion that the big, expensive computer was their own personal machine. Time sharing offered another avenue to get computing power into the hands of new groups of users, but the promise of a cheap "computer utility," analogous to the electric power grid that supplied power to one's home, did not materialize at that time.

An important component of this movement toward interactive computing was the development in 1964 of the BASIC programming language at Dartmouth College in New Hampshire, where students from liberal arts as well as science or engineering backgrounds found the computer more accessible than those at other colleges, who had to submit their programs as decks of punched cards, coded in less-friendly languages, and wait for the computer to come around to their place in the queue.

The personal computer

These assaults on the mainframe method of computing converged in 1975, when an obscure company from New Mexico offered the "Altair"—billed as the world's first computer kit and selling for less than $400. This kit was just barely a "computer," and one had to add a lot more equipment to get a practical system (Kidwell and Ceruzzi 1994). But the Altair's announcement touched off an explosion of creative energy that by 1977 had produced systems that could do useful work. These systems used advanced silicon chips both for processing and memory; a floppy disk (invented at IBM) for mass storage; and the BASIC programming language to allow users to write their own applications software. This version of BASIC was written by a small group led by Bill Gates, who dropped out of Harvard and moved to New Mexico to develop software for the Altair. The net result was to topple IBM's dominance of the computer industry. None of the giants doing battle with IBM did very well in the following decade either. Even Digital Equipment Corporation, in many ways the parent of the personal computer, faced near bankruptcy in the early 1990s.

The personal computer brought the cost of computing way down, but machines like the Altair were not suitable for anyone not well-versed in digital electronics and binary arithmetic. By 1977 several products appeared on the market that claimed to be as easy to install and use as any household appliance. The most influential of them was the Apple II. Apple's founders, Steve Jobs and Steve Wozniak, were the Silicon Valley counterpart to Eckert and Mauchly: one a first-rate engineer, the other a visionary who saw the potential of the computer if made accessible to a mass market (Rose 1989). In 1979 a program called "Visicalc" appeared for the Apple II: it manipulated rows and columns of figures known to accountants as a "spread sheet," only much faster and easier than anyone had imagined possible. A person owning Visicalc and an Apple II could now do things that even a large mainframe could not do easily. Finally, after decades of promise, software—the programs that get a computer to do what one wants it to do—came to the fore where it really belonged. A decade later it would be software companies, like Bill Gates' Microsoft, that would dominate the news about computing's advances.

Although it had a reputation as a slow-moving, bloated bureaucracy, IBM was quick to respond to Apple's challenge, and brought out its "PC" in 1981. In a radical departure for IBM, but typical of minicomputers and other personal computers, the PC had an open architecture that encouraged other companies to supply software, peripheral equipment, and plug-in circuit cards. The IBM PC was more successful in the marketplace than anyone had imagined. The IBM name gave the machine respectability. It used an advanced processor from Intel that allowed it to access far more memory than its competitors. The operating system was supplied by Microsoft. A very capable spreadsheet program, Lotus 1-2-3, was offered for the PC and its compatible machines.

Apple competed with IBM in 1984 with its "Macintosh," which brought advanced concepts of the so-called "user interface" out of the laboratories and into the popular consciousness. The metaphor of treating files on a screen as a series of overlapping windows, with the user accessing them by a pointer called a "mouse," had been pioneered in military-sponsored labs in the 1960s. In the early 1970s had been further developed by a brilliant team of researchers at the Silicon Valley laboratory of the Xerox Corporation. But it remained for Apple to make that a commercial success; Microsoft followed with its own "Windows" operating system, introduced around the same time as the Macintosh but not a market success until 1990. For the next decade the personal computer field continued this battle between the Apple architecture and the one pioneered by IBM that used Intel processors and Microsoft system software.

Altair Personal Computer, 1974. A small hobbyist company, MITS, of Albuquerque, New Mexico, introduced this computer as a kit in 1974, and it sparked the revolution in personal computing. Smithsonian Institution.

The beginnings of networking

During the 1980s personal computers brought the topic of computing into the popular consciousness. Many individuals used them at work, and a few had them at home as well. The technology, though still somewhat baffling, was no longer mysterious. While personal computers dominated the popular press, the venerable mainframe computers continued to dominate the industry in terms of the dollar value of installed equipment and software. Mainframes could not compete with PC programs like spreadsheets and word processors, but any applications that required handling large amounts of data required mainframes. Beginning in the 1970s, these computers began to move away from punched cards and into interactive operations, using keyboards and terminals that superficially resembled a personal computer. Large, on-line database systems became common and gradually began to transform business and government activities in the industrialized world. Some of the more visible of these applications included airline reservations systems, customer information and billing systems for utilities and insurance companies, and computerized inventory and stocking programs for large retail. The combination of on-line database and billing systems, toll-free telephone numbers, and credit card verification and billing over the telephone transformed the once-humble mail order branch of retailing into a giant force in the American economy.

All of these activities required large and expensive mainframe computers, with software custom written at great expense for each customer. One was tempted to hook up an array of cheap personal computers running inexpensive software packages, but this was not feasible. Hitching another team of horses to a wagon might allow one to pull more weight, but the wagon will not go faster. Even that has its limits as it becomes increasingly difficult for the teamster to get the horses all to pull in the same direction. The problem with computing was similar and was expressed informally as "Grosch's Law": for a given amount of money, one gets more work out of one big computer than out of two smaller ones (Grosch 1991).

But that would change. At the Xerox Palo Alto Research Center in 1973, where so many advances in the user interface were made, a method of networking was invented that finally overturned this law. Its inventors called it "Ethernet," after the medium that nineteenth-century physicists thought carried light. Ethernet made it practical to link smaller computers in an office or building to one another, thereby sharing mass memory, laser printers (another Xerox invention), and allowing computer users to send electronic mail

Xerox "Alto" Workstation, ca. 1973. The Alto was designed and built at the Xerox Palo Alto Research Center (PARC) in California. It pioneered in the use of a mouse and a graphical user interface, which eventually became common in personal computers. Altos were linked to one another by Ethernet and sent their output to laser printers, both Xerox innovations. Smithsonian Institution.

to one another. At the same time as Ethernet was making local networking practical, an effort funded by the Defense Department's Advanced Research Projects Agency (ARPA) was doing the same for linking computers that were geographically dispersed. ARPA was concerned with maintaining secure military communications in the event of war, when sections of a network might be destroyed. Early military networks descended from Project Whirlwind had central command centers, and as such were vulnerable to an attack on the network's central control. These centers were housed in windowless, reinforced concrete structures, but if they were damaged the network was inoperable (Abbate 1999).

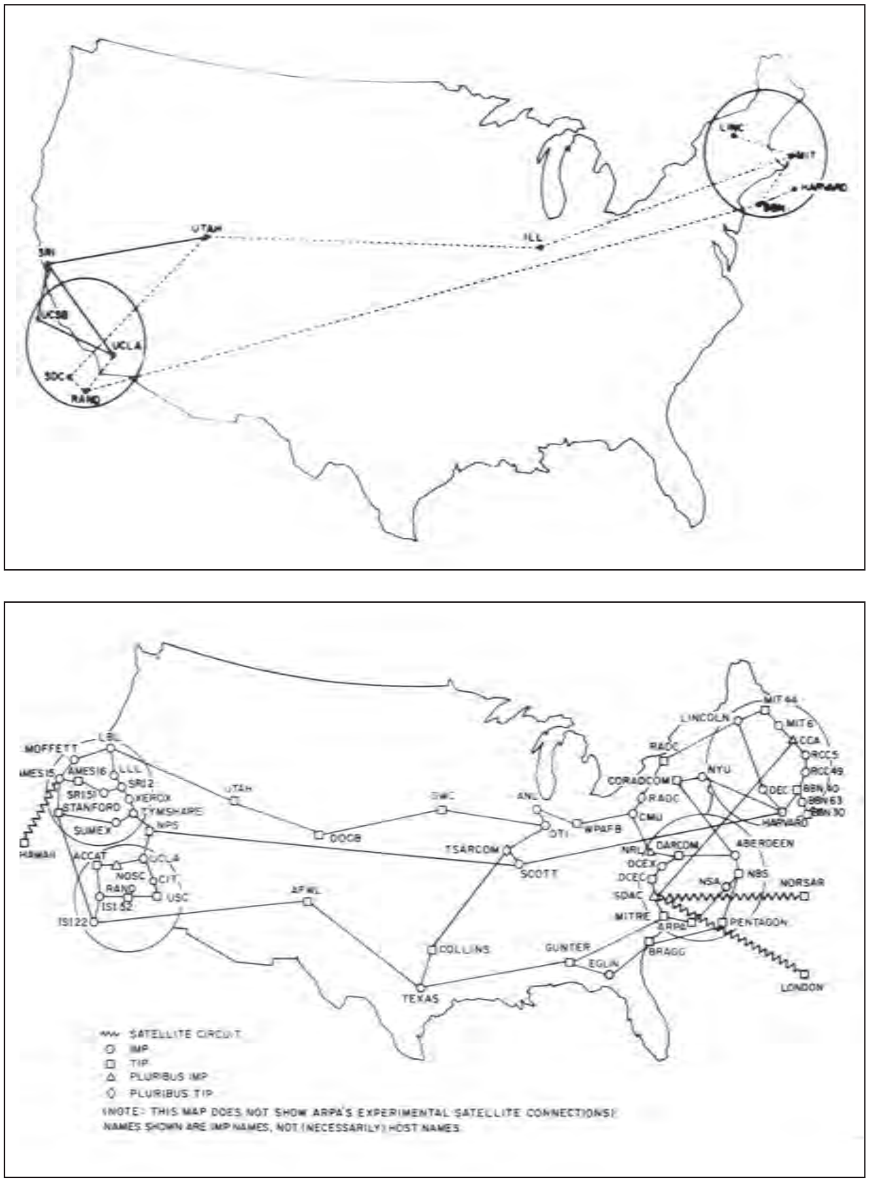
With funding from ARPA, a group of researchers developed an alternative, in which data was broken up into "packets," each given the address of the computer to receive it, and sent out over a network. If one or more computers on the network were inoperable, the system would find an alternate route. The computer at the receiving end would re-assemble the packets into a faithful copy of the original transmission. By 1971 "ARPANET" consisted of 15 nodes across the country. It grew rapidly for the rest of that decade. Its original intent was to send large data sets or programs from one node to another, but soon after the network came into existence people began using it to send brief notes to one another. At first this was an awkward process, but in 1973 that was transformed by Ray Tomlinson, an engineer at the Cambridge, Massachusetts firm Bolt Beranek and Newman. Tomlinson came up with a simple notion of separating the name of a message's recipient and that person's computer with an "@" sign—one of the few non-alphabetic symbols available on the Teletype console that ARPANET used at the time. Thus was modern e-mail conceived, and with it, the symbol of the networked age.

The pressure to use ARPANET for general-purpose e-mail and other non-military uses was so great that it was split up. One part remained under military control. The other part was turned over to the US-funded, civilian National Science Foundation, which sponsored research not only to expand this network but also to allow interconnection among different types of networks (for example, networks that used radio instead of wires). Researchers began calling the result an "internet," to reflect its heterogeneous nature. In 1983 the networks adopted a set of standards for data transmission, called "Transmission Control Protocol/ Internet Protocol" (TCP/IP), with such interconnection. These protocols are still in use today and are the basis for the modern Internet (Aspray and Ceruzzi 2008).

These local and remote networking schemes fit well with other developments going on in computer hardware and software. A new type of computer emerged, called a "workstation," which unlike the personal computer was better suited for networking. Another critical distinction was that they used an operating system called "UNIX," which though difficult for consumers was well-suited to networking and other advanced programming. UNIX was developed at Bell Laboratories, the research arm of the US government-regulated telephone monopoly AT&T. Groups of workstations, linked locally by Ethernet to one another, and by the Internet to similar clusters world-wide, finally offered a real alternative to the large mainframe installation for many applications.

The Internet Age

The National Science Foundation, an agency of the US government, could not allow commercial use of the Internet that it controlled. It could, however, offer the use of the Internet protocols to anyone who wished to use them at little or no cost, in contrast to the networking protocols offered by computer companies like IBM. As Internet use grew, the NSF was under pressure to turn it over to commercial firms to manage it. A law passed by the US Congress in 1992 effectively ended the prohibition against commercial use, and one could say that with the passage of that law, the modern Internet Age began. That was not entirely true, as the US government retained control over the addressing scheme of the Internet—e.g. the suffixes ".com," ".edu," and so on, which allow computers to know where an electronic message is sent. By the turn of the twenty-first century, a number of countries asked that this control be turned over to the United Nations, but so far the US has resisted. The Internet is truly a resource offered freely to all countries of the world, but its master registry of domain names is managed by an American private company whose authority is given by the US Department of Commerce.

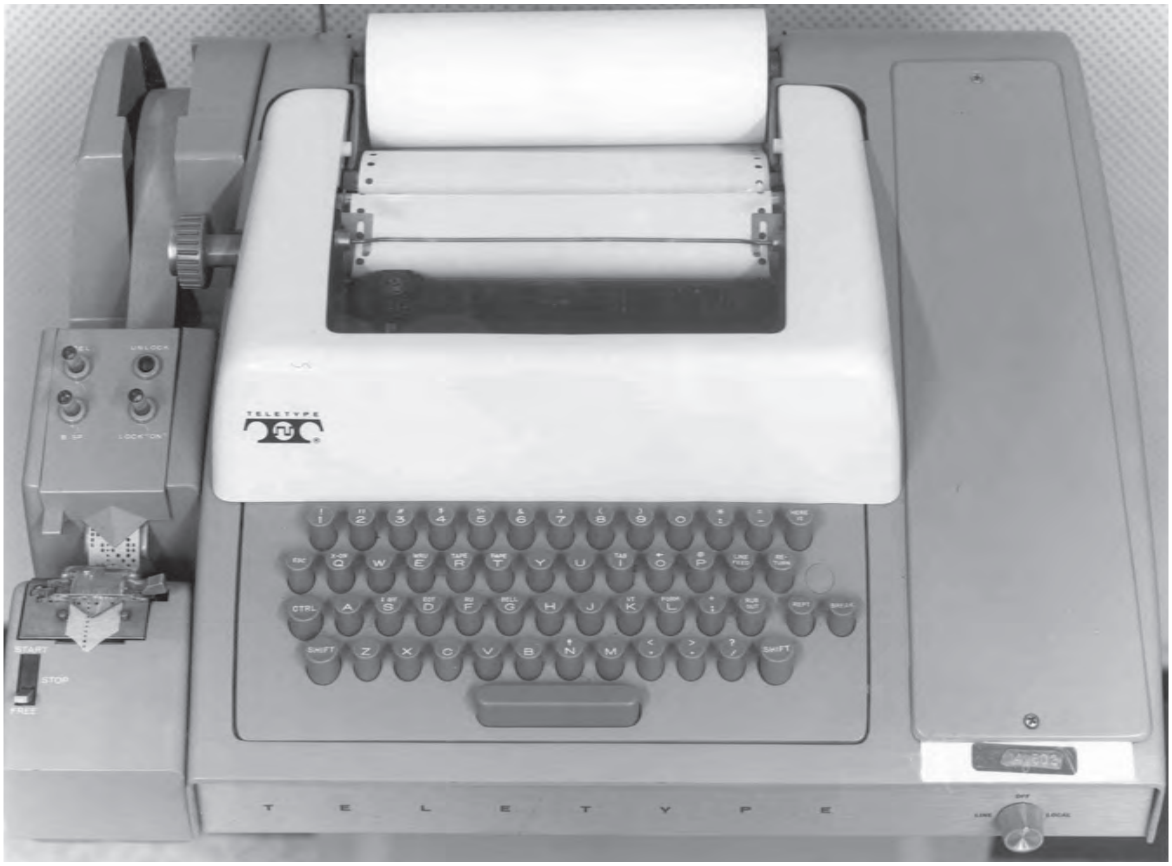


ARPANET, 1970, and 1974. The modern Internet is descended from this military-sponsored network, which grew rapidly from its inception during the 1970s. Credit: U.S. Department of Defense, Defense Advanced Research Projects Agency.

This political activity was complemented by dramatic advances in computer technology, which further led to the rapid spread of the Internet. By 1990 the expensive UNIX workstations had given way to personal computers that used advanced processors, especially a processor called the "Pentium," supplied by Intel. On the software side, new versions of the Microsoft Windows operating system came with the Internet protocols and other networking software installed. This combination gave PCs the equivalent power of the workstation. UNIX is rarely found on the PC, although the more powerful servers and so-called "routers" that perform the basic switching for the Internet continue to use it. A variant of UNIX called "Linux," developed in 1991 by Linus Torvalds in Finland, was offered as a free or low-cost alternative to the Microsoft Windows system. It and related software gained a small but significant market share. These came to be called "open source" software, defined as "free" but not without restrictions (Williams 2002).

While this activity was going on at government and university laboratories, personal computer users were independently discovering the benefits of networking. The first personal computers like the Apple II did not have much ability to be networked, but resourceful hobbyists developed ingenious ways to communicate anyway. They used a device called a "modem" (modulator-demodulator) to transmit computer data slowly as audio tones over ordinary telephone lines. In this they were helped by a ruling by the US telephone monopoly, that data sent over a telephone line was not treated any differently than voice calls. Local calls were effectively free in the United States, but long-distance calls were expensive. Personal computer enthusiasts worked out ways of gathering messages locally, and then sending them across the country to one another at night, when rates were lower (the result was called "FidoNet," named after a dog that "fetched" data). Commercial companies arose that served this market as well; they rented local telephone numbers in most metropolitan areas, and charged users a fee for connecting to them. One of the most influential of these was called "The Source," founded in 1979; after some financial difficulties it was reorganized and became the basis for America Online, the most popular personal networking service from the late 1980s through the 1990s.

These personal and commercial systems are significant because they introduced a social dimension to networking. ARPANET was a military network. Its descendents frowned on frivolous or commercial use. But the personal networks, like the house telephone over which their messages ran, were used for chats,

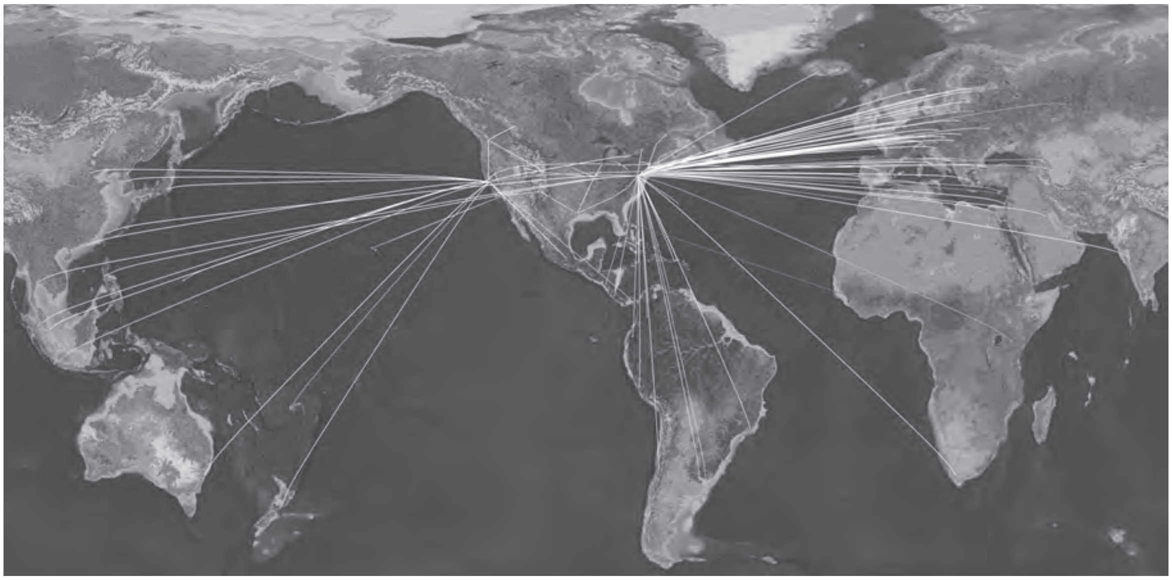


Teletype Model ASR-33. The ARPANET used this modified Teletype as a terminal. Note the "@" sign, which was adopted for e-mail and has become the icon of the networked age. Smithsonian Institution.

freewheeling discussions, news, and commercial services right from the start. One of the commercial networks, Prodigy, also incorporated color graphics—another staple of today's Internet. The histories of the Internet that concentrate on ARPANET are correct: ARPANET was the technical ancestor of the Internet, and the Internet protocols emerged from ARPA research. But a full history of the Internet must include the social and cultural dimension as well, and that emerged from Prodigy, AOL, and the community of hobbyists.

By the late 1980s it was clear that computer networks were desirable for both the home and the office. But the "Internet," the network that was being built with National Science Foundation support, was only one of many possible contenders. Business reports from those years were championing a completely different sort of network, namely the expansion of cable television into a host of new channels—up to 500, according to one popular prediction. The reconfigured television would also allow some degree of interactivity, but it would not be through a general-purpose, personal computer. This concept was a natural outgrowth of the marketing aims of the television and entertainment industry. Among the scientists and computer professionals, networking would come in the form of a well-structured set of protocols called "Open Systems Interconnection" (OSI), which would replace the more freewheeling Internet. None of this happened, largely because the Internet, unlike the competing schemes, was designed to allow disparate networks access, and it was not tied to a particular government-regulated monopoly, private corporation, or industry. By the mid-1990s private networks like AOL established connections to the Internet, and the OSI protocols fell into disuse. Ironically, it was precisely because the Internet was available for free and without any specific commercial uses in mind, that allowed it to become the basis for so much commercial activity once it was released from US government control after 1993 (Aspray and Ceruzzi 2008).

In the summer of 1991, researchers at the European particle physics laboratory CERN released a program called the World Wide Web. It was a set of protocols that ran on top of the Internet protocols, and allowed a very flexible and general-purpose access to material stored on the Internet in a variety of formats. As with the Internet itself, it was this feature of access across formats, machines, operating systems, and standards



National Science Foundation (NSF) Network, ca. 1991. The National Science Foundation supported the transition of networking from military to civilian use. As a government agency, however, it still restricted the network to educational or research use. When these restrictions were lifted shortly after this map was produced, the modern commercial Internet began. U.S. National Science Foundation

that allowed the Web to become popular so rapidly. Today most consumers consider the Web and the Internet to be synonymous; it is more accurate to say that the later was the foundation for the former. The primary author of the Web software was Tim Berners-Lee, who was working at CERN at the time. He recalled that his inspiration for developing the software came from observing physicists from all over the world meeting together for scientific discussions in common areas at the CERN buildings. In addition to developing the Web, Berners-Lee also developed a program that allowed easy access to the software from a personal computer. This program, called a "browser," was a further key ingredient in making the Internet available to the masses (Berners-Lee 1999). Berners-Lee's browser saw only limited use; it was soon replaced by a more sophisticated browser called "Mosaic," developed in 1993 at the University of Illinois in the United States. Two years later the principal developers of Mosaic left Illinois and moved to Silicon Valley in California, where they founded a company called Netscape. Their browser, called "Navigator," was offered free to individuals to download; commercial users had to pay. Netscape's almost instant success led to the beginning of the Internet "bubble" whereby any stock remotely connected to the Web was traded at absurdly high prices. Mosaic faded away, but Microsoft purchased rights to it, and that became the bases for Microsoft's own browser, Internet Explorer, which today is the most popular means of access to the Web and to the Internet in general (Clark 1999).

Conclusion

The history of computing began in a slow orderly fashion, and then careened out of control with the advent of networking, browsers, and now portable devices. Any narrative that attempts to chart its recent trajectory is doomed to failure. The driving force for this is Moore's Law: an observation made by Gordon Moore, one of the founders of Intel, that silicon chip memory doubles in capacity about every 18 months (Moore 1965). It has been doing this since the 1960s, and despite regular predictions that it will soon come to an end, it seems to be still in force. The capacities of mass storage, especially magnetic disks, and the bandwidth of telecommunications cables and other channels have been increasing at exponential rates as well. This puts engineers on a treadmill from which there is no escape: when asked to design a consumer or commercial product, they design it not with the capabilities of existing chips in mind, but with what they anticipate will be the chip power at the time the product is brought to the market. That in turn forces the chip makers to come up with a chip that meets this expectation. One can always find predictions in the popular and trade press that this treadmill has to stop some day: at least when the limits of quantum physics make it impossible to design chips that have greater density. But in spite of these regular predictions that Moore's Law will come to an end, it has not. And as long as it holds, it is impossible to predict a "trajectory" for computing for even the next year. But that does make this era one of the most exciting to be living in, as long as one can cope with the rapidity of technological change.

Bibliography

Abbate, J. *Inventing the Internet.* Cambridge, Massachusetts: MIT Press, 1999.

Aspray, W., ed. *Computing Before Computers.* Ames, Iowa: Iowa State University Press, 1990.

—, and P. E. Ceruzzi, eds. *The Internet and American Business*. Cambridge, Massachusetts, 2008.

Berners-Lee, T. and M. Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. San Francisco: Harper, 1999.

Bird, P. *LEO: The First Business Computer*. Berkshire, United Kingdom: Hasler Publishing, 1994.

Burks, A. R. and W. Arthur. *The First Electronic Computer: The Atanasoff Story*. Ann Arbor, Michigan: University of Michigan Press, 1988.

Ceruzzi, P. E. *Reckoners: the Prehistory of the Digital Computer, From Relays to the Stored Program Concept, 1935-1945*. Westport, Connecticut: Greenwood Press, 1983.

—, *A History of Modern Computing.* Cambridge, Massachusetts: MIT Press, 1998.

Clark, J. and O. Edwards. *Netscape Time: The Making of the Billion-Dollar Start-Up that Took on Microsoft.* New York: St. Martin's Press, 1999.

Eames, Ch. and R. Offices of. *A Computer Perspective: Background to the Computer Age*. Cambridge, Massachusetts: Harvard University Press, 1990.

Eckert, W. J. *Punched Card Methods in Scientific Calculation*. New York: IBM Corporation, 1940.

Grosch, H. R. J. *Computer: Bit Slices from a Life*. Novato, California: Third Millennium Books, 1991.

Kidwell, P. A., and P. E. Ceruzzi. *Landmarks in Digital Computing: A Smithsonian Pictorial History*. Washington, D. C.: Smithsonian Institution Press, 1994.

Merzbach, U. *Georg Scheutz and the First Printing Calculator*. Washington, D. C.: Smithsonian Institution Press, 1977.

Moore, G. E. «Cramming More Components onto Integrated Circuits», *Electronics*, April 19, 1965, 114-117.

Randall, B., ed. *The Origins of Digital Computers: Selected Papers*. Berlin, Heidelberg and New York: Springer-Verlag, 1975.

Redmond, K. C. and Th. M. Smith. *Project Whirlwind: The History of a Pioneer Computer*. Bedford, Massachusetts: Digital Press, 1980.

Rose, F. *West of Eden: The End of Innocence at Apple Computer*. New York: Penguin Books, 1989.

Stern, N. *From ENIAC to UNIVAC: An Appraisal of the Eckert-Mauchly Computers.* Bedford, Massachusetts: Digital Press, 1981.

Williams, S. *Free as in Freedom: Richard Stallman's Crusade for Free Software*. Sebastopol, California: O'Reilly, 2002.

computers and space exploration

PAUL E. CERUZZI

The Soviet Union's successful launch of two Sputnik satellites in the fall of 1957 came as a shock to many Americans. Although the US intelligence community was not surprised, ordinary Americans were, and the two launches demonstrated without any doubt that the Soviet Union had a lead over the US not only in satellites, but in booster rockets, which could deliver weapons as well. Among the responses to Sputnik was the founding of agencies, one an arm of the US Defense Department, the other a civilian agency. One was the (Defense) "Advanced Research Projects Agency," or "ARPA," more recently known as "DARPA." ARPA's mission was plain: support long-term research that will make it unlikely that the US would ever again to be caught off guard as it was when the Sputniks were lau ched. One of ARPA's research areas was in missiles and space exploration; by the end of 1958 most of that work was transferred to another agency, under civilian control: the National Air and Space Administration (NASA). Both were in 1958 (Norberg and O'Neil 1996).

In the fifty years since their founding, one can list a remarkable number of achievements by each, but chief among those achievements are two. Beginning in the mid-1960s, DARPA designed and build a network of computers, known as ARPANET, which was the technical inspiration for today's Internet. And NASA, responding to a challenge by President John F. Kennedy in 1961, successfully landed a dozen astronauts on the Moon and retuned them safely to Earth between 1969 and 1972.

In the mid-1990s, the Internet moved rapidly from a network known only to computer scientists or other specialists, to something that was used by ordinary citizens across the industrialized world. In the US, the non-profit Public Broadcasting Service produced a multi-part television program to document the meteoric rise of this phenomenon. It was given the whimsical title "Nerds 2.0.1: A Brief History of the Internet" (Segaller 1998). The title suggested that the Internet was a creation of "nerds": mostly young men, few of them over thirty years old, whose obsessive tinkering with computers led to this world-changing social phenomenon. In nearly every episode of the television program, the narrator noted the contrast between the accomplishments of the two agencies founded at the same time: the Internet as a descendant of ARPA's work, the manned landings on the Moon the result of NASA's.

The body of the program elaborated further on this theme. The program—correctly—noted that the Internet descended from the ARPANET, a computer network designed for, and sponsored by the US military. The show went a step further: it argued that the Moon landings were a one-time stunt, with little or no long-term impact on society, while the Internet was a world-changing technology that did, and continues, to affect the lives or ordinary people around the world.

A half-century after the founding of those two agencies, we can revisit the relative achievements in computing and space exploration, and ask about the relationship those two technologies have had with each other. In both aerospace and computing, there has been tremendous progress, but the future did not turn out at all the way people thought it would.

In the late 1960s, many influential computer scientists predicted that computers would attain "Artificial Intelligence" (AI), and become our personal servants, perhaps even companions (McCorduck 1979). Science fiction writers embraced this theme and portrayed AI-enabled computers either as our beneficial servants, as found in the robots in the *Star Wars* movie series, or to our detriment, as found in the malevolent computer "HAL" in the movie *2001: A Space Odyssey*. But in spite of this recurring theme, that did not happen. Artificial Intelligence remains an elusive goal. However, outside of the narrow confines of the AI community of computer scientists, this "failure" does not bother anyone. The reason is simple: the advent of the personal computer, the Internet, the wireless telephone, and other advances have brought computing technology to the world at levels that surpass what most had envisioned at the time of the Moon landings. We cannot converse with them as we would another person, but these systems exhibit a surprising amount of what one may call "intelligence," more from their brute-force application of processing power and memory than from their inherent design as artificial substitutes for the human brain.

CRAY-1 Supercomputer, ca. 1976. The CRAY-1, designed by Seymour Cray, was the first "supercomputer," which could compete with wind tunnels in analyzing air and spacecraft design. CRAY Research, Inc.

In the realm of space exploration, the Apollo missions to the Moon generated predictions that also failed to come to pass: permanent outposts on the Moon, tourist hotels in Earth orbit, manned missions to Mars. None of these have happened yet, but advances in space technology have been remarkable. The Earth is now encircled by communications and weather satellites that are integrated into our daily lives. The Global Positioning System (GPS), and the planned European and Asian counterparts to it, provide precise timing and location services at low cost to the world. Robotic space probes have begun an exploration of Mars and the outer planets that rival the voyages of any previous age of exploration. Space telescopes operating in the visible and other wavelengths have ushered in a new era of science that is as exciting as any in history (Dick and Launius 2007).

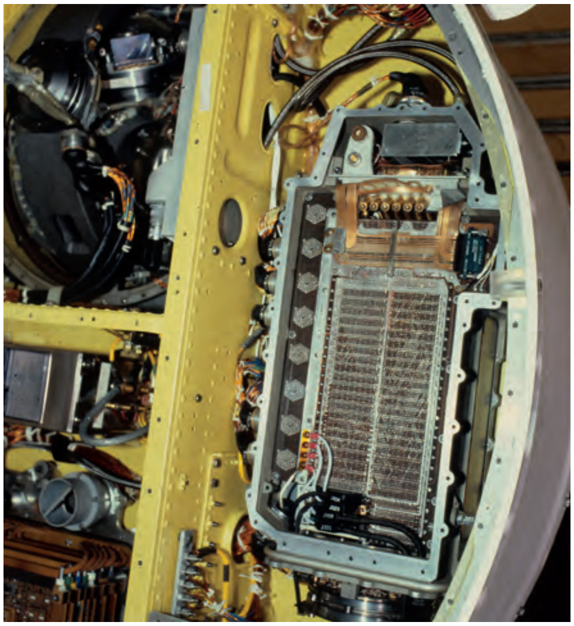
In the realm of computing, the advances in sheer memory capacity and processing power, plus networking, have more than covered any frustrations over the failure of computers to acquire human-like intelligence. In the realm of space exploration, the advances described above have not erased the frustration at not achieving a significant human presence off our planet. (In the related realm of aircraft that fly within the Earth's atmosphere, recent decades have likewise seen frustrations. Aircraft broke through the sound barrier in the late 1940s, but outside of a few specialized military systems, most aircraft today fly below the speed of sound. Commercial jetliners fly at about the same speed, and about the same altitude, as the first commercial jets that were introduced into service in the 1950s. The supersonic Concorde, though a technical marvel, was a commercial failure and was withdrawn from service.)

Hence the thesis of that television program: that the little-noticed computer network from ARPA overwhelms the more visible aeronautics and space achievements of NASA. Many viewers apparently agreed, regardless of whatever counter arguments NASA or other space enthusiasts raised against it.

For the past sixty years, computing and aerospace have been deeply interconnected, and it is hardly possible to treat the history of each separately. The invention of the electronic digital computer, which occurred in several places between about 1940 and 1950, was often connected to the solution of problems in the sciences of astronomy and aerodynamics, or in support of the technologies of aircraft design and production, air traffic control, anti-aircraft weapons, and later guided missile development. One of the inspirations for the development of ARPANET was the need to adapt communications networks to the crisis of control brought about by the development of ballistic missiles and jet-powered bombers. It was not simply a matter of designing a network that could survive a nuclear attack, as many popular histories assert; it was also a need to have a communications system that cold be as flexible and robust in keeping with the new military environment of aerospace after World War II (Abbate 1999).

After 1945, the US aerospace community had the further attribute of commanding large sums of money from the military arm of its government, as the US waged a Cold War with the Soviet Union. That pushed the development of digital computing much faster in the US than it progressed in England, the home of the first code-breaking computers, the first stored-program computers, and the first commercial computer. Some of that money was wasted, but US military support, mainly although not exclusively to support aerospace, was a powerful driver of the technology.

By its nature, a digital computer is a general-purpose device. If one can write a suitable program for it—admittedly a significant condition—then one can use a computer to serve a variety of ends. This quality, first described in theoretical terms by the English mathematician Alan Turing in the 1930s, set the computer apart from other machines, which are typically designed and optimized for one, and only one, function. Thus aerospace was but one of many places where computers found applications. The decade of the 1950s saw a steady increase in the power and memory capacity of mainframe computers, coupled with a development of general purpose software such as the programming language FORTRAN, and special-purpose software that was used for computer-aided design/computer-assisted manufacturing (CAD/CAM), stress analysis, or fluid dynamics.

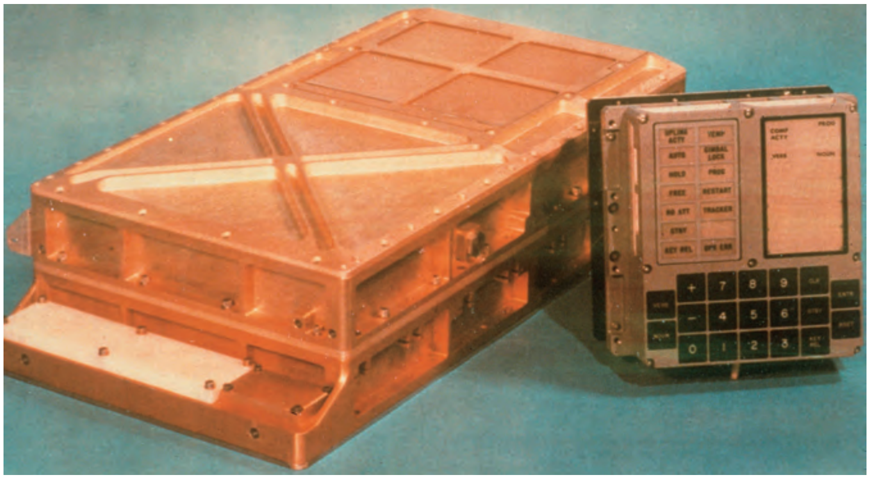


Minuteman III Guidance System, ca. 1970. The Minuteman, a solid fuel ballistic missile developed for the U.S. Air Force beginning in the 1960s, was a pioneer in its use of electronic components. For the first Minuteman, the Air Force developed what they called a "High Reliability" program for its electronic components. The Minuteman III, a later model, was a pioneer in the use of the newly-invented integrated circuit. Smithsonian Institution.

Unlike computer applications in, say, banking or finance, aerospace applications have an additional constraint. Until about 1960, computers were large, fragile, and consumed large amounts of power. That restricted their applications in aerospace to the ground—to airline reservations, wind-tunnel analysis, CAD/CAM, and the like. For aerospace, the computer's potential to become a universal machine as implied by Turing's thesis was thwarted by the hard reality of the need to adapt to the rigors of air and space flight. The aerospace and defense community, which in the 1950s in the US had vast financial resources available to it, was therefore in a position to shape the direction of computing in its most formative years. In turn, as computing addressed issues of reliability, size, weight, and ruggedness, it influenced aerospace as well during a decade of rapid change in flight technology (Ceruzzi 1989).

The transistor, invented in the late 1940s, was the first technological advance to address the issues of reliability, size, and weight. It took a long period of development, however, before the silicon transistor became reliable enough to allow computers to become small, rugged, and less power consuming. Transistorized computers began to appear in missile-guidance systems around 1960. In 1959 two engineers, Jack Kilby at Texas Instruments and Robert Noyce at Fairchild Instruments, went a step further and developed circuits that placed several transistors and other components on a single chip of material (at first germanium, later silicon). The integrated circuit, or silicon chip, was born. Neither Noyce nor Kilby was working on an aerospace application at the time. But aerospace needs provided the context for the chip's invention. In the dozen years between the invention of the transistor and the silicon chip, the US Air Force mounted a campaign to improve the reliability of electronic circuits in general. The Air Force was at the time developing ballistic missiles: million-dollar weapons that would sometimes explode on the launch pad because of the failure of an electronic component that may have cost less than one dollar. The electronic industry of the 1950s based its economic models on a consumer market, where low manufacturing costs, not high quality, were the way to achieve profits. Consumers at the time simply accepted the occasional failure of components, much as today they accept personal computer software that occasionally "crashes" (Ceruzzi 1998, 177-206).

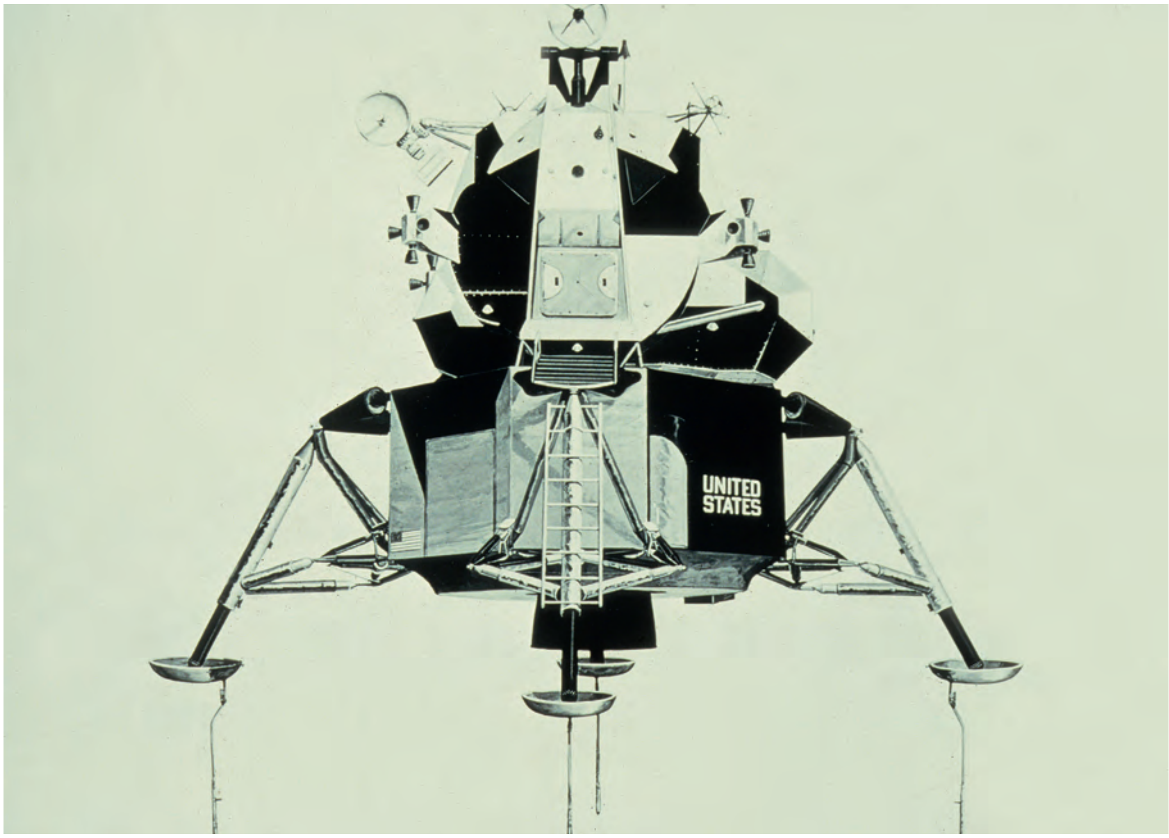
For aerospace applications, this model had to be abandoned. Computer crashes were not metaphorical but real. The Air Force's "High-Reliability" program of the late 1950s accomplished that goal. Manufacturers developed statistical quality control techniques; every step in a manufacturing process was rigorously documented. Devices were assembled in "clean rooms" (invented at a US weapons laboratory in New Mexico): more sterile than the finest hospital operating room. In them, workers wore suits that prevented hair or skin flakes from contaminating the assemblies, and filters screened out the tiniest particles of dust. Also during the 1950s, chemists developed ways of producing ultra-pure crystalline silicon, into which they could introduce very small and precise quantities of other elements to yield a material with the desired electronic properties (a process called "doping"). Much of this activity took place in what was once an agricultural valley south of San Francisco, soon dubbed "Silicon Valley" by a local journalist. The Fairchild Semiconductor Company, where Robert Noyce worked, was at the center of this creative activity. There, in addition to developing the silicon handling techniques mentioned above, engineers also developed a method of manufacturing transistors by photographic etching. All these advances took place before the Integrated Circuit was invented, but without them, what followed could not have happened.



Apollo Guidance Computer, 1969. The Apollo Guidance Computer performed critical guidance, navigation, and control functions for the missions that took a total of 12 astronauts to the Moon and back between 1969 and 1972. These computers, along with the Minuteman II guidance computers, were among the world's first to use integrated circuits. Next to the computer is a keyboard with buttons large enough to be pushed by an astronaut wearing a space suit. Smithsonian Institution.

The integrated circuit placed more than one device on a piece of material. At first the number of circuits on a chip was small, about five or six. But that number began to double, at first doubling every year, then at a doubling rate of about every 18 months. That doubling rate has remained in force ever since. It was christened "Moore's Law," by Gordon Moore, a colleague of Robert Noyce's at Fairchild, who was responsible for laying much of the material foundation for the chip's advances (Moore 1965). That law—really an empirical observation—has driven the computer industry ever since, and with it the symbiotic relationship with aerospace. In this context, it is not surprising that the first contract for large quantities of chips was for the US Air Force's Minuteman ballistic missile program, for a model of that missile that first flew in 1964. Following closely on the Minuteman contract was a contract for the computer that guided Apollo astronauts to the Moon and back, in a series of crewed missions that began in 1968 (Ceruzzi 1998, 182). By the time of the Apollo missions, Moore's Law was beginning to have a significant impact on aerospace engineering and elsewhere. The last Apollo mission, an Earth-orbit rendezvous with a Soviet Soyuz capsule, flew in 1975. Onboard was a pocket calculator made by the Silicon Valley firm Hewlett-Packard. That hand-held calculator had more computing power than the onboard Apollo Guidance Computer, designed a decade earlier when the chip was new. One could find numerous examples of similar effects.

The spectacular advances in robotic deep-space missions, and other accomplishments mentioned above, are largely a result of the effect of Moore's Law on spacecraft design—especially spacecraft that do not carry humans (who, for better or worse, have the same physical dimensions and need for food, water, and oxygen today as we had in 1959, when the silicon chip was invented). The direct comparison of the ARPANET with Project Apollo misses the nuances of this story. One of the ironies of history is that advances in space exploration have had an effect on aircraft design as well. The Apollo Lunar Module—the gangly craft that took two astronauts the final 100 kilometers from Lunar orbit to the Moon's surface—had to have computer control, as no human being could manage the delicacy of a lunar landing in the absence of an atmosphere, and ground controllers in Houston were too far away to be of help (Mindell 2008). At the end of the Apollo program, Apollo guidance computers were removed from spacecraft and installed in an experimental NASA aircraft, to see if aircraft could benefit from this technology as well. It was no coincidence that NASA choose as the



Lunar Module. The Lunar Module operated entirely in the vacuum of space, and it landed on the moon by use of its rocket engines. No human being could fly it, and because of the distance from Earth, it had to be controlled by an on-board computer. NASA.

manager of this program none other than Neil Armstrong, the first person to walk on the Moon in 1969, and thus one of the first whose life depended intimately on the correct operation of a digital computer (Tomayko 2000).

The NASA tests were successful, but American aircraft companies were slow to adopt the new technology. The European consortium Airbus, however, embraced it, beginning in the late 1980s with the Airbus A-320. Aircraft do not require, as the Lunar Module did, such "fly-by-wire" controls, but by using a computer, the A-320 had better comfort and better fuel economy than competing aircraft from American suppliers Boeing and McDonnell-Douglas. Fly-by-wire, along with "glass cockpits" (instrument panels that use computer displays) are now commonplace among all new commercial, military, and general aviation aircraft. The Space Shuttle, too, uses fly-by-wire controls in its design, as without such controls it would be impractical to have a human pilot fly it to an unpowered, precise landing on a runway after entering the atmosphere at over 27,000 kilometers per hour.

Another direct influence of the Air Force and NASA on computing was the development of Computer Aided Design (CAD). Air Force funding supported an effort at the Massachusetts Institute of Technology (MIT) that led to the control of machine tools by a sequence of digital computer controls, coded as holes punched into a strip of plastic tape. The results of this work transformed machine tooling, not just for aerospace, but for metalworking in general. At the same time, NASA engineers, working at the various centers, had been using computers to assist in stress analysis of rockets and spacecraft. Launch vehicles had to be strong enough to hold the fuel and oxygen, as well as support the structure of the upper stages, while enduring the vibration and stress of launch, and they had to be light-weight. Aircraft engineers had grappled with this problem of stress analysis for decades; in a typical aircraft company, for every aerodynamicist on the payroll there might have been ten engineers involved with stress analysis. Their job was to ensure that the craft was strong enough to survive a flight, yet light enough to get off the ground. NASA funded computer research in this area, and among the results was a generalized stress analysis program called "NASTRAN"—an shortening of "NASA Structural Analysis" and based on the already-popular FORTRAN programming language. It has since become a standard throughout the aerospace industry.

One obvious issue that arose in the transfer of fly-by-wire to commercial aircraft from Project Apollo was the issue of reliability, already mentioned. The invention of the silicon chip, combined with the Air Force's High-Reliability initiatives, went a long way in making computers reliable for aerospace use, but reliability was still an issue. If the Apollo computers failed in flight, the astronauts could be guide home by an army of ground controllers in Houston. No Apollo computer ever failed, but during the Apollo 13 mission in 1970, the spacecraft lost most of its electrical power, and the crew was indeed saved by ground controllers (Kranz 2000). During the first moon landing—Apollo 11 in 1969—the crew encountered a software error as they descended to the surface; this was resolved by ground controllers, who advised the crew to go ahead with a landing. Having a battery of ground controllers on call for every commercial flight is obviously not practical. Likewise the Space Shuttle, intended to provide routine access to space, was designed differently. For the A-320, Airbus devised a system of three, identical computers, which "vote" on every action. An in-flight failure of one computer would be outvoted by the other two, and the craft can land safely. The Shuttle has five—the failure of one Shuttle computer would allow the mission to continue. The fifth computer is there in case of a software error—it is programmed by a different group of people, so there is little chance of all five computers having a common "bug" in their software (Tomayko 1987, 85–133). This type of redundancy has become the norm in aircraft design. Many spacecraft adopt it, too, but in more nuanced ways, especially if the craft is not carrying a human crew.

Whereas the onboard computing capabilities of commercial aircraft have transformed the passenger jet, the situation on the ground has not progressed far beyond the vacuum-tube age. Commercial air traffic is very safe, and its safety depends on air traffic controllers directing traffic through virtual highways in the sky. Because the US was a pioneer in this activity, it accumulated a large investment in a technology that relies on relatively old-fashioned mainframe computers on the ground, with communications to and from the pilots via VHF radio operating in classical AM voice mode—likewise old fashioned technology. The advent of the Global Positioning System (GPS)—as good an example of the power of Moore's Law as any—should allow air traffic controllers to dispense with much of this infrastructure, and replace it with onboard information sent directly to pilots from satellites. In other words, rather than have ground controllers keep track of the location and route of a plane, the pilots themselves will do that, in a method that does not compromise safety yet increases the capacity of the airways. The pilots would obtain information about their location, and the location of any potentially interfering traffic, using onboard computers that process data from the constellation of GPS or other navigation satellites, plus other satellites and a few select ground stations. That is beginning to happen, but the continental US may be the last to fully adopt it.

Airbus A-320. The computer controls developed to land the Lunar Module were transferred to commercial aircraft. The Airbus A-320, which entered service in the mid-1980s, was the first to adopt this—fly by wire—technology. It is now standard on nearly all commercial jets. Lufthansa.

If there is a common theme among these stories, it is that of how best to utilize the capabilities of the human versus the capabilities of the computer, whether on the ground, in the air, or in space. That issue is never settled, as it is affected by the increasing sophistication and miniaturization of computers, which obviously imply that the craft itself can take on duties that previously required humans. But it is not that simple. Ground-based computers are getting better, too. Human beings today may have the same physical limits and needs as the Apollo astronauts, but they have a much more sophisticated knowledge of the nature of space flight and its needs.

The needs of Aerospace computing

At this point it is worthwhile to step back and examine some specific aspects of space flight, and how "computing," broadly defined, is connected to it.

The Wright brothers' patent for their 1903 airplane was for a method of control, not lift, structure, or propulsion. Spacecraft face a similar need. For spacecraft and guided missiles, control is as important as rocket propulsion. Guided missiles are controlled like airplanes, although without a human pilot. Spacecraft face a different environment, and their control needs are different. An aircraft or guided

World-War II era Air Traffic Control Console, in use into 1990s. The United States' Air Traffic Control system shared many of the characteristics of the SAGE air-defense system. Because of the large initial investment made by the United States in these systems, they remained in use even though they were technically obsolete. This round radar screen, derived from a World War II era defense system, was being used for U.S. commercial air traffic control into the 1990s. Smithsonian Institution.

missile must operate its engines constantly, to work against atmospheric drag, while the forward motion of the wings through the air generates lift to counter the force of gravity. A rocket, by contrast, counters the force of gravity not by lift but by the direct application of thrust. And once a spacecraft enters space, there is little or no atmospheric drag. At that point, its rocket engines are shut off. Thus for many space missions, the rocket motors are active for only a fraction of the total mission time. A spacecraft still requires control, however, but in a different way depending on the phase of its mission. During the initial phase of powered flight, which may last only a few minutes or less, the critical issue is to align the thrust vector of the rocket against the launch vehicle's center of gravity. The configuration of most rockets, with their engines at the bottom and the fuel tanks and payload above, is unstable. The vehicle "wants" to topple over and will do so in an instant, unless its thrust is actively and constantly guided as it ascends. Once that first-order stability is achieved, the vehicle's guidance system may direct the thrust to deviate from that alignment—at first slightly, then more and more as it gains velocity. That will cause the vehicle to tilt, eventually to an optimum angle where the rocket's thrust not only counters gravity but also propels it horizontally: to achieve orbit, to return to Earth some distance away, or to escape the Earth entirely.

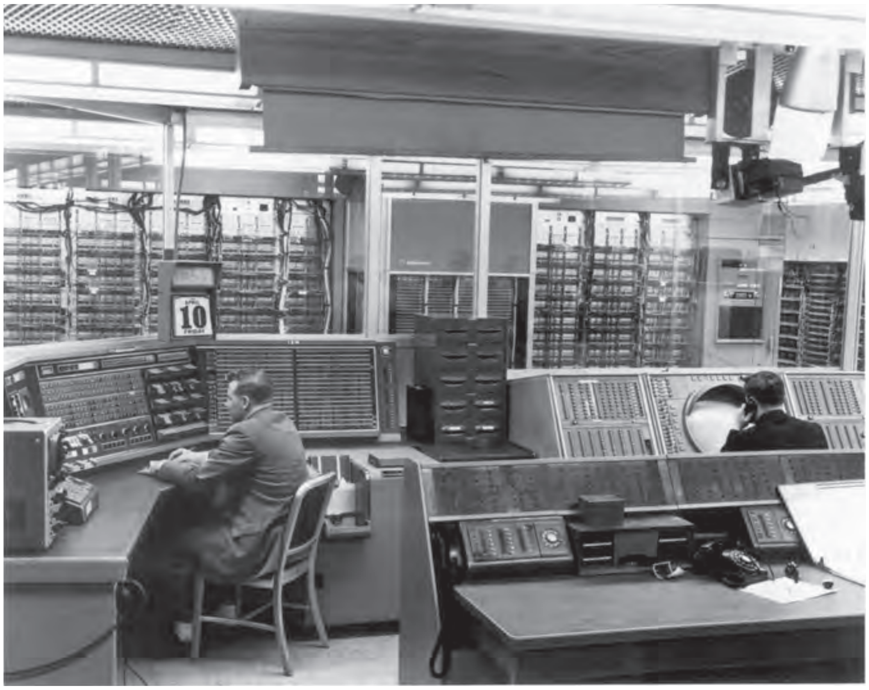
Controlling a rocket's thrust in this, the powered phase of a mission, we call "guidance," although the aerospace community does not always agree on the definition of this term. Note also that this form of guidance is also required for nearly the whole trajectory of an air-breathing guided missile, which is powered through most of its flight.

Once a spacecraft reaches its desired velocity, it may coast to its destination on a "ballistic" trajectory, so-called because its path resembles that of a thrown rock. This assumes that the desired velocity was correct at the moment the engines were cut off. If not, either the main engines or other auxiliary engines are used to change the craft's trajectory. This operation is typically called "navigation," although once again it is not strictly defined. Again in contrast to ships at sea or aircraft on long-distance missions, a spacecraft may fire its onboard rockets only occasionally, not continuously (ion and electric propulsion systems are an exception to this rule). But the process is the same: determine whether one is on a desired course, and if not, fire the onboard engines to change the velocity as needed.

Finally, a craft operating in the vacuum of space feels no atmospheric forces. Once the rocket motors have shut down, it is free to orient itself in any direction and will fly the same no matter how it is pointed. In practice a mission requires that a craft orient itself in a specific way: to point its solar panels at the Sun, to point a camera to a spot on Earth, to aim an antenna, etc. The process of orienting a spacecraft along its x, y, and z axes in space we will call the "control" function. Spacecraft achieve control by using rocket motors with very small thrust, by magnetic coils, momentum wheels, gravity-gradient, or other more exotic devices. The term "control" also encompasses operational aspects of a space mission, such as turning on a camera, activating an instrument, preparing a vehicle for capture by another planet, etc. These actions can be done automatically, by crew members onboard, or from "mission control" stations on the ground.

The Wright brothers' aircraft was unstable by design and required constant attention from its pilot. Moving the horizontal stabilizer to the rear of an airplane provided greater stability; just as tail feathers stabilize an arrow. But controlled aeronautical flight was still difficult. To assist a pilot in maintaining control, in the early twentieth century the American inventor, Elmer Sperry, devised a system of gyroscopes, which augmented the inherent stability of the airplane and reduced the workload on the pilot. This combination of aft-placement of aerodynamic control surfaces, plus a self-correcting system based on gyroscopes, was carried over into rocket research and development. Elmer Sperry's original insight, much extended, is still found at the heart of modern rocket guidance systems. Of those extensions, one was especially important for rocket guidance and came from the German V-2 program: the design of a "pendulous" gyro to measure the time integral of acceleration, which (by Newton's calculus) indicates the craft's velocity (MacKenzie 2000).

During the powered phase of flight, guidance must be performed at speeds commensurate with the action of the rocket. That precludes any off-line work done by humans stationed at the launch point, other than simple decisions such as to destroy a rocket that is going off course. Control functions may also be performed by onboard systems, but if there is no urgency to orient a craft, that can be done by commands from the ground. Navigation often can proceed at a slower pace, with time to process radar or telemetry data through powerful mainframe computers, which can then radio up commands as needed. Thus, while guidance is typically performed by onboard gyroscopes and accelerometers operating with no external communication in either direction, navigation and control may combine signals from onboard systems with radio signals to and from ground stations. Some early ballistic missiles were also guided by radio from the ground, although at real-time speeds with no direct human input at launch. This form of radio or beam-riding guidance has fallen from favor.



SAGE Air-Defense Computer, ca. 1959. The SAGE air-defense system was enormously influential in the direction of computing, even if many consider it obsolete by the time it became operational in the late 1950s. It pioneered in the use of graphic displays, in real-time, interactive computer use, and computer networks. Smithsonian Institution.

Translating the signals from an integrating gyro or accelerometer required what we now call "computing." Early systems used electro-mechanical systems of gears and relays. These were analog computers, using a design that was a mirror (or analog) of the flight conditions it was to control. The V-2, for example, used a pendulous gyro to compute the integral of acceleration, thus giving the velocity; at a certain velocity the motor was shut off to hit a predetermined target. The early mechanical or pneumatic devices were later replaced by electronic systems, using vacuum tubes. Vacuum tubes, though fast acting, remained inherently fragile and unreliable, and were only used in a few instances.

Electronic systems became practical with the advent of solid-state devices, beginning with the invention of the transistor and then the Integrated Circuit, as described above. These circuits were not only small and rugged, they also made it possible to design digital, rather than analog, controls and thus take advantage of the digital computer's greater flexibility. Digital technology has completely taken over not only rocketry and space vehicles but also all new guided missiles, as well as commercial and military aircraft. Although properly heralded as a "revolution," the change was slow to happen, with digital controls first appearing only in the mid-1960s with systems like the Gemini onboard computer.

Long before that, however, the digital computer had an impact on flight from the ground. The V-2 operated too rapidly to be controlled—or tracked and intercepted—by a human being during flight. New jet aircraft were not quite as fast but still challenged the ability of humans to control them. Beginning around 1950, it was recognized that the electronic digital computer, located on the ground where its size and weight were of less concern, could address this problem. Project Whirlwind at MIT successfully tracked and directed an Air Force plane to intercept another aircraft over Cape Cod in April 1951. Whirlwind led to SAGE, an acronym for "Semi-Automatic Ground Environment." SAGE was a massive system of radars, computers, and communications links that warned the US of any flights of Soviet bombers over the North Pole. Critics have charged that SAGE was obsolete by the time it was completed, as the ballistic missile replaced the bomber as a method of delivering a weapon. SAGE could not defend against ballistic missiles, but the system was the inspiration for many ground-control systems, including those used today by the US Federal Aviation Administration to manage commercial air traffic (Ceruzzi 1989).

American Airlines manual reservation system, ca. 1957. Another spin off of SAGE technology was for airline reservations. American Airlines and IBM developed the "SABRE" system, still in use, to replace the labor-intensive manual methods shown here. American Airlines.

By the 1960s, space operations were tightly bound to the ground. The initial design of Project Mercury, for example, had the astronaut simply along for the ride, with ground stations scattered across the globe doing all the mission control. The first Project Mercury capsules did not even have a window. From that beginning, manned spacecraft gradually acquired more onboard control and autonomy, but no crewed spacecraft to this day is ever allowed to operate without inputs from mission controllers on Earth. The rescue of the Apollo 13 crew in 1970 drove home the importance of ground control. Today, most space operations, from the piloted Shuttle and Space Station, to commercial communications satellites, to unmanned military and scientific spacecraft, require more ground-control facilities than commercial or military aviation.

SAGE was designed to look for enemy aircraft. A decade later the US began the development of BMEWS (Ballistic Missile Early Warning System), to provide a warning of ballistic missiles. Air defenses for the continent were consolidated in a facility, called NORAD, at Colorado Springs, Colorado, where computers and human beings continuously monitor the skies and near-space environment. Defense against ballistic missiles continues to be an elusive goal. At present these efforts are subsumed under the term National Missile Defense, which has developed some prototype hardware. A few systems designed to intercept short-range missiles have been deployed at a few sites around the world. Computers play a crucial role in these efforts: to detect launches of a missile, to track its trajectory, to separate legitimate targets from decoys, and to guide an interceptor. These activities require enormous computational power; they also require very high computation speeds as well. Missile defense pushes the state of the art of computing in ways hardly recognized by consumers, however impressive are their portable phones, laptops, and portable media players.

Similarly elaborate and expensive control systems were built for reconnaissance and signals-intelligence satellites. Although the details of these systems are classified, we can say that many US military systems tend to be controlled from ground facilities located near Colorado Springs, Colorado; human spaceflight from Houston, Texas; and commercial systems from various other places in the country. All may be legitimately called descendants of Project Whirlwind.

One final point needs to be made regarding the nature of ground versus onboard spacecraft control. SAGE stood for "Semi-Automatic Ground Environment." The prefix "semi" was inserted to make it clear that human beings were very much "in the loop"—no computer system would automatically start a war without human intervention. An inertial guidance system like that used on the Minuteman is completely automatic once launched, but prior to launch there are multiple decision points for human intervention. Likewise in the human space program, the initial plans to have spacecraft totally controlled from the ground were not adopted. Project Mercury's initial designs were modified, first under pressure from the astronauts, and later more so after the initial flights showed that it was foolish to have the astronaut play only a passive role. A desire for human input is also seen in the controls for the Space Shuttle, which cannot be operated without a human pilot.

The Future

It should be clear from the above discussion that a simple comparison of the advances in computing and advances in space travel since 1958 is not possible. Nevertheless, the producers of the television show "Nerds 2.0.1" made a valid point. The Internet has enjoyed a rapid diffusion into society that aerospace has not been able to match. A factor not mentioned in that show, but which may be relevant, is an observation made by networking pioneer Robert Metcalfe. According to Metcalfe (and promoted by him as "Metcalfe's Law" as a counterpart to Moore's Law), the value of a network increases as the square of the number of people connected to it. Thus the Internet, which adds new connections every day, increases in value much faster than the cost of making each of those new connections. Space exploration has no corresponding law, although if deep space probes discover evidence of life on other planets, that equation will be rewritten.

One facet of history that is forgotten when writing about the Internet is that the aerospace community was among the world's pioneers in computer networking, but to different ends. The SAGE system was the world's first large-scale computer network, for example. And the first use of a computer network for private, as opposed to military or government use, was the airline reservations system "SABRE," developed by IBM in the early 1960s for American Airlines. These networks were significant but were not the technical antecedents of the Internet. In fact, the ARPANET was developed in partial response to the deficiencies of SAGE. In the latter system, the entire network could have been rendered inoperative if a central control node were destroyed; with the Internet that cannot happen as it has no central control point, by design. The Internet's ability to link disparate computer systems by a set of common protocols likewise sets it apart from aerospace networks, which often are unable to communicate with one another. An embarrassing example of this happened recently during the development by Airbus of its superjumbo transport, the Airbus A-380. Airbus made heavy use of a CAD program called "CATIA," developed by the French aerospace company Dassault Systemes. CATIA allowed engineers from different laboratories and plants to work to a common set of virtual "drawings," as if they were in the same building. For the A-380, one group of designers was using a different version of CATIA to the others, and when the parts were brought together for final assembly at the Airbus plant in Toulouse, France, they did not fit. Boeing has likewise experienced problems integrating assemblies from different places for its new jet, the 787 Dreamliner. In fairness to Airbus and Boeing, the Internet, as it is presently configured, would be unable to handle the complexities of designing a modern airplane, in spite of its ability to scale up to large numbers of nodes from all over the world.

Was NASA's Project Apollo a technological dead-end, however impressive an engineering accomplishment it was? And was the network developed by NASA's companion agency, ARPA, the true defining technology of the modern age? Neither question admits of an easy answer. The two technologies have grown in a symbiotic relationship with each other, and they will continue to do so in the future. The notion of a computer as an artificially-intelligent agent in service to humanity has given way to a notion of the computer as a device to "augment human intellect," in the worlds of computer pioneer Douglas Engelbart. Engelbart is best known for his invention of the mouse as a computer pointing device, but he is also known as one of the first to recognize this place for computers among us. Before inventing the mouse, Engelbart worked at the NASA Ames Research Center in Mountain View, California, and later for a commercial computer networking company owned by the aerospace company McDonnell-Douglas. So he was no stranger to the real-world limitations, and potential, of networked computing, and to aerospace applications.

The limitations of the human body will remain as a drag on progress in the human exploration of deep space. Given the laws of physics as we currently know them, it is difficult to envision human travel beyond the orbit of Mars with even the most optimistic extrapolations of current chemical rocket propulsion. One intriguing way out of this dilemma is suggested by Moore's Law. If current trends continue, computers will contain the equivalent number of circuits as there are neurons in the human brain by about the year 2030. If one assumes an equivalence, then one could envision transferring the nature of human consciousness to a computer, which could then explore the cosmos unconstrained by a human body that currently is required to support it. This is the argument made by inventor Ray Kurzweil, who believes such a transfer of consciousness is inevitable (Kurzweil 1999). Of course the assumption of equivalence makes all the difference. We have already seen how the early predictions of artificially intelligent computers fell short. Having more and more circuits may not be enough to cross the threshold from "intelligence," however defined, to "consciousness." In this area it is best to leave such speculation to the science fiction writers. One may feel disappointed that the human exploration of space seems to be so constrained, but it is hard to maintain that feeling in the face of all the other exciting developments in aerospace that are happening all around that fact.

Bibliography

Abbate, Janet. *Inventing the Internet.* Cambridge, Massachusetts: MIT Press, 1999.

Ceruzzi, Paul. *Beyond the Limits: Flight Enters the Computer Age.* Cambridge, Massachusetts: MIT Press, 1989.

—, *A History of Modern Computing.* Cambridge, Massachusetts: MIT Press, 1998.

Dick, Steven J., y Roger Launius, eds. *Societal Impact of Spaceflight.* Washington, DC: NASA, 2007.

Kranz, Gene. *Failure is not an Option.* New York: Berkeley Books, 2000.

Kurzweil, Ray. *The Age of Spiritual machines: When Computers Exceed Human Intelligence.* New York: Viking, 1999.

Mackenzie, Donald. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance.* Cambridge, Massachusetts: MIT Press, 1990.

McCorduck, Pamela. *Machines Who Think: A personal Inquiry into the History and Prospects of Artificial Intelligence.* San Francisco, California: W.H. Freeman, 1979.

Mindell, David. *Digital Apollo: Human and Machine in Spaceflight.* Cambridge, Massachusetts: MIT Press, 2000.

Moore, Gordon. "Cramming More Components onto Integrated Circuits." *Electronics*, April 19, 1965, 114-117.

Noble, David F. *Forces of Production: A Social history of Industrial Automation.* New York: Oxford University Press, 1986.

Norberg, Arthur, y Judy O'Neill. *Transforming Computer Technology: Information Processing for the Pentagon, 1962-1986*. Baltimore, Maryland: Johns Hopkins University Press, 1996.

Segaller, Stephen. *Nerds: A Brief History of the Internet.* New York: TV Books, 1998.

Tomakyo, James E. *Computers in Spaceflight: the NASA Experience.* New York: Marcel Dekker, Inc., 1987.

—, *Computers Take Flight: A History of NASA's Pioneering Digital Fly-By-Wire Project.* Washington, DC: NASA, 2000.

the internet: global evolution and challenges

JANET ABBATE

The Internet, a global network of networks, is a remarkably complex technical system built on the creative contributions of scientists around the world from the 1950s to the present. Throughout its evolution, the Internet and other networks have been promoted by governments, researchers, educators, and individuals as tools for meeting a range of human needs. A combination of high-level policy and grassroots improvisation has produced social benefits including easier and more widespread access to computers and information; increased scientific collaboration; economic growth; the formation of *virtual communities* and an increased ability to maintain social ties over long distances; the democratization of content creation; and online political and social activism. The Internet's rapid growth has also spawned technical crises, such as congestion and a scarcity of network addresses, and social dilemmas, including malicious and illegal activities and persistent *digital divides* based on income, location, age, gender, and education. Such problems continue to demand creative solutions from scientists, policy makers, and citizens.

Several general themes characterize the technical development of the Internet. First, from the 1950s to the present there has been a steady increase in the size of data networks and the variety of services they offer. Rapid growth and diversity have forced network designers to overcome incompatibilities between computer systems and components, manage data traffic to avoid congestion and chaos, and reach international agreement on technical standards. These challenges have led to fundamental advances in research areas such as operating systems and queuing theory. A second trend has been the modeling of network functions as a series of *layers*, each of which behaves according to a standard *protocol*, a set of rules for interaction that is implemented in software or hardware. Layering reduces the complexity of the network system and minimizes the amount of standardization necessary, which makes it easier for networks to join the Internet. A third important feature of the Internet's technical development has been an unusually decentralized and participatory design process. This has opened the system to innovation from a variety of directions and has encouraged informal worldwide collaboration. The following sections describe some of the major milestones in the evolution of the Internet and its predecessors.

Beginnings: early terminal networks

The first electronic digital computers, invented during the World War II and commercialized immediately afterward, were solitary machines: they were not designed to interact with their human users or to communicate with other computers. Within a few years, however, computer scientists began experimenting with ways to access computers from a distance or transmit data from one machine to another. The data networks of the 1950s and early 1960s were systems to connect terminals to computers, rather than connecting computers to each other. Experiments with terminal networks provided an intriguing research area for computer scientists, but they were also a response to contemporary political and economic realities, including the Cold War and the growth of global economic, transportation, and communication networks.

Computer science research in the United States was largely funded by the military and reflected that country's rivalry with the USSR. For example, an important US development of the 1950s was Project SAGE, a computerized early-warning defense system designed to detect missile attacks. Each SAGE center had an IBM computer that received data through telephone lines from dozens of radar installations and military bases. A key technology developed for SAGE by AT&T Bell Laboratories was the *modem*, which converts digital computer data into analog signals that can be sent over the telephone network. AT&T began to offer modems for general use in 1958, and for several decades modems would provide the chief means of network access for home users.

Demand for terminal networks was driven by another technical milestone of the early 1960s: *time sharing* operating systems. Invented independently in 1959 by Christopher Strachey of the UK and John McCarthy of the US, time sharing allowed multiple users to run programs on a single computer simultaneously. Because the cost of the computer could be shared among a much larger number of users, time sharing made it practical to allow individuals to use a computer interactively for long stretches of time, rather than being restricted to running a single program and receiving the results offline. Commercial time sharing services took advantage of these economies of scale to provide affordable computing to many academic and business customers. By the mid-1960s, commercial time sharing services were developing their own data networks to give their customers low-cost access to their computers.

Global capitalism and the growth of transportation and communication systems provided the impetus for large-scale commercial terminal networks. In the early 1960s, data-intensive industries, such as aviation and stock trading, built cooperative networks to enable firms to share a common pool of information. For example, in the early 1960s American Airlines and IBM created the SABRE on-line reservation system (based on IBM's work on SAGE), which connected 2,000 terminals across the United States to a central computer. Similarly, the US National Association of Securities Dealers Automated Quotation System (NASDAQ) created a network for stock quotations in 1970. In an early example of international collaboration in networking, a cooperative of airlines called SITA (Société Internationale de Télécommunications Aéronautiques) built a network in 1969 using the *packet switching* technique (see below). The SITA network handled traffic for 175 airlines through computer centers in Amsterdam, Brussels, Frankfurt, Hong Kong, London, Madrid, New York, Paris, and Rome (SITA, 2006). Such financial and commercial networks helped accelerate the integration of the global economy.

Research networks

Terminal networks were based on a relatively simple hub-and-spoke model that connected numerous users to a single central computer resource. More complex networks involving multiple computers were built by computer scientists from the late 1960s to the late 1970s. Experimenting with new technologies, researchers aimed to break the barriers to sharing data between dissimilar computer systems. Scientists and their government sponsors saw a threefold promise in networking: the ability to share scarce and expensive computers, which would increase access while decreasing costs; the ability to share data and work collaboratively with colleagues in other locations; and the opportunity to advance the theory and practice of computer science.

Three of the most important early research networks were the ARPANET (US, 1969), the NPL Mark I (UK, 1969), and CYCLADES (France, 1972). A key innovation of these experimental networks was a communications technique called *packet switching*. Previous communication systems, such as the telephone and the terminal networks, provided dedicated circuits between the two ends of a connection. In contrast, a packet switching network divides the data to be transmitted into small units called *packets* that are sent out individually, sharing the network circuits with packets from other connections. Packet switching allows communications links to be used more efficiently, thus conserving an expensive resource. In addition, packets from the same connection can be sent to their destination by different routes, making it possible to distribute traffic among multiple links or respond to a breakdown in one part of the network by routing traffic elsewhere. This flexibility helps prevent congestion and increases the reliability of the network.

The concept of packet switching was invented independently in early 1960s by Paul Baran of the US and Donald Davies of the UK; Davies put the technique into practice in the one-node Mark I network at the National Physical Laboratory. In the US, the Defense Advanced Research Projects Agency (DARPA) sponsored the first large-scale packet switching network, ARPANET. One of the theorists contributing to this project was Leonard Kleinrock, who developed some of the first methods for analyzing packet network behavior. In France, Louis Pouzin pioneered connectionless or *datagram* networking techniques in the packet-switched CYCLADES network. Datagram networks were simpler than connection-oriented networks such as ARPANET, and this simplicity made it more feasible to interconnect different networks—an important step toward developing a worldwide Internet. As Pouzin noted: "The more sophisticated a network, the less likely it is going to interface properly with another." (Pouzin 1975, 429.) Experiments in *internetworking* (connecting multiple networks) were already taking place by the early 1970s. For example, the NPL network was connected to CYCLADES in 1974, and in 1976 both CYCLADES and NPL were connected with the new European Informatics Network. EIN had grown out of a 1971 science and technology study group of the European Economic Community (now the European Union), which recommended the building of a multinational network to help member countries share computer resources and promote computer science research. By 1976 the EIN was providing network service to ten countries, with hubs in Italy, France, Switzerland, and the United Kingdom (Laws and Hathway 1978). The convergence of networking systems thus mirrored the political convergence of the cooperating states.

A number of experimental techniques besides packet switching were featured in the ARPANET. This network connected researchers across the United States working in areas such as time sharing, artificial intelligence, and graphics; because of generous government funding and the large pool of computer science talent involved, the ARPANET builders were able to experiment with promising but extremely challenging techniques. For example, rather than limiting the network to a single type of computer, as had most other experiments in computer-to-computer communication, the ARPANET included a variety of extremely diverse computers. This drove the team of computer scientists, graduate students, and industry engineers to find ways of bridging the incompatibilities between computers, and their hard work made it much easier to build the next generation of networks. The ARPANET also had a *distributed* topology featuring many switching nodes with multiple interconnections, rather than a single central node. Distributed communications, first described by Baran (1964), could spread out the traffic load and potentially increase reliability by creating multiple paths between any two computers. However, adopting this untried technique greatly increased the complexity of the routing system, forcing the ARPANET designers to analyze and manage unexpected network behavior. In another risky move, the network design called for the routing operations to be decentralized and *adaptive*: each node would make its routing decisions independently and would change its behavior in response to changes in traffic conditions or network configuration (for example, if an adjacent node became disabled). The ARPANET's decentralized design and autonomous routing behavior increased the difficulty of analyzing network behavior; at the same time, these techniques would contribute to the future success of the Internet, because they would allow the network to grow without being limited by a central bottleneck.

One of the most novel features of the ARPANET project was not technical but organizational: an informal, decentralized decision-making process. The network software was developed by a loose confederation of researchers and students called the Network Working Group. Any member of the group could suggest a new feature by circulating a *Request For Comments*; after a period of discussion and trial implementations, the suggestion would be modified, abandoned, or adopted by consensus as a network standard. This collaborative process continues to be used for Internet standards (Bradner 1996) and has helped the system grow and adapt by encouraging free debate and wide participation in its technical development.

By far the most successful application of the early research networks was electronic mail, which became a standard service in the early 1970s. The popularity of email came as a surprise to the ARPANET builders, who had expected that research-oriented networks would focus on sophisticated, computationally-intensive applications such as mathematics or graphics. While email was adopted in part because it was simple to use, its popularity also reflected the realization that scientific research depended as much on human collaboration as on access to machines. Email provided an unprecedented opportunity for ongoing interaction with remote colleagues.

Though they were not open to the general public, the early research networks went beyond providing computer access for a small group of scientists. They produced solutions to formidable technical obstacles and established vital resources for future innovation, including standard techniques and a community of researchers and engineers experienced in networking (Quarterman 1990). Early efforts to build multi-national networks and internets also sowed the seeds of global cooperation, without which today's Internet could not exist.

Expanding access: proprietary, public, and grassroots networks

In the mid-1970s, the emergence of research networks was paralleled by three other trends: proprietary networking systems offered by computer manufacturers; public data networks built by national telecommunications carriers (PTTs); and grassroots networks that were improvised by individuals with little funding. Companies such as IBM had provided limited networking capabilities since the 1960s, but after the research networks had demonstrated the viability of packet switching, computer firms began offering their own packet-switching technologies. Widely used systems included IBM's Systems Network Architecture (1974), Xerox Network Services (1975), and Digital Equipment Corporation's DECNET (1975). Unlike research networks, these proprietary systems had many corporate users. Corporate networks enabled businesses to be both more distributed —because branch operations could access the data they needed to operate independently—and more centralized, because data from far-flung operations could be instantly monitored by the head office. Thus computer networking reflected and augmented the trend toward economic globalization that accelerated in the 1980s and beyond.

While proprietary systems provided a vital service to organizations with many computers from the same manufacturer, these networks were generally not compatible with computers from rival manufacturers. This could be a problem within a single organization and certainly raised an obstacle to building a national or international network. In addition, these commercial systems were under the control of private corporations and did not adhere to publicly established technical standards. This was of particular concern outside the United States, where most of the large computer manufacturers were located. To provide the public with an alternative, in 1974–75 the national telecommunications carriers in Europe, Canada, and Japan announced plans to build data networks that would be available to any user, regardless of the brand of computer they used.

The PTTs' vision of data networking, modeled on the phone system, included not only universal access but also international connections. Realizing that this would require agreement on a shared network protocol, in 1975–76 the Consultative Committee on International Telegraphy and Telephony of the International Telecommunications Union developed a packet-switching network standard called X.25. X.25 provided a reliable connection called a *virtual circuit* between two points on a network, allowing terminal users to access online resources without having to install complex networking software. Early adopters of the new standard included Canada's Datapac network (1977), France's Transpac (1978), Japan's DDX (1979), the British Post Office's PSS (1980), and the multinational Euronet (1979). While X.25 was later superseded by other technologies such as *frame relay*, it provided a base for the rapid development of public networks around the world and avoided the chaos of competing incompatible standards. Another influential standards effort in the late 1970s was the Open Systems Interconnection model created by the International Standards Organization. This defined the functions for seven layers of network services, ranging from low-level hardware connections to high-level applications and user interfaces. Although there was much debate over these standards (Abbate 1999), adopting a common model helped computer scientists and manufacturers move closer to creating fully interoperable network systems.

Public data networks provided the first online access for much of the world's population. They also sponsored new types of content and services that made data networks relevant to non-technical users. For example, in the early 1980s France Telecom achieved widespread public use of its Transpac network by offering the innovative Minitel system: a free terminal, given to customers in place of a telephone directory, with access to a free online directory and a variety of paid services. Minitel was in use for almost three decades and served nearly half the French population. With payments securely handled by the phone company, Minitel provided some of the world's first e-commerce, including airline and train ticketing, mail-order retail, banking and stock trading, information services, and message boards (McGrath 2004).

The development of public data networks reflected an emerging view—by both individual users and the highest levels of government—that access to computer communications was a public good, a resource that would be necessary for full citizenship in the twenty-first century. In serving this mission, public data networks were complemented by a third trend of this period: improvised grassroots networks. These low-cost networks used existing software and simple dial-up connections to exchange mail and discussion lists among an informal community of users. The most well-known were USENET, which was established in 1979 using UNIX protocols, and BITNET, created in 1981 using IBM protocols. These networks played an important role in providing communication to people who had no access to formal networking infrastructure.

Designing the Internet

How did these disparate data communications systems become united into the global network that we know as the Internet? While some connections between networks were established in the 1970s, design incompatibilities generally limited their services to the exchange of mail and news. The technologies that allow the full range of network services to be shared seamlessly across systems were initially created for the ARPANET. DARPA's explorations in internetworking stemmed from its desire to connect the ARPANET with two new networks it had built, which extended packet switching techniques to radio and satellite communications. Since these media did not have the same technical characteristics as telephone lines—radio links were unreliable; satellites introduced delays—existing techniques such as X.25 or the original ARPANET protocols were not suitable for such a diverse interconnected system. In the early 1970s, therefore, DARPA started an Internet Program to develop a more comprehensive solution.

Another technical development that helped drive the demand for internetworking was *local area networks*. Ethernet, the most influential of these, was invented in 1973 by Robert Metcalfe, drawing on an earlier network called Alohanet that was created by Norman Abramson, Frank Kuo, and Richard Binder (Metcalfe 1996; Abramson 1970). Ethernet and Alohanet pioneered a technique called *random access* that allowed many users to share a communication channel without the need for complex routing procedures.[1] The simplicity of the random access design helped make LANs affordable for a broad range of users. Ethernet became formally standardized and commercially available in the early 1980s and was widely adopted by universities, businesses, and other organizations. Another popular LAN system, token ring, was invented by IBM researchers in Zurich and commercialized in 1985. The popularity of LANs would create many new networks that potentially could be interconnected; but, like the packet radio network, these random access systems could not guarantee a reliable connection, and therefore would not work well with existing wide-area network protocols. A new system was needed.

The Internet Program was led by Vinton Cerf and Robert Kahn, with the collaboration of computer scientists from around the world. In addition to US researchers at DARPA, Stanford, the University of Southern California, the University of Hawaii, BBN, and Xerox PARC, Cerf and Kahn consulted networking experts from University College London, the NPL and CYCLADES groups, and the International Network Working Group (Cerf 1990). The INWG had been founded in 1972 and included representatives from many national PTTs that were planning to build packet-switching networks. By sharing concerns and pooling ideas, this inclusive team was able to design a system that could serve users with diverse infrastructural resources and networking needs.

The Internet architecture had two main elements. The first was a set of protocols called TCP/IP, or Transmission Control Protocol and Internet Protocol (Cerf and Kahn 1974).[2] TCP was an example of a *host protocol*, whose function is to set up and manage a connection between two computers (*hosts*) across a network. The insight behind TCP was that the host protocol could guarantee a reliable connection between hosts even if they were connected by an unreliable network, such as a packet radio or Ethernet system. By lowering the requirement for reliability in the network, the use of TCP opened the Internet to many more networks than it might otherwise have accommodated. To ensure dependable connections, TCP was designed to verify the safe arrival of packets, using confirmation messages called *acknowledgments*; compensate for errors by retransmitting lost or damaged packets; and control the rate of data flow between the hosts by limiting the number of packets in transit. In contrast, the Internet Protocol performed a much simpler set of tasks that allowed packets to be passed from machine to machine as they made their way through the network. IP became the common language of the Internet, the only required protocol for a network wishing to join: member networks had the freedom to choose among multiple protocols for other layers of the system (though in practice most eventually adopted TCP for their host protocol). Reflecting the diverse needs and preferences of the experts who participated in its design, the Internet architecture accommodated variation and local autonomy among its member networks.

The second creative element was the use of special computers called *gateways* as the interface between

[1] Metcalfe's improved version of the random access system was called Carrier Sense Multiple Access with Collision Detection (CSMA/CD).

[2] Originally there was a single protocol, TCP; it was split into two protocols, TCP and IP, in 1980.

different networks (Cerf 1979). Gateways are now commonly known as *routers*; as the name implies, they determine the route that packets should take to get from one network to another. A network would direct non-local packets to a nearby gateway, which would forward the packets to their destination network. By dividing routing responsibility between networks and gateways, this architecture made the Internet easier to scale up: individual networks did not have to know the topology of the whole Internet, only how to reach the nearest gateway; gateways needed to know how to reach all the networks in the Internet, but not how to reach individual hosts within a network.

Another notable invention that would make the worldwide growth of the Internet manageable was the Domain Name System, created in 1984 by Paul Mockapetris (Cerf 1993; Leiner et al 1997). One challenge of communicating across a large network is the need to know the address of the computer at the far end. While human beings usually refer to computers by names (such as "darpa"), the computers in the network identify each other by numerical addresses. In the original ARPANET, the names and addresses of all the host computers had been kept in a large file, which had to be frequently updated and distributed to all the hosts. Clearly, this mechanism would not scale up well for a network of thousands or millions of computers. The Domain Name System decentralized the task of finding addresses by creating groups of names called *domains* (such as .com or .org) and special computers called *name servers* that would maintain databases of the addresses that corresponded to each domain name. To find an address, the host would simply query the appropriate name server. The new system also made it possible to decentralize the authority to assign names, so that, for example, each country could control its own domain.

The World Wide Web and other applications

The Internet architecture made it possible to build a worldwide data communications infrastructure, but it did not directly address the question of content. In the 1980s, almost all content on the Internet was plain text. It was relatively difficult for users to locate information they wanted; the user had to know in advance the address of the site hosting the data, since there were no search engines or links between sites. The breakthrough that transformed how Internet content was created, displayed, and found was the World Wide Web.

The World Wide Web was the brainchild of Tim Berners-Lee, a British researcher at CERN, the international physics laboratory in Geneva. He envisioned the Internet as a collaborative space where people could share information of all kinds. In his proposed system, users could create pages of content on computers called *web servers*, and the web pages could be viewed with a program called a *browser*. The Web would be able to handle multimedia as well as text, and Web pages could be connected by *hyperlinks*, so that people could navigate between sites based on meaningful relationships between the ideas on different pages. This would create a web of connections based on content, rather than infrastructure. Berners-Lee formulated his ideas in 1989, and he and collaborator Robert Cailliau created the first operational version of the Web in 1990. The technical underpinnings of the new system included *html* (hypertext markup language, used to create web pages), *http* (hypertext transfer protocol, used to transmit web page data), and the *url* (uniform resource locator, a way of addressing web pages).

The Web was popular with the physicists who used it at CERN, and they spread it to other research sites. At one such site, the US National Center for Supercomputer Applications, Marc Andreessen led the development of an improved browser called Mosaic in 1993. Mosaic could run on personal computers as well as on larger machines, and NCSA made the browser freely available over the Internet, which led to a flood of interest in the Web. By 1994 there were estimated to be a million or more copies of Mosaic in use (Schatz and Hardin 1994).

The Web's hyperlinks were designed to solve a long-standing problem for Internet users: how to find information within such a large system? To address this need, various finding aids were developed in the 1990s. One of the earliest tools for searching the Internet was Archie (1990), which sent queries to computers on the Internet and gathered listings of publicly available files. Gopher (1991) was a listing system specifically for the Web, while Yahoo (1994) was a directory of Web pages organized by themes. Yahoo's staff categorized Web pages by hand, rather than automatically; given the vast amount of data accumulating on the Web, however, a variety of new services tried to automate searching. The most successful of these *search engines* was Google (1998). Search engines transformed the way users find information on the Web, allowing them to search a vast number of sources for a particular topic rather than having to know in advance which sources might have relevant information.

Like the Internet itself, the Web was designed to be flexible, expandable, and decentralized, inviting people to invent new ways of using it. The spread of the World Wide Web coincided with the transition in 1995 of the US Internet backbone from government to private-sector control. This removed many barriers to commercial use of the Internet and ushered in the "dot-com" boom of the 1990s, in which huge amounts of capital were invested in e-commerce schemes. While the dot-com bubble burst in 2000, it was significant in creating a popular understanding of the Internet as an economic engine and not merely a technical novelty. The beginning of the twenty-first century also saw the proliferation of *social media* that provided new ways for people to interact and share information and entertainment online. These included weblogs (1997), wikis (1995), file sharing (1999), podcasting (2004), social networking sites, and a variety of multi-player games.

The Internet and society: successes and challenges

After half a century of research and innovation, the Internet was firmly established as a widely available resource offering an array of potential benefits. Users had greater access to information of all kinds, and governments and businesses had a new platform for providing information and services. E-commerce brought economic growth, greater choices for consumers, and opportunities for producers in disadvantaged areas to reach new markets. A variety of communications options, from email to elaborate social networking sites, made it easier for friends and family to stay in touch over long distances and for strangers to form "virtual communities" around common interests. Grassroots organizers adopted the Internet for political and social activism and used it to mobilize worldwide responses to natural disasters and human rights abuses. Users of all ages embraced the Internet as a medium for personal expression, and new applications helped democratize the technology by making it easier for ordinary people to independently produce and disseminate news, information, opinion, and entertainment.

However, many challenges remained as the Internet entered the twenty-first century. Users faced abusive practices such as spam (unwanted commercial email), viruses, identity theft, and break-ins. Technical experts responded with solutions that attempted to minimize these ongoing dangers, providing anti-virus systems, filters, secure web transactions, and improved security systems. But other issues were too divisive for a technical solution to satisfy conflicting public opinion, especially when activities crossed national boundaries. Some governments severely limited and closely monitored the online activities of their citizens; while human rights groups protested this as censorship and intimidating surveillance, the governments in question asserted their right to protect public safety and morality. Other groups complained that the Internet was *too* open to objectionable or illegal content such as child pornography or pirated songs, movies, and software. Filters and copyright protection devices provided means to restrict the flow of such information, but these devices were themselves controversial. Internet governance was another thorny issue, with many of the world's nations calling for a more international, less US-dominated mechanism for managing the Internet's name and address system. Another technical issue with political ramifications was the proposed transition from the old Internet Protocol, called IPv4, to a new protocol called IPv6 that would provide a much larger number of addresses (Bradner and Mankin 1995); this was in part a response to the fact that the United States held a disproportionate share of the IPv4 addresses. Ipv6 was proposed as an Internet standard in 1994, but due to technical and political disagreements the protocol was still only used for a tiny percentage of Internet traffic 15 years later (DeNardis 2009). Given these many obstacles, the Internet's decentralized, consensus-based development process continued to work remarkably well to keep the system thriving amid rapid growth and change.

Perhaps most troubling was the persistent inequality of access to the Internet and its opportunities for economic development, political participation, government transparency, and the growth of local science and technology. Significant gaps remained between rich and poor regions, urban and rural citizens, young and old. The United Nations reported in 2007 that the global digital divide was still enormous: "Over half the population in developed regions were using the Internet in 2005, compared to 9 per cent in developing regions and 1 per cent in the 50 least developed countries." (UN, 2007, 32.) To help address this issue, the UN and International Telecommunications Union sponsored a two-part World Summit on the Information Society in Geneva (2003) and Tunis (2005) to devise a plan of action to bring access to information and communication technologies to all of the world's people (WSIS 2008). Computer scientists also devoted their ingenuity to making the Internet more accessible to the world's poor. For example, in 2001 a group of Indian computer scientists reversed the paradigm of expensive, energy-consuming personal computers by creating the Simputer: a simple, low-cost, low-energy computer

[3] "Internationalizing" the governance of the Internet was a central issue at the UN-sponsored World Summit on the Information Society in 2005.

[4] The creators and trustees of the Simputer project were Vijay Chandru, Swami Manohar, Ramesh Hariharan, V. Vinay, Vinay Deshpande, Shashank Garg, and Mark Mathias (http://www.simputer.org/simputer/people/trustees.php).

that would provide a multilingual interface and could be shared among the residents of a village (Sterling 2001). Similarly, Nicholas Negroponte initiated the One Laptop Per Child project in 2005 to serve educational needs in developing countries. To help fit the technology to local needs, lead designer Mary Lou Jepsen invented an inexpensive, power-efficient screen readable in outdoor light, and software designer Walter Bender created an intuitive graphical user interface (One Laptop Per Child 2008; Roush 2008). The Stockholm Challenge, an annual event since 1995, showcases hundreds of innovative projects from around the world that use ICTs to promote development (Stockholm Challenge 2008).

No longer simply the domain of scientists, pushing the frontiers of the Internet increasingly involves social as well as technical innovation and the collaboration of researchers, businesses, civil society organizations, governments, and ordinary people. The values guiding the Internet's social and technical development have been complementary: increasing access, accommodating diversity, decentralizing authority, making decisions by consensus with a wide range of participants, and allowing users to take an active role in adding features to the network. On the technical side, these goals have been achieved through layered architecture, open protocols, and a collaborative process for approving design changes, while social goals have been advanced through government leadership and the inspiration of individuals who saw the Internet's potential for communication, cooperation, and self-expression.

Bibliography

Abbate, Janet. *Inventing the Internet*. Cambridge: MIT Press, 1999.

Abramson, Norman. "The ALOHA System —, Another Alternative for Computer Communications." *Proceedings, AFIPS Fall Joint Computer Conference*. Montvale, NJ: AFIPS Press, 1970, 281–285.

Baran, Paul. *On Distributed Communications.* Santa Monica, CA: RAND Corporation, 1964.

Berners-Lee, Tim. *Weaving the Web*. New York: HarperCollins, 1999.

Bradner, Scott, and A. Mankin. "The Recommendation for the IP Next Generation Protocol." *Network Working Group Request for Comments 1752*, January 1995. Available on the Internet at http://www.rfc-editor.org/

Bradner, Scott. "The Internet Standards Process—Revision 3." *Network Working Group Request for Comments 2026*, October 1996. Available on the Internet at http://www.rfc-editor.org/

Campbell-Kelly, Martin, and William Aspray. *Computer: A History of the Information Machine*. New York: Basic Books, 1996.

Cerf, Vinton G. "DARPA Activities in Packet Network Interconnection." In K. G. Beauchamp, ed. *Interlinking of Computer Networks*. Dordrecht, Holland: D. Reidel, 1979.

Cerf, Vinton G. Oral history interview by Judy O'Neill (Reston, VA, April 24, 1990), OH 191. Minneapolis, MN: The Charles Babbage Institute, University of Minnesota, 1990. Available on the Internet at http://www.cbi.umn.edu/oh/display.phtml?id–118

—, "How the Internet Came to Be." In Bernard Aboba, ed. *The Online User's Encyclopedia*. Addison-Wesley, 1993.

Cerf, Vinton G., and Robert E. KAHN. "A Protocol for Packet Network Intercommunication." *IEEE Transactions on Communications* COM-22, May 1974, 637–648.

Denardis, Laura. *Protocol Politics: The Globalization of Internet Governance.* Cambridge: MIT Press, 2009.

Laws, J., and V. Hathway. "Experience From Two Forms of Inter-Network Connection." In K. G. Beauchamp, ed. *Interlinking of Computer Networks*. NATO, 1978, 273–284.

Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. 1997. "A Brief History of the Internet." Available on the Internet at http://www.isoc.org/internet-history. Revised February 1997.

McGrath, Dermot. "Minitel: The Old New Thing." *Wired,* January 18, 2004. Available on the Internet at http://www.wired.com/science/discoveries/news/2001/04/42943.

Metcalfe, Robert M. *Packet Communication*. San Jose: Peer-to-Peer Communications, 1996.

One Laptop Per Child. http://www.laptop.org/ (accessed September 5, 2008).

Pouzin, Louis. "Presentation and Major Design Aspects of the CYCLADES Computer Network." In R. L. Grimsdale and F. F. Kuo, eds. *Computer Communication Networks*. Leyden: Noordhoff, 1975.

Quarterman, John S. *The Matrix: Computer Networks and Conferencing Systems Worldwide*. Burlington, MA: Digital Press, 1990.

Roush, Wade. "One Laptop Per Child Foundation No Longer a Disruptive Force, Bender Fears." *Xconomy*, April 24, 2008. Available on the Internet at http://www.xconomy.com/boston/2008/04/24/

Schatz, Bruce R., and Joseph B. Hardin. "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet." *Science* 265 (1994), 895–901.

SITA. "SITA's history and milestones." http://www.sita.aero/News_Centre/Corporate_profile/History/ (updated 21 July 2006, accessed September 5, 2008).

Sterling, Bruce. "The Year In Ideas: A to Z; Simputer." *The New York Times*, December 9, 2001.

Stockholm Challenge. www.challenge.stockholm.se <http://www.challenge.stockholm.se> / (accessed September 5, 2008).

United Nations. *The Millennium Development Goals Report 2007.* New York: United Nations, 2007.

World Summit on the Information Society. http://www.itu.int/wsis/ (accessed September 5, 2008).

the century of the gene. molecular biology and genetics

GINÉS MORATA

The twentieth century was the century in which human society incorporated technological development in a massive way. For much of that century, the greatest technological contributions grew out of the physical sciences: automobiles, the telephone, airplanes, plastics, computers, and so on. The introduction of those factors has changed society and human behavior more than even political and social events.

During the second half of the twentieth century, however, and especially during the last two decades, biological technology with enormous medical and social potential has emerged. This technology offers a new image of the evolution of life on our planet and is destined to revolutionize the very structure of human society.

Perhaps the person who has most lucidly delved into these ideas is Sydney Brenner. One of the most brilliant scientists of the twentieth century, he will be remembered by the history of science for his enormous contributions to molecular biology, a science he was decisively involved in creating. Brenner says that new biology offers us greater comprehension of ourselves, and a new understanding of humans as organisms: "… for the first time, we can pose the fundamental problem of man and begin to understand our evolution, our history, our culture and our biology as a whole."

In the present text, I will deal with the history of those scientific events that led to this situation and briefly speculate about the implications these new discoveries have for future society, and even for our own understanding of human nature.

Turning points in biological knowledge

Over the course of its history, biology has undergone three major revolutions. And here we use the term "revolution" to refer to the emergence of a discovery that is important unto itself but also leads to a radical change in the general approach that characterized this discipline until then.

The first revolution took place in 1860 with the evolutionist theories of Darwin and Wallace, who defended the universality of the origin of all living beings. The second revolution was the discovery of the universality of the biological information mechanism proposed by Watson and Crick in 1953. The third revolution has been the discovery of the universality of animal design and that of the basic processes that regulate biological functions. This last revolution took place in the twentieth century, between 1985 and 2000. Unlike the previous ones, it is the result of contributions by a relatively large number of researchers. These three events have led to a new understanding of evolution and of the biology of human beings themselves.

Evolutionary fact

The idea that species change over time is very old and certainly earlier than Darwin's proposal. In the year 520 B.C., in his treatise, *On Nature,* Anaximander of Miletus introduced the idea of evolution, stating that life began in the oceans. In his book, *Historia Plantarum,* published in 1686, John Ray catalogs 18,600 types of plants and proposes the first definition of species based on common descent. And Darwin's own grandfather, Erasmus Darwin, explicitly proposed that animal species change over time.

What differentiates Darwin and Wallace from their predecessors is that they proposed a *plausible* mechanism of evolution based on the idea of natural selection. Darwin in particular proposed that the strength of natural selection lay in the survival of the fittest, since their greater capacity for survival also insured them a greater capacity to transmit their characteristics to their progeny. Through this process, the characteristics of populations of each particular species were gradually modified over the course of successive generations.

Darwin also had access to information unknown to his predecessors, and that information contributed considerably to his comprehension of the evolutionary phenomenon. It was known, for example, that the Earth was much older than had previously been thought, which allowed much more time for the gradual change prophesized by the theory of natural selection. By Darwin's time, there was also a very well cataloged roster of fossils, which made it possible to verify the existence of gradual change in many lines of animals and plants. This clearly supported Darwin's proposal. It was also known that artificial selection is able to generate very profound morphological changes in a very short period of time. That becomes clear when we consider, for example, the vast variety of breeds of dogs now in existence. They all derive from the wolf, but over the course of five to ten thousand years of artificial—not natural—evolution, man has managed to create a great diversity of canine breeds. This indicates the degree to which biological material is versatile when subjected to selection.

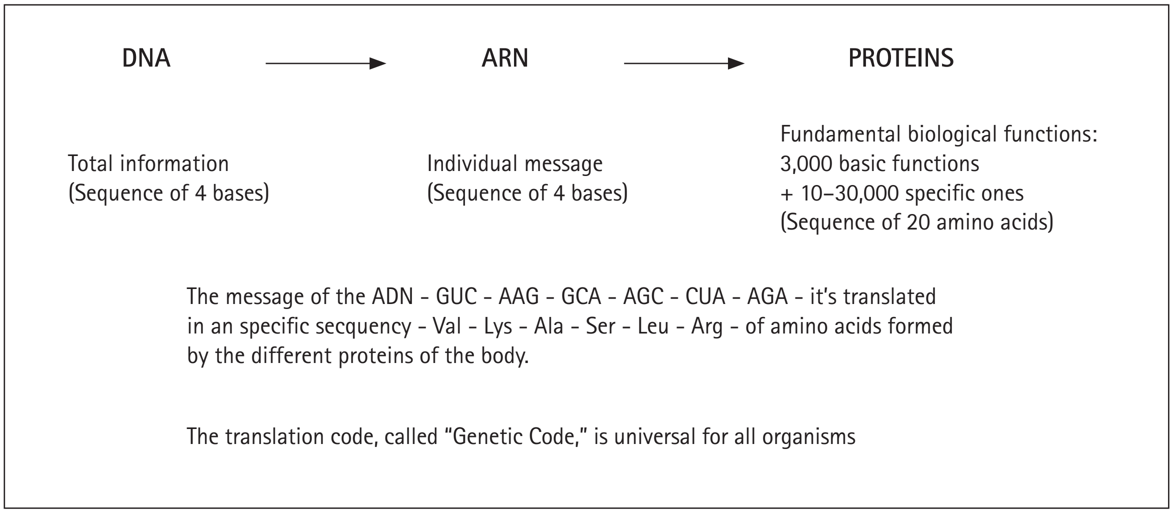
If we were to summarize the implications of evolutionary theory, we would concentrate on three points: 1) all living beings have a shared origin; 2) there has been a process of gradual change over many millions of years that has led to all biological diversity on this planet; and finally, 3) the human species is simply one more of the hundreds of thousands of species that exist or have existed. Darwin's proposal reflects a Copernican change in the approach to the position of the human species as a biological entity. Man is no longer the center of creation. Instead, he is simply one more species among the millions created by evolution. It is no surprise that there was a great social reaction to this in Darwin's time. Even now, evolution is not accepted by all members of society. According to the Gallup institute, in 2004, more than half of the United States believed that man was literally *created,* exactly as the Bible states, some 10,000 years ago.

Genetics and evolution: an operational definition of the gene

Darwin offered a descriptive explanation of biological diversity that was plausible, but not mechanistic. The question is: if all living organisms have a shared origin, what biological function is common to all of them, transmitted from parents to offspring and modifiable in order to generate biological diversity? In his time, Darwin was unable to answer these questions. It was precisely the posing of such questions that led to Genetics, the discipline that studies how biological information is transmitted and modified. We owe the first evidence of the existence of inheritable genetic information to Gregor Mendel, an Augustinian monk who demonstrated that the shape or color of peas is faithfully transmitted from one generation to the next.

But the progress of Genetics in the twentieth century owes much to the fruit fly, *Drosophila melanogaster,* an organism that has become a classic object of study for genetic research because it breeds easily in laboratory settings, has a very short biological cycle (which is very useful when studying the transmission of diverse traits from one generation to the next) and is totally innocuous to human beings. Drosophila studies revealed many concrete inheritable traits (genes), demonstrating that they are located and aligned in cell nucleae—in organules called chromosomes—and that each gene is situated in a specific position in the chromosome. They also showed that inheritable variations (mutations) naturally appear in genes, and that these mutations are the source of the biological variation that is essential to the evolutionary process. These mutations can also be artificially induced using radiation or chemical compounds. In sum, what Drosophila genetics discovered is that the real motivating force for evolution are the genes, which make up the inheritable genetic information, and which can be modified.

After over a century of research on this fly, knowledge of its genetics is the most complete of all the animal kingdom, and a series of concepts and technologies have been developed to carry out experiments that are not possible with any other species.



The message of the ADN - GUC - AAG - GCA - AGC - CUA - AGA - it's translated in an specific secquency - Val - Lys - Ala - Ser - Leu - Arg - of amino acids formed by the different proteins of the body.

The translation code, called "Genetic Code," is universal for all organisms

Figure 1. Translation of the genetic text.

The nature of genetic information

The problem that arose after that, around the nineteen forties, was to discover the gene's physical nature. What was its chemical composition? The solution to this problem led to what I call the second revolution in biology: Watson and Crick's explanation of the nature and structure of genetic information as DNA. The famous article published in *Nature* magazine in 1953 was the beginning of a biological revolution destined to change the very course of humanity. DNA is a molecule with a double-helix structure consisting of two large chains of molecules of a sugar (deoxy-ribose) linked by phosphates. Connecting the two chains, like rungs of a ladder, are other molecules called nitrogenated bases that maintain the structure's stability. Watson and Crick immediately noticed that the structure of the molecule itself explains the mechanism of replication, leading to identical molecules and thus insuring faithful transmission of biological information for generations.

Moreover, the structure of DNA indicates that biological information lies in the sequence of four nitrogenated bases running throughout the molecule. These bases are called thymine (T), guanine (G), adenine (A), and cytosine (C). What an organism inherits from its progenitors, and which will determine its biological characteristics, is simply a sequence written in a language of four letters.

The discovery of the structure and function of DNA modified biology's experimental focus: all organisms are encoded in a language of four letters: A, T, C, and G. From then on, biology concentrated on the study of the properties and structure of DNA. The first complete sequence of DNA obtained for an organism, bacteriophage ØX174, contains 5,000 letters (called bases). By comparison, the DNA sequence of a nematode worm consists of 90 million pairs of bases and the sequence of the Drosophila fruit fly contains 120 million pares of bases, while the human sequence has no less than 3,300 million pairs of bases. Each of these sequences represents a sort of formula for the construction of the species in question.

A universal genetic code

The problem is that life processes are not catalyzed by DNA, but instead by proteins. DNA is simply a recipe that has to be translated into the full variety of proteins—some 3,000 basic ones—that control life processes, including the replication and expression of DNA itself.

Proteins consist of combinations of 20 amino acids, so each protein is different from the others because it is made up of a specific sequence of amino acids. Therefore, the sequence of 4 bases inherited from progenitors has to be translated into sequences of 20 amino acids in order to produce the proteins that support biological functions. Deciphering the translation code, the genetic code, was one of the first great successes of molecular biology. The laboratories of Ochoa, Nuremberg, and Brenner were decisive in deciphering the translation mechanism. Those researchers demonstrated that each amino acid is codified by a specific sequence of three bases (triplet), thus insuring that each gene, which is a particular sequence of the complete DNA, is translated into a specific protein. The AAG triplet codifies for the amino acid, lysine, while GCA codifies alanine, and AGA codifies arginine. Thus, the DNA sequence, AAGGCAAGA would translate into the amino-acid sequence lysine-alanine-arginine (see figure 1).

What is interesting about the genetic code is that it is universal for all organisms. The universality of this code is, itself, proof of evolution. All organisms have the same genetic code simply because we have inherited it from an ancestral forebear. In this context, a gene is simply a concrete sequence of DNA codified for a specific protein that handles a concrete function, for example, the hemoglobin needed for breathing, or myosine for muscles.

The development of molecular biology

The discovery that DNA is an instruction manual for making a living being, and the deciphering of the basic mechanisms of genetic functions—the genetic code and the manufacturing of proteins—mark the beginnings of molecular biology. Beginning in the nineteen seventies, the study of DNA, its structure, and properties, became the main focus of this discipline. That concentration of efforts has led to extraordinarily powerful concepts that make it possible to manipulate DNA with great efficiency. These are the techniques that allow the cloning of genes, the generation of transgenic animals and plants, the possibility of gene therapy, and the Genome Projects. The generation of transgenic organisms—those in which genes from another species have been inserted—springs from the fact that all DNA, no matter what its origin, is chemically identical, and a gene is simply a fragment of DNA. This makes it possible to use chemical methods to mix fragments of DNA (genes) from different origins. Once methods were developed for inserting those fragments into a receiving organism, that organism could have a gene with a different origin. A clear example of this are strains of yeast into which the human gene that codifies for insulin has been inserted. This procedure has created transgenic yeast that manufactures human insulin.

The great development of these procedures in recent years has made it possible to generate transgenic plants (wheat, soy, rice, and others already on the market) and animals of many species, including rats, mice, pigs, flies, and so on. It is important to note that the methods used for the different animal species are very similar and constitute the basis of applications for their therapeutic use in humans. The goal is to use gene therapy to cure genetic diseases. In 2000, *Science* magazine published the first test of gene therapy in which several children were cured of a severe immunodeficiency. Unfortunately, those tests had to be interrupted because of harmful side effects. Three of the cured children later developed cancer. This example simultaneously shows both the potential of such new methods and the fact that they are in a very early stage of development. Given the speed with which they are progressing, it is to be hoped that they will be available in the not-too-distant future.

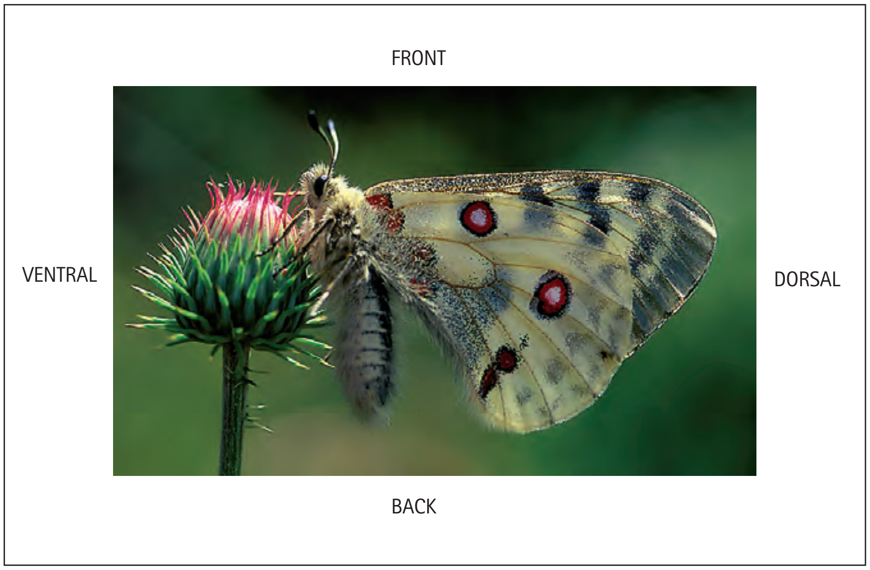


Figure 2. Axes of the animal body.

The genetic design of animal bodies

One of the areas in which molecular biology has progressed significantly, and with considerable applications for human biology, is the field of genetic design of animal bodies. Initially, molecular biology experiments used unicellular organisms, bacteria or viruses, to study the properties and functions of DNA. Those studies produced very important results, as described above, but their very nature made it impossible to draw conclusions about genetic control of the development of complex organisms, such as a fly or a mouse, in which associations of cells have to be grouped in the proper fashion as part of a three-dimensional structure.

Let us consider, for example, a butterfly (figure 2). Each individual cell has to carry out the primary biological functions—protein synthesis, replication of DNA, and so on—but it must also be able to form groups with other cells and differentiate itself in order to make specific organs such as eyes, wings, legs, and so on. Those organs have to be assembled with the other organs in order for each to appear in the right place. An animal design calls for the various parts of the body to be properly situated in space's three dimensions: the anterior-posterior, dorsal-ventral, and proximo-distal axes. This problem of body design has been one of the great challenges to the genetics of superior organisms: how genes specify positional information for different parts of the body so that the cells that are going to make an eye know they have to do so in the upper part of the body, and those that make the legs have to be in the ventral part. In other words, what is the genetic description of a three-dimensional organism? In an insect like a butterfly,

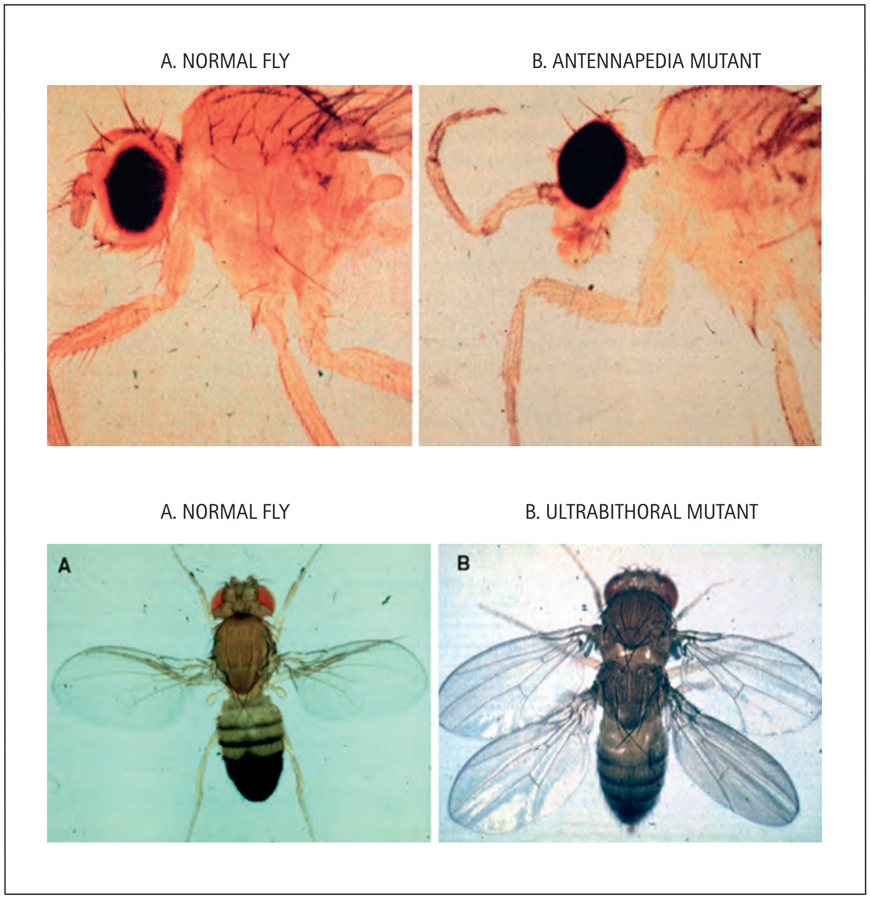


Figure 3. Homeotic mutations of Drosophila.

we can morphologically distinguish a cephalic part, a thoracic part and an abdominal part, but there is no guarantee that this description corresponds to the true genetic description of the organism.

There has been notable progress on the question of the genetic description of animals in the last thirty years. The keys to its genetic design are in what are called homeotic genes, now called Hox. These make up genetic mechanisms that have been studied in great detail in the Drosophila fruit fly. What is characteristic of these genes is that their mutations transform some parts of the body into others (figure 3). A mutation such as Antennapedia (Antp), for example, transforms an antenna into a leg, while a mutation such as Ultrabithorax (Ubx) transforms the halter into a wing, creating a fly with four wings. What is interesting about these transformations is that, despite the fact that the general construction of the body is erroneous, the morphology of the parts is normal: the leg that appears in the antenna of Antp is normal, only its location is anomalous. Likewise, the transformed wings that appear in Ubx flies have the normal size and shape of wings. The only abnormality is where they appear. The implication of those phenotypes is that what Hox genes control is not the morphology of the structures, but rather the general design of the body, the positional information I mentioned above, which insures that each organ appears in its proper place.

Homeotic genes are thus high-level regulator genes that determine the type of development of the different parts of Drosophila's body. A very important question that arose in the nineteen eighties was: how many homeotic genes exist? It was hoped that the identification of all of them would make it possible to clarify the genetic logic underlying the body's design. Studies carried out in the United States and in Spain demonstrated that the number of Hox genes is surprisingly small. There are only nine such genes in Drosophila, where they establish the spatial coordinates of the anterior-posterior axis, recognize the positional value along each axis, and determine the acquisition of the proper development program for generating each part of the body. These results were certainly interesting, but they referred to the fruit fly. At first, we did not suspect that they might have a general value in explaining the body design of other animals, including the human species.

Nevertheless, the progress of molecular biology in the nineteen seventies and eighties permitted the molecular isolation (cloning) and sequencing of Drosophila's Hox genes. By late 1985, all of those genes had already been cloned and sequenced. An extraordinarily important discovery was made when their sequences were compared: they all had one sequence in common, which was named homeobox. The discovery of the homeobox sequence had very important implications: 1) this sequence codifies for a motif of union to DNA, indicating that homeotic proteins function as transcription factors and regulate the activity of other subsidiary genes; 2) the presence of the same sequence in all Hox genes indicates that those genes have a shared origin; and 3) the homeobox sequence is a molecular marker for Hox genes that makes it possible to identify these genes in organisms (the human species, for example) in which it is impossible to detect them with conventional genetic procedures. As we will see below, this last aspect proved very significant.

A universal genetic design

The fact that the homeobox is a molecular marker in Hox genes made it possible to identify the Hox complex in many groups in the animal kingdom, making these genes a fundamental subject of biological research in the nineteen eighties and early nineties. The general result is that the Hox complex has been found in all animal groups in which it was sought. It is thus a universal characteristic of the genome of all animals, including the human species.

Humans have a Hox complex that is very similar to that of Drosophila, except that, instead of having one copy in each genome, we have four.

Drosophila studies had previously established that the function of those genes was to determine the development of the different parts of the body, but there was no evidence as to their function in other organisms. The difficulty of studying this aspect is that the genetic analyses carried out on Drosophila are not possible in many vertebrates and totally impossible in the human species. Therefore, other methods had to be employed.

The molecular technologies developed in the nineteen eighties and nineties made it possible to generate individuals—in this case, Drosophila fruit flies—into which the gene of another species could be inserted in order to study its function in that foreign system. Various experiments of this kind made it possible to conclude that the Hox genes of humans and other vertebrates work in identical or similar ways to those in Drosophila. The conservation of functions is such that human or mouse genes are able to replace their counterparts in Drosophila. That is the case of mouse gene, Hoxd13. When inserted into the fly, it is just as capable of programming the development of the rear part of Drosophila as the fly's own gene. Other very striking examples are, for example, genes of apterous and eyeless Drosophila, which have known counterparts in humans. Apterous is needed to make wings and its mutations produce individuals without wings. Eyeless is needed to program the development of the eye, and individuals in which this gene has mutated do not have eyes.

A. Normal fly. The asterisks indicate where mouse eye genes will be activated.

B. Transgenic fly showing eyes at the base of the wings and halters. Ectopic eyes have been induced by the small-eye mouse gene.

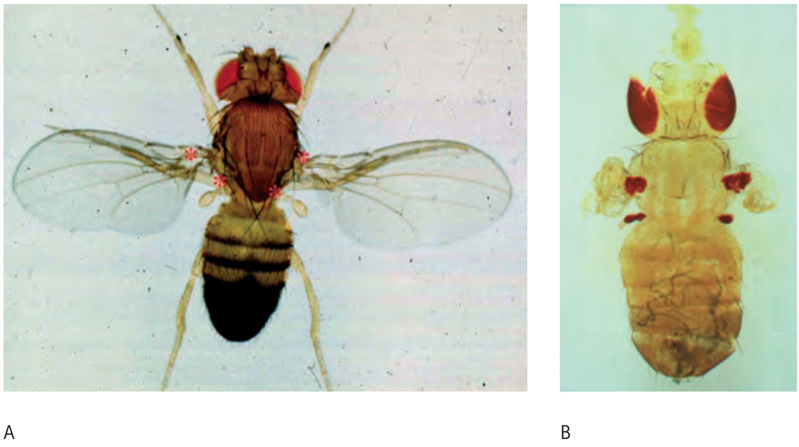


Figure 4. Functional conservation of the genetic programming of the eye.

When a mutant apterous fly receives the human gene, it is capable of forming fly wings. Thus, even though humans do not have fly wings, we have a gene capable of replacing the Drosophila gene that programs the formation of fly wings, thanks to a mechanism of functional conservation. In that same sense, the mouse gene that is homologous to eyeless, called *small eye*, is capable of inducing fly eyes (figure 4). Similar experiments with genes from other organisms have led to the conclusion that the genetic design of eyes is the same in all animals, be they flies, octopi or human beings. The evolutionary invention of a light-receiving organ connected to the brain took place around 540 million years ago and has been inherited by all multi-cellular organisms. These experiments illustrate a general principle of the phenomenon of evolution: when a mechanism that works adequately appears, the genetic programming of that mechanism remains fixed in the genome and stays the same, or with only slight modifications, from then on.

The general conclusion of all the above is that the overall mechanism of genetic design of animals, based on Hox genes and their derivatives, is common throughout the animal kingdom.

The explosion during the Cambrian era, that is, the sudden apparition of bilateralia with organs arranged along all three of the spatial axes, is almost certainly the result of the apparition of the Hox complex and its derivatives during the lower Cambrian. The similarity of sequences among these genes indicates they come from an ancestral gene that underwent various duplications in tandem, thus generating the set of linked genes that make up this complex. We can thus state that all living beings share the same basic biological functions. Together, these studies have given rise to a unifying view of biological processes based, ultimately, on the evolutionary process. As Darwin and Wallace proposed, organisms have a common origin, sharing the mechanism that stores and releases genetic information based on the universality of the function of DNA, RNA, and the mechanism of genetic code. Finally, all members of the animal kingdom share the same genetic process of body design.

An important implication of these observations is that many aspects of the human body's design can be studied in model organisms such as flies, worms, or mice. It is understood that the genetic/molecular base of those processes is common to all species, and therefore, many of the processes involved will be so as well. A typical example of this approach can be found in regeneration studies being carried out on amphibians and chickens. It has long been known that amphibians and reptiles are

Figure 5. The official presentation of the Human Genome by President Clinton. Completion of the Human Genome Project was possible thanks to the collaboration of the public sector—led by Collins—and the private one, led by Venter.

able to regenerate their limbs, while birds and mammals cannot. The studies underway are making it possible to identify genes related to the regenerative process, several of which are also present in species that do not regenerate. It seems that the capacity to regenerate an organ or not depends less on the presence or absence of one or more genes than on the mechanism that regulates common genes. Regenerating species are able to activate these genes following physical trauma while non-regenerators are not. A well-founded speculation is that when the process that regulates those genes is understood, it will be possible to intervene in the control of its functioning in order to artificially induce the regenerative process in species like the human one, which cannot do it naturally.

The genome projects

What has been set out above is, itself, proof of the entire phenomenon of evolution, as it clearly shows the functional universality of biological phenomena. Furthermore, new molecular technology has offered us a more direct demonstration of this universality. In recent years, the complete sequences of DNA (the genome projects) for many animal and plant species have been completed, making it possible to directly compare the degrees of similarity or difference in the biological information of different species.

Particularly relevant in that context are the genomes of the nematode *Caenorabditis elegans,* which contains DNA with 90 million pairs of bases; of the Drosophila fly, with 120 million pairs of bases; and of the human species, with 3,300 million pairs of bases. The Human Genome Project (figure 5) used the DNA of five people (three women and two men) from four different ethnic groups (Hispanic, Asian, Afro-American, and Caucasian). It is interesting to note that no significant differences were detected among them. These projects have managed to identify all the genes in each species, determining their sequence and accumulating that information in databases. Along with the development of very sophisticated software and powerful computers, this has made it possible to compare the significant sequences. That comparison has produced many interesting results, one of the most important of which (figure 6) is the discovery that the human species shares approximately 50% of its genes with those of the nematode *Caenorabditis elegans* and about 60% with the Drosophila fruit fly. This observation is a healthy reminder of our biological origins, which we share with the rest of the animals. Naturally, this is reflected in the DNA that is the common evolutionary record linking us all.

The nematode worm C. Elegans (1998)	90 x 10^6 bp	19,000 genes
The Drosophila fruit fly (2000)	120 x 10^6 bp	14,000 genes
The Human Species (2001)	3,300 x 10^6 bp	40,000 genes

Computer techniques allow us to compare the DNA sequences of different species, that is, their degree of genetic similarity.

Humans share	50% genetic identity with the C. elegans worm. 60% genetic identity with the Drosophila fly.

Figure 6. A comparison of some important genomes.

The study of human illness in model organisms

The high degree of genetic similarity among the species mentioned and, in fact, throughout the animal kingdom, not only validates the phenomenon of evolution; it also has powerful implications for the study of human biology and pathology. Because we share so many genes with organisms such as Drosophila, there are many aspects of biology and human illness that can be studied in flies without the experimental and ethical limitations imposed by human material. The philosophy underlying this is that much of the knowledge obtained by working with Drosophila will also be applicable to us. As we saw above, the study of Hox genes of flies is shedding very important light on the function of those same genes in our own species.

With regard to pathological processes, the latest estimates indicate that 75 percent of genes related with human illness are present in Drosophila. That makes it an enormously important source of information for basic knowledge of human illness. Currently, numerous laboratories around the world are using Drosophila as an organism for studying pathologies such as cancer, Alzheimer's disease, ataxias, and so on. One example of this approach is the experiments that seek to induce the molecular syndrome of Alzheimer in Drosophila. Deposits of the protein, amyloid (Aß), in neurons is a characteristic of that illness. The pathological form contains 42 amino acids rather than 40 and forms aggregates called amyloid plaques. Drosophila technology makes it possible to induce this illness in the eyes and brains of a fly and to study its evolution. It is possible to produce hundreds of individuals and test a large number of possible remedies or compounds that interfere with the development of the illness. Those experiments have made it possible to identify a drug (Congo Red) that considerably mitigates the effect of this illness in flies. Although the drug is toxic for humans and cannot be used to treat the illness, it clearly indicates the potential of this type of technology. Experiments of this kind have already identified various drugs aimed at treating cancer and other degenerative processes.

Can the duration of human life be changed?

The extremely high degree of conservation of fundamental biological phenomena throughout the animal kingdom allows us to speculate on the possibility of manipulating processes only recently considered inaccessible to human intervention. One of the fundamental paradigms of human society and culture is the idea that aging and death are inevitable biological processes. The supposition is that there is internal programming that establishes the maximum lifespan of members of each species within a relatively narrow range.

During the twentieth century, the average human lifespan increased considerably, due mainly to improved living conditions, hygiene, and medical progress. Even so, the estimated maximum lifespan is about 120-125 years. Could this limit be surpassed? That is a subject that has received considerable attention in international science magazines (*Nature* 458, 2008, 1065-1071), fundamentally because of recent discoveries directly related to the genetic programming of lifespan.

The fundamental fact is that, in both the nematode worm, *Caenorhabditis elegans*, and in the Drosophila fly, various genes have been identified whose function is directly related to the aging program of those species. Given the ease with which those organisms can be genetically manipulated, it has been possible to substantially prolong the life of individuals from those species. In the case of nematodes, lifespan has been successfully multiplied by six- or even seven-fold. If this were extrapolated to the human species, it would offer an average human lifespan of some 350 years, and some individuals would live over half a millennium.

What is important about these discoveries is that the aging genes identified in the nematode worm and in Drosophila are also present in the human species. The most studied of those genes, called DAF-16 in worms and FOXO in Drosophila and humans, is related to the insulin path and some of the variant forms of FOXO appear to be particularly frequent in individuals over one hundred years old. Mutations in the human species that affect the activity of the insulin path have also been detected in individuals who are over one hundred. DAF-16/FOXO has been cloned and genetically modified worms have been created in which alterations in the levels of this gene's functions result in alterations that double the lifespan of those worms. The fact that such results can be obtained by altering just one gene illustrates the potential of such techniques. As we mentioned above, this gene is present in our own species, which suggests the possibility that its manipulation could be used to modify the lifespan of human beings.

The future evolution of the human species: technological man

In closing, I would like to briefly reflect upon the evolution of life on our planet, and the life of the human species. Life on our planet began around 2,000 to 3,000 million years ago. Bilateralia animals, the animals that exist today, appeared around 540 million years ago. Around 100,000 to 200,000 years ago, Darwinian selection led to the human species, along with many millions of others, living or extinct. However, the intellectual and technological development of our species has made it especially immune to the process of natural selection. As a result, normal rules of evolution have little or no effect on us nowadays.

Human civilization began some 10,000 years ago and technological development about 200 years ago. DNA technology is about 25 years old. This technology has progressed extremely rapidly, leading to very powerful methods of manipulation. In sum, the vehicle of evolution, DNA, is being modified directly by human intervention. These methods, though still very crude, are being used on experimental animals—flies, mice, worms, and so on—whose great genetic similarity to us indicates that the day is not far off when they can be applied to the human species. These methods have enormous potential, especially when we consider that they only began twenty-five years ago. It is impossible to imagine what they will be able to achieve in another fifty years, not to mention 500 or 5,000. The human species will be able to genetically modify itself in a controlled manner. That perspective offers enormous possibilities for determining our own biological future and evolution. DNA technology offers a new social paradigm, and will be able to completely change the very essence of the human being.

biomedicine at the turn of the century

JESÚS AVILA & JOSÉ M. MATO

In the last century, descriptive biology gave way to quantitative biology, while medicine based on simple visual diagnosis became a medicine of analysis and protocols. In both cases, what was sought was a new, more rigorous methodology offering reproducible data. Quantification and measurement were used in the quest to increase the scientific nature of biology and medicine. These criteria based on correct measurement led to the development of biochemistry, and later, molecular biology. At first, analyses were based on simple models, but these became increasingly more complex and could be extrapolated to help understand how human beings function. Thus, interpretations of growing complexity were carried out by studying simple processes, with few variables, and then studying how those processes interact. At the turn of the twentieth to twenty-first century, this became known as systems biology.

New discoveries in the transition

During the period of transition between the two centuries, scientists managed to decipher the information (or sequence) of the genomes of different organisms, including the human genome (Venter et al. 2001, 1304), carrying out a not—especially—rigorous description of the proteins expressed in those organisms, in terms of both the quantity and nature of those proteins.

Moreover, more was learned, at a fundamental level, about the basic elements of life. The central dogma of molecular biology indicates that the genome (DNA) is transcribed as RNA, one form of which—messenger RNA—is transposed, leading to the synthesis of proteins. Beyond the previously described messenger RNA, transfer RNA and ribosomic RNA, in recent years, a new type of RNA has been described: interference RNA, which acts to regulate genetic expression (Fire et al. 1998, 806).

Nevertheless, the second part of the genetic code, the key to how proteins are folded, remains unknown. This is important because there are various illnesses—called proteinopathies—that are based on the defective, nonfunctional folding of a protein. In many cases, this leads to the aberrant aggregation of those proteins.

Levels of study

Studies have been carried out at both molecular and cellular levels. While basic molecular processes are shared not only by the cells in a single organism, but also between one organism and another, the great variety of cells in superior organisms such as mammals has led to specific studies focused on different cell types. Some types of cells proliferate in a determined cycle whose aberrations can lead to tumors, while others survive for a long time without dividing, in a differentiated state, such as neurons. There are cells that interact intimately among each other, like the ones that form epithelial tissue; while those that make up connective tissue surround themselves with an extra-cellular matrix. But in general, the process of differentiation begins with precursor cells that undergo various modifications, leading to mature cells. Those prolific precursor cells are in turn derived from more primitive cells called stem or mother cells. At the turn of this new century, an important advance has been made in the knowledge of those mother cells. Protocols have been developed to transform mother cells originating in human embryos into different cell types—muscle tissue, nerve tissue, and so on (Thompson et al. 1998, 1145)—suggesting that this process may be used for cellular regeneration therapy. Moreover, there is considerable discussion about the possibility that mother cells from certain adult tissues might be pluripotent, that is, that they might be able to transform into different types of tissue. This discussion continues today. More recently, a discovery of potential interest has been described. This discovery, reprogramming, consists in reversing the differentiation of mature cells from specific tissues, converting them into cells with the characteristics of embryonic mother cells. The expression of transcription factors, Oct4, Sox2, KLF4, and cMyc, converts—reprograms—differentiated fibroblasts into cells with characteristics similar to those of mother cells (Takahashi and Yamanaka 2006, 663). This reprogramming may occur naturally in the transition between epithelial and mesenquimal cells that takes place during development (Mani et al. 2008, 704). Afterwards, these cells with stem-cell characteristics may again differentiate themselves, becoming cell types other than the initial one. In rats, neurons obtained from reprogrammed fibroblasts can be transplanted into a mouse with the symptoms of Parkinson's disease, producing a functional recovery (Wernig et al. 2008, 5856).

New therapies

Beginning with the previously mentioned discovery of interference RNA (RNAi), and the characterization of mother cells, molecular therapy methods have been established that used RNAi to impede the expression of a protein that could be toxic for the organism; as well as cellular therapy methods that use precursor cells in regenerative processes. An alternative source of stem cells is blood from the umbilical cord (Kurtzberg et al. 1996, 157). These cells can be used for cellular therapy in illnesses such as leukemia or hemoglobinopathies, and banks have been set up to store such cells.

The regeneration of organs and tissues varies drastically depending on the nature of the cells that make up those tissues. It is know that, following a hepatomy, the liver regenerates itself, recovering its original size, just as cells change in the blood, skin and bones in a natural way. Artificially, in the case of skin cells, it has been possible to grow skin for regenerative purposes (O'Connor et al. 1981, 75), after cultivating keratinocyte and fibroblast cells. Similarly, the implantation of condriocytes has permitted the regeneration of cartilage (Brittberg et al. 1994, 889). The regeneration of other organs, such as the brain, is more difficult. The development of areas such as molecular biology, cellular biology and biotechnology has permitted better diagnosis of different diseases, and in some cases, the proposal of personalized treatment for patients. Clear differences can be observed among different human beings; different genders, races, hair colors, statures, body sizes, and so on. These differences are due in part to their genomes—DNA—and it has recently been observed that, while what are called polymorphisms of a single nucleotide give rise to small difference in the same gene in different individuals, the largest differences appear to be due to the existence of deletions and/or inversions of genetic material in an individual, as compared to the genetic material of another individual. This observation was considered by *Science* magazine as one of last year's most important scientific achievements (Kennedy 2007, 1833).

Publication of the results

It is important to point out the media's influence on the greater or lesser visibility of discoveries. As we mentioned above, at the end of each year, *Science* magazine announces its favorite discoveries. *Nature* does the same in the magazine, *Nature Methods,* where it chooses the methods of the year. This year, the main method of the year was new technology for faster and cheaper sequencing of genomes, including those of different human beings (Wold and Myers 2008, 19). It has been suggested that this was the technique used to sequence the genome of scientists, Watson and Venter.

Cell also publicizes discoveries published in its magazine. A few years ago, it celebrated its thirtieth anniversary with a special edition (Cell 30th Anniversary Supplement, January 23, 2004) featuring the most relevant works previously published in *Cell.* These included several that earned the Nobel Prize for their authors. One example is the work on the mechanisms of decay of cytoplasm proteins that was carried out with a machine called the proteasoma (Rader and Daugherty 2008, 904). It is possible that some future Nobel Prize Laureates will be authors whose work was featured in that issue of *Cell.*

The most prevalent diseases

Moreover, molecular knowledge of the different types of cells found in different tissues can facilitate a greater knowledge of diseases. From a very general standpoint, we could say that there are four main types of tissues in an organism: muscle, nerve, epithelia, and connective tissue. The first is very related to cardiovascular problems, as the heart is a muscle. The second is associated with neurodegenerative diseases, the third with infectious processes, and both the third and fourth are associated with increased tumor formation. The solution for these four types of problems —cardiovascular, neurodegenerative, oncological, and infectious, are twenty-first century medicine's fundamental challenges. In this chapter, we will deal mainly with problems related to metabolic defects and neurodegenerative processes. For aspects relating to cardiovascular diseases, we would refer readers to a specific issue of *Nature* (volume 451, issue 7181), which covers this subject almost monographically. It includes discussion of possible new therapies for arteriosclerosis (Rader and Daugherty 2008, 904), possible treatment for thrombosis (Mackman 2008, 914) and the use of mother cells for heart disease (Segers and Lee 2008, 937).

For aspects relating to cancer, we would recommend that readers consult *Nature* (volume 441, issue 7092), which includes the supplement, *Insight: signaling in cancer,* with interesting articles including one by Benson JD et al, about the validation of new targets for testing anti-tumor compounds.

Before closing this introduction, we would like to briefly mention the use of animal models for studying illness, and some improvements in medical diagnosis from a general viewpoint.

Animal models

In many cases, efforts have been made to reproduce some pathological aspects of various diseases by using animal models that minify all or some of the disease's characteristics. Those models include worms, mice, and flies. Fundamentally, studies that employ mice as models use them as targets for drug testing, as a first step towards their posterior clinical use. The use of mice models was awarded the Nobel Prize for Medicine in 2007.

Improvements in medical diagnosis

Despite the advantages of models, it is very important to know specifically what happens in a diseased human body. In that sense, knowledge of medical diagnosis has grown considerably. Such knowledge can be obtained by analyzing the components of fluids such as blood, plasma, urine, and cephalorachidian liquid. But the greatest advance may well have been the development of new imaging technique, such as functional-image magnetic resonance, which offers excellent information about parts of the human body that are difficult to analyze, including the brain (Kerr and Denk 2008, 195).

As we mentioned above, in the present text, we will focus mainly on certain aspects of neurodegenerative processes and metabolic problems.

Alzheimer's disease

As an example of a neurodegenerative disease, let us start with Alzheimer's disease, a senile dementia characterized initially by memory loss, followed by great disorientation and dementia by the patient as the disease progresses (Alzheimer 1907, 146).

Aging is the greatest risk factor for suffering Alzheimer's disease. From the age of 65 on, its prevalence doubles every five years. Almost 2.5% of 65-year-olds suffer it, while the incidence is between 40 and 50% among persons over the age of 85.

Different problems arise around the fundamental ones associated with this disease—the patient's loss of memory, understanding, and will. As a consequence of this deterioration, the patient requires care by others, generally the closest relatives, who may well be those most harmed by the disease. Those are the fundamental problems, but they are accompanied by others in the social and economic spheres.

World Heath Organization data from 2001 (Vas 2001) estimated that Alzheimer's disease affected 18 million people on the planet at that time. Approximate data for the year 2006 raise that figure to 26 million people. It has been estimated that the number may triple by the year 2050, given the increased longevity of human beings today, and the fact that the greatest risk factor is aging. This was published on the web page of the Alzheimer's Association in 2007 under the title, *Alzheimer's Disease Facts and Figures* (http://www.alz.org/alzheimers_disease_facts_figures.asp). In that sense, it has also been suggested that, among persons aged one-hundred or more, the probability of suffering senile dementia could be greater than 50%. In other words, neurodegeneration could be considered a normal process occurring at very advanced ages.

From an economic standpoint, in the United States it has been calculated that caring for 5.6 million patients could have a minimum cost of one hundred thousand million dollars (Glenner and Wong 1984, 1131; Masters et al. 1985, 4245). Moreover, data from 2006 indicates expenditure of 4,600 million dollars on palliative drugs, which are the only ones available at this time, as no cure has yet been found.

So this is a chronic illness, and it is devastating, with great human, social, and economic cost. Moreover, it is becoming increasingly prevalent, due to the population's increasing average age.

From 1907—when A. Alzheimer discovered the first case of this illness—to the nineteen eighties, knowledge of Alzheimer's disease was derived fundamentally from the anomalous behavior of its sufferers and from histopathological studies of their autopsies. Autopsies of patient's brains indicate an uncommon abundance of two types of aberrant structures: senile plaques and neurofibrillary tangles. Patient autopsies also reveal considerable neuronal death, especially in the hippocampus, the limbic zone, and the cerebral cortex.

In the nineteen eighties, researchers began studying what happens in Alzheimer patients' brains that leads to the formation of senile plaques and neurofibrillary tangles. They discovered that the main component of senile plaques was a peptide (Glenner and Wong 1984, 1131) and they obtained its sequence of amino acids. Since those plaques took the form of amyloid aggregates, they named its peptide *beta amyloid.* This peptide was later found to be a fragment of a protein (Masters et al. 1985, 4245), which was named Amyloid Precursor Protein (APP). This protein is obliquely present in different cell types, but the amyloid beta-peptide only produces aberrant aggregates in the nervous system. Studying the characteristics of the amyloid peptide, researchers concluded that the cuts occurring in APP to produce it were not the ones that usually occur in this protein in non-pathological circumstances. In normal cases, APP, which is located in the plasmatic membrane of cells, remains united to the membrane, or becomes disconnected when a protease called secretease alpha is cut. After the cut, the larger fragment is secreted into the extra-cellular medium. But in pathological conditions (Price et al. 1998, 461), APP is not cut by secretease alpha, but rather by another protease called secretease, and then by another protease, called secretease. When APP is cut by secretease alpha and secretese, it generates the amyloid peptide that leads to different types of aggregates, the largest of which are senile plaques. While the cut produced by beta secretease is precise, occurring between two specific amino acids, the cut produced by secretease γ is imprecise, falling in a particular region of the APP. This lack of precision generates amyloid beta-peptides of different sizes, the most usual of which contain forty $A\beta_{40}$ and forty-two $A\beta_{42}$ residues. The latter of these has a greater capacity to aggregate than the former, so $A\beta_{42}$ is considered the amyloid peptide with the greatest toxic potential. Once amyloid beta-peptides have formed, they can be broken down by proteases such as neprelysin or the enzyme that breaks down insulin, called IDE. By breaking them down, those proteases prevent the aberrant aggregation of amyloid beta-peptides, which can be toxic. That toxicity has been observed when the peptide is added to a culture of neuronal cells. The toxicity of aggregate amyloid beta-peptides may be due to the fact that they facilitate the entrance of calcium into the cellular cytoplasm and/or act as the antagonist in some neuronal signaling paths. Moreover, it has been suggested that they act on microglial cells, facilitating the secretion by those cells of cytokines and other factors, which would provoke an inflammatory process that could end in neuronal death.

It has also been suggested that, in order for the amyloid peptide's toxic effect to take place in neurons, the latter would require the presence of the tau protein (see below). The main component of neurofibrillary tangles is a protein called *tau* (Grundke-Iqbal et al. 1986, 4913) associated with microtubules. Modified by hyperphosphorylation, this protein is present in what are called paired helicoidal filaments (PHF), whose aggregates are neurofibrillary tangles. Using the isolated tau protein, it was possible to determine that it, alone, was sufficient to produce aggregates similar to PHFs (Montejo De Garcini, Serrano, and Avila 1986, 790). Moreover, it has been reported that different kinase proteins can modify—phosphorilate—the tau protein, and that the one that modifies the most residues of the tau protein is the kinase known as GSK3 (Avila et al. 2004, 361). It has been suggested that both phosphorilated tau protein and aggregates made out of tau protein can be toxic for those cells in which they are present.

More recently, it has been suggested that, following neuronal death, intracellular tau moves to the extra-cellular medium, and that this extra-cellular tau may be toxic for nearby neurons, thus contributing to the propagation of the pathology (Gómez-Ramos et al. 2008, 673).

Alzheimer's disease has been divided into two types—family Alzheimer's disease, of genetic origin; and Alzheimer's disease of sporadic origin. The former is very rare. Possibly, less than one percent of the total of Alzheimer's disease cases are of family origin, so the most common type of this illness is the sporadic one.

Nevertheless, knowledge of the mechanism of family Alzheimer's disease offers important clues about this illness in general. That is significant because, if the process by which plaques and tangles are formed were known, it might be possible to design therapies to combat the disease.

Family cases of Alzheimer's disease are due to mutations in three different genes that codify three proteins: APP, the precursor protein of the amyloid peptide; presenilin 1 (PS-1) and presenilin 2 (PS-2). Mutations in APP, which induce the development of the illness, facilitate its cutting by beta and gamma secreteases and inhibit its cutting by secretease alpha. In all such cases, the formation of the amyloid beta-peptide is facilitated (Price et al. 1998, 461).

Because mutations in APP, PS-1, and PS-2 almost always lead to increased production of the amyloid beta-peptide, it was suggested in the "amyloid cascade hypothesis" (Hardy and Selkoe 2002, 353), that the first step in the development of the apparition of the disease was the presence of a specific amount of the amyloid peptide. One it aggregated, it could trigger posterior pathological processes, including hyperphosphorylation and the aggregation of the tau protein. Nevertheless, anatomical-pathological analyses of the development of this disease did not confirm that hypothesis. Apparently, what most correlates to this disease's pathological process is the pathology related to the tau protein, and not that related to the amyloid protein (Braak and Braak 1991, 239; Arriagada et al. 1992, 631). Therefore, analyses were carried out to see if the result of mutations in APP, PS-1, and PS-2 could converge in the modification of any other protein. One possible protein could be the kinase protein, GSK3, because mutations of APP leading to the apparition of amyloid beta-peptide facilitate the activation of GSK3's kinase activity. This is because the amyloid beta-peptide acts as an antagonist for signal paths, insulin, or WNT, leading to an inactivation of GSK3 (Avila et al. 2004, 361). On the other hand, mutations in PS-1 or PS-2 that lead to an increase in the amount of amyloid peptide can have the same consequences as those indicated for mutated APP, while those mutations in PS-1/PS-2 that do not lead to an increase in amyloid beta-peptide may augment GSK3 activity in other ways (Baki et al. 2004, 2586).

Given the confluence of APP, PS-1, and PS 2 mutations in the effect of activating GSK3, a transgenic mouse model was developed that overexpressed kinase in those parts of the hippocampus and cortex most affected by Alzheimer's disease (Lucas et al. 2001, 27). This mouse reproduced some aspects of the tau pathology and also showed memory deficits (Hernández et al. 2002, 1529). It has therefore been used as a target for testing drugs that could inhibit kinase activity and could therefore repair the cognitive deficit. In those genetically modified mice, the most evident lesion is the degeneration of the dentate gyrus (Engel et al. 2006, 1258), which also occurs with Alzheimer patients and may be responsible for the observed memory loss in both the animal model and Alzheimer patients.

Clinically, patients initially have a growing loss of memory, and a slight cognitive deterioration, which have been related to lesions in the region of the hippocampus where the dentate gyrus is located. Later, the pathology extends to the limbic zone of the temporal lobe, and even later, to the frontal cortex, leading to problems of memory consolidation, behavior, and language. Even later, neuronal death can be observed in the parietal cortex, which can lead to visual-spatial problems or problems of disorientation, for example, in the use of utensils, or incapacity to make decisions, which involves both the parietal and frontal cortexes. All of the problems related to disorientation are clinically called dementia. So this disease can be divided in a very general way into two large stages: an initial one characterized by memory loss, and a posterior one in which dementia appears. The two problems are of the utmost importance, but the second one requires greater attention by those persons caring for the patients. In the transition between the twentieth and twenty-first centuries, there has been a great advance in basic-level knowledge of this disease, but there is still not a good therapeutic application of that knowledge to combat it.

Until now, possible therapies have included palliative drugs, rather than curative or modifying ones. Those drugs rather timidly slow the disease's development, but tragically, it still develops in the end. The ones approved by the United States Food and Drug Administration (FDA) are: Tacrine, Donepezil, Galantamine, Rivastigmine, and Memantine. Annual sales of Donepezil alone are close to one-thousand million dollars, while sales of Memantine, the most recent to enter the market, are close to five-hundred million dollars (Mount and Downton 2006, 780; Stephen Salloway 2008, 65). The first four are inhibitors of the acetilcholinesterase enzyme (Stephen Salloway 2008, 65). This enzyme breaks down the neurotransmitter, acetilcholine, which is needed for perfect neuronal transmission. In Alzheimer's disease, there is preferential damage to cholinergic neurons, which use acetylcholine as a neurotransmitter. Those drugs are therefore used to attempt to maintain high levels of the neurotransmitter, although the first one, Tacrine, was taken off the list because of its toxicity. Memantine is an antagonist to the receptor of another neurotransmittor, glutamate. It has been observed that, among elderly patients, there is an excessive activation of a type of glutamate receptors called NMDA. That activation can be toxic, damaging neurons. In order to protect the neurons of those elderly people, they are given Memantine, which is an antagonist to NMDA receptors (Parsons, Danysz, and Quack 1999, 735; Reisberg et al. 2003, 1333).

Those are the current drugs. Below, we will briefly discuss some diagnostic methods, and some possible future drugs.

As biomarkers for this disease, it is possible to determine levels of the tau protein with different levels of phosphorilylation, and of the amyloid beta-peptide in cephalorachidian fluid. More recently, the levels of as many as 18 components of plasma have been determined as possible indicators of the disease (Ray et al. 2007, 1359), but the diagnostic methods that have probably received the most attention are those that employ imaging techniques such as the PET (Blennow and Zetterberg 2006, 753) and functional magnetic resonance (Logothetis 2008, 869). With these techniques, it is possible to follow the expansion of the ventricles following the neuronal death that affects Alzheimer patients. Of these two techniques, functional magnetic resonance seems to be the most advantageous. It measures hemodynamic changes in different parts of the brain, is non-invasive and has good time-space resolution that can show results correlated with a specific activity being carried out by an individual (Logothetis 2008, 869).

The new drugs are already, or very close to being clinically tested and may be modifiers (Stephen Salloway 2008, 65) rather than palliatives for this disease. In other words, these possible future drugs have a mechanism based on the different observations of this disease's pathology that have been carried out at a basic level. Some are being developed to reduce levels of amyloid beta-peptide, including Bapineuzumab, which involves the development of specific antibodies—vaccines—against the amyloid peptide. Inhibitors of beta and gamma secretease are also being developed, some of which modulate secretease in an effort to reduce the A $_{1\text{-}42}$/ A $_{1\text{-}40}$ relation. Some, such as Flurizan and Tarenflumbol, also show a possible anti-inflammatory effect. Still, recent news offers negative data for Flurizan. There are other compounds, such as Clioquinol, that prevent beta amyloid aggregation, or polyphenol extracts that may also prevent the oligomerization of the amyloid beta-peptide (Wang et al. 2008, 6388; Stephen Salloway 2008, 65). Others may maintain high levels of those enzymes, such as IDE, the enzyme that breaks down insulin, which can break down the amyloid peptide. One of these compounds is Rosiglitazone, an agonist of PPAR (Pedersen et al. 2006, 265). There has also been a search for inhibitors that do not link directly to secretease but rather to its substrate, impeding the proteolithic cut of the enzyme in that substrate, but not in others (Kukar et al. 2008, 925).

Moreover, it has been shown that the toxic effect of the excessive activity of NMDA receptors could increase levels of amyloid beta-peptide (Harkany et al. 2000, 2735) and of the tau protein (Amadoro et al. 2006, 2892). Therefore, there is a search for antagonists to NMDA receptors, other than Memantine. One of these is Dimebon. Another type of study has been the search for antioxidants that could act as neuroprotectors, but it has not shown significant results.

Finally, regarding pathology related to the tau protein, there is a search for specific inhibitors of kinases like GSK3, that primarily modify that tau protein. It has recently been observed that methyl blue, an antiaggregant of the tau protein, might have a therapeutic effect. Those are just a few examples of the tremendous effort being made to prevent this terrible disease.

Non-alcoholic fatty liver disease

Non-alcoholic fatty liver disease (NAFLD) is the most frequent cause of liver disease in the Western world and is thus a world health problem. It affects around 20 million people in the United States and a similar number in the European Union (Adams and Lindor 2007; Ahmed and Byrne 2007). NAFLD is frequent in patients with obesity, diabetes, hyperlipidemia, and hypertension (Abdelmalek and Diehl 2007). Increasing obesity in Westernized countries justifies growing interest in the study of NAFLD. Approximately 50% of obese individuals have NAFLD (Angulo 2007). NAFLD is a clinical-pathological term used to describe a broad range of situations running from a simple accumulation of fat on the liver (non-alcoholic steatosis) to non-alcoholic steatohepatitis (NASH, an accumulation of fat with inflammation, necrosis, and fibrosis). NAFLD is generally an asymptomatic disease, although in a minority of NAFLD patients, it leads to the development of cirrhosis, liver failure, and hepatocellular carcinoma (HCC). Approximately 10% of patients with NAFLD develop NASH, and of these, between 10–20% develop cirrhosis and HCC. Excessive alcohol consumption also produces fatty liver and, as happens with NAFLD, the liver disease can also lead to steatohepatitis, cirrhosis, and HCC. Nevertheless, it is important to emphasize that NAFLD and alcohol-induced liver disease are two different diseases. NASH is also different than other forms of hepatitis caused by various viral infections, such as hepatitis B and C. While clinical diagnosis for NAFLD is based on increases of transaminases in the blood, on body mass index (BMI is calculated by dividing a person's weight in kilos by the square of their height in meters. Values of between 18.5 and 25 are considered normal), on accumulations of fat on the liver visible through sonograms or magnetic resonance, and on the presence of other factors, such as obesity, diabetes, hypertension, and hyperlipidemia. Confirmation of the presence of NASH and the degree of fibrosis and necrosis requires a liver biopsy.

Despite the fact that NAFLD is a world health problem, it is not known why it leads to NASH in some individuals and not in others. There are various hypotheses. The leading one states that two *hits* are necessary. The first hit is the accumulation of fat on the liver, but the nature of the second hit is unknown, though studies carried out in the last few years on genetically modified animals have offered new clues as to the second hit. Nor do we know why some NASH patients progress to cirrhosis and HCC and others don't.

About 25% of the adult population of the United States is obese and another 55% is overweight (Sturm 2002). While those numbers are somewhat better in Europe's adult population, obesity in the European Union is also a public health problem. Obesity is caused fundamentally by an excessive ingestion of calories relative to energy consumption. On an individual level, obesity can be more complex, involving genetic factors that regulate metabolism and lipid storage, the brain's control of eating and exercise habits, and other unknown factors. It is certain that some individuals have genetic factors that have a decisive influence on their tendency to accumulate fat, or that favor the sensation of hunger rather than satiation. For such individuals, weight control is very difficult, but for the majority of the population, genetic factors are less decisive and can be more easily compensated by changes in eating habits. Any obese individual will lose weight if he or she is obliged to maintain a low-calorie diet in combination with exercise. We know this to be true because patients subjected to a gastric by-pass (a surgical intervention to drastically reduce the size of the stomach from around one liter to only 30-60 milliliters) markedly reduce the amount of fat in their adipose tissue.

For most individuals, adipose tissue is nothing but an inert mass of fat, but since the mid nineteen nineties, we know that, biologically, it is very active tissue. In 1994, Friedman and his collaborators identified the leptin hormone, discovering that its lack was the cause of the extreme obesity in a mutant mouse called *obese* (Zhang et al. 1994). Those mice are enormous. While a normal mouse weights around 30 grams, *obese* mice can weigh as much as 90 grams. They have high levels of lipids in their blood and develop fatty livers. Leptin is synthesized in adipose tissue. In normal animals, the amount of leptin in the blood is proportional to the amount of adipose tissue. That is the mechanism used by adipose tissue to inform the brain that enough food has been ingested. *Obese* animals have a mutation in the leptin gene that leads them to synthesize a hormone that is not biologically active. Thus, their brains do not receive an adequate signal to stop eating. Unfortunately, in most obese individuals, the concentration of leptin is abnormally high, rather than low. But the few obese patients that have a genetic leptin deficiency respond well to treatment with this hormone, reducing the accumulation of body fat.

While the discovery of leptin did not lead to the curing of obesity, it did mark a permanent change in how we think about the physiopathology of obesity. Since the discovery of leptin, researchers have discovered other hormones and cytokines (proteins whose main activity is the control of inflammation) originating in adipose tissue that regulates appetite and/or lipid metabolism. These include adiponectine (a hormone synthesized by adipose tissue that favors the oxidation of fatty acids and the metabolism of glucose and whose levels in the blood are inversely proportional to BMI); resistine (a hormone synthesized by adipose tissue and related to inflammation and diabetes), and tumor necrosis factor alpha (TNF-alpha, a cytokine involved in the regulation of cell death, differentiation and proliferation that plays an important role in the etiology of diverse diseases, including diabetes). In other words, adipose tissue is much more than a mere fat deposit. It also plays a fundamental role in appetite control and the metabolism of lipids and carbohydrates.

Signals that encourage us to eat must be balanced with the brain's appetite control center so that the latter initiates the desire to eat when a negative energy balance occurs. Studies carried out over the last fifteen years with genetically modified mice have made it possible to identify obesity genes. Most genetically modified animals that develop obesity do so because they eat more. Their extreme obesity is caused by mutations that affect their eating habits. It is the increased ingestion of food, and not the capacity to metabolize fats, that causes obesity. In other words, most obesity genes regulate appetite, not lipid metabolism. For example, an obese mutant mouse called *diabetic* has a leptin receptor deficiency. This mouse's problem is that the lack of a biologically active leptin receptor impedes leptin from informing the brain cells that it is time to stop eating. Another obese mouse, known as *yellow mouse,* has a mutation that affects the route of pro-opiomelanocortine (POMC), which is important for appetite reduction. The mutation of those obesity genes produces an accumulation of fat in those mice that is independent of the existence of other genetic or environmental factors. Generally speaking, the situation in humans is more complex, and obesity rarely occurs as a consequence of the mutation of a single gene. In humans, there is usually more than one gene involved in the development of obesity, and the environment generally has an important role in the body's fat accumulation. In other words, on an individual level, eating habits are the main cause of obesity, although an individual's genetic characteristics can also play an important role in that behavior.

Fatty acids are the fat components of triglyceride molecules. Triglycerides are the main component of the fat in adipose tissue and in food. Fatty acids thus come from food, but can also be synthesized from carbohydrates, mostly in the liver. Fatty acids are an important energy reserve, not only because they have a greater caloric density per gram than sugars or proteins, but also because they are hydrophobic. Rather than attracting water, they repel it, and can thus be stored in the body in a more compact fashion than carbohydrates or proteins, which do attract water. Thus, while the caloric density of fat is 9 kcal per gram, that of sugars and proteins is around 4 kcal per gram. Moreover, while fats do not accumulate water, carbohydrates accumulate 2 grams of water per gram of sugar. In other words, there is approximately six times more energy stored up in a gram of fat than in a gram of sugar. This means that if the human body were to store energy in the form of carbohydrates, rather than as fat, an individual would need to store around 100 kilos of glucogen in order to have the energy equivalent of 15 kilos of fat, which is the approximate amount of fat on a non-obese adult human.

When triglycerides from food enter the stomach, the stomach acids, bile salts, and digestive enzymes known as lipases break down those triglycerides into their two components: fatty acids and glycerol. Lipases are synthesized by the pancreas and bile salts come from the liver through the gall bladder. Once freed from the triglycerides, fatty acids enter the cells that make up the intestinal walls and are again converted into triglycerides. There, along with cholesterol esters, phospholipids, and proteins, they create nanoparticles called kilomicrons. The latter are carried to the blood by the lymph system. In the blood, they come into contact with high-density lipoproteins, with which they exchange triglycerides for cholesterol esters. As kilomicrons pass through capillaries, adipose tissue, muscles, the heart, and other non-hepatic tissues, they lose their load of fatty acids as a result of the activity of the lipase lipoprotein enzyme. The fatty acids generated in that way are oxidized to produce energy needed by each of those tissues to fulfill its biological function, or they accumulate in the form of triglycerides. Finally, the kilomicrons remaining in the blood, almost totally freed of their load of triglycerides, are transported into liver cells to be metabolized.

Food intake also leads to the secretion of insulin. This secretion by the pancreas's beta cells stimulates the synthesis of glucogen in muscles and in the liver. In adipose tissue, insulin also stimulates the metabolism of glucose and the synthesis of glycerol, the molecule to which fatty acids link to form triglycerides. Also, in the liver, insulin suppresses gluconeogenesis (the synthesis of glucose and glucogen) and accelerates glucolysis (the metabolism of glucose), which increases the synthesis of fatty acids that accumulate in the form of triglycerides. If intake of fat and carbohydrates surpasses their consumption in a chronic way, the excess energy accumulates in adipose tissue and the blood carries it to the liver in the form of free fatty acids linked to albumin. Finally, those fatty acids accumulate in the liver in the form of triglycerides, producing NAFLD.

We have known for at least the last 500 years that when ducks and geese are overfed, they develop fatty livers. In 1570, Bartolomé Scappi, Pope Pius V's chef, published a cookbook titled *Opera,* in which he wrote that, "…the livers of domestic geese raised by the Jews reach the extreme size of 3 pounds." Overfeeding not only produces NAFLD in birds. In the laboratory, overfeeding rats and mice with a diet rich in fatty acids and carbohydrates to induce the generation of fatty livers continues to be a very widespread experimental method.

Research carried out over the last ten years with genetically modified mice has been fundamental in showing that the deactivation of certain enzymes needed for hepatic synthesis of fatty acids and triglycerides—acetil coenzyme A carboxilase, diacilglycerol aciltranferase, elongase of long-chain fatty acids, glycerol 3-phosphate mitochondrial aciltransferase, and stearoil-coenzyme A desaturase—prevents the formation of fatty acids induced by a diet rich in fat and carbohydrates (Postic and Girard 2008). These data suggest that the decrease in hepatic synthesis of triglycerides may possibly be an important therapeutic target for the treatment of NAFLD. Nevertheless, it is important to emphasize that the accumulation of triglycerides in the liver is not necessarily toxic. Instead, it may be a way of protecting the liver from toxicity caused by free fatty acids—fatty acids not linked to glycerol molecules to form triglycerides. For example, in *obese* mice, the inhibition of triglyceride synthesis improves steatosis but worsens liver damage (necrosis, inflammation, and fibrosis) (Yamaguchi et al. 2007). If free fatty acids are not oxidized to produce energy, they are metabolized by the microsomal system called cytochrome P450 2E1 (CYP2E1 is particularly active in the liver. Not only does it metabolize exogenous substances such as alcohol, drugs, and pro-carcinogens, it also participates in the metabolism of cholesterol, bile acids, and fatty acids). The metabolism of fatty acids by CYP2E1 generates cytotoxic substances, such as those that react to oxygen (ROS) and peroxidized lipids, that produce hepatic inflammation and necrosis.

In the liver, triglyceride molecules accumulate in the cytoplasm of hepatocytes, forming small drops of lipids. Those drops of lipids are not a simple accumulation of triglycerides—like the drops that form when oil is mixed with water—they are organules whose creation requires the presence of certain specific proteins. One of these proteins is called ADFP. Mice lacking ADFP do not develop NAFLD when overfed with a fat-rich diet (Chang et al. 2006). While ADFP may be a therapeutic target for treating NAFLD, it is not yet known whether the inhibition of the accumulation of triglycerides in obese animals through the inhibition of ADFP may increase liver damage. Another experimental approach, which has been used to prevent NAFLD, is to block the activity of certain transcription factors (proteins that link DNA and regulate the expression of specific genes) that control the synthesis of lipids. One of those transcription factors, called SREBP-1c, mediates the effect of insulin on the expression of those enzymes that regulate the synthesis of fatty acids. Steatosis improves in *obese* mice that are deficient in SREBP-1c (Yahagi et al. 2002), but it is not yet know whether the inhibition of SREBP-1c can increase liver damage in the long run. In sum, even though the inhibition of the synthesis of triglycerides, or of their accumulation in the form of vesicles in the liver, are theoretically good therapeutic approaches to the prevention of NAFLD, it is important to recall that those procedures are not free of possible side effects that might produce liver damage. Therefore, it is not clear that they can have any clinical application.

Surprisingly, malnutrition can also provoke fatty liver. It is not entirely known how this happens, although studies carried out in recent years with genetically modified animals offer new data about the importance of certain nutrients in the development of NAFLD.

In 1930, Banting and Best, the discoverers of insulin, observed that diabetic dogs treated with insulin developed fatty livers, and that this situation could be corrected by administering choline—a micro-nutrient that is a precursor to the synthesis of methionine. Some years later, Best, du Vigneaud, and other research groups observed that when mice or rats are fed a diet lacking methionine and choline, in just a few weeks they also develop steatosis that leads to NASH and, in some animals, even HCC, if the diet is maintained. Those animals fed with a diet lacking methionine and choline not only are not obese; in general, they weigh less than mice fed with a normal diet. Those experiments not only related steatosis to diabetes, they also provided the very first evidence of the importance of a group of micro-nutrients known as methyl-group donors (choline, methionine, betaine, and folic acid) in the prevention of steatosis (Mato et al. 2008).

In mammals, including humans, methionine is an essential amino acid, that is, it cannot be synthesized by the body and must be taken in through food. When methionine is administered to a person orally, the blood levels of this amino acid increase transitorily, returning to their basal levels in two or three hours. The speed with which a person returns to basal levels of methionine after ingesting it is an indicator of the metabolism of this amino acid in the body. In cirrhotic patients, the metabolism of methionine is markedly slower than in individuals with normal hepatic functions. The first step in the metabolism of methionine is its conversion into S-adenosilmethionine (SAMe), a molecule discovered by Giulio Cantoni in 1953 (Cantoni 1975). SAMe has a special place in biology due to its capacity to modify other molecules and their biological activity by adding a methyl group (a methyl group is a carbon atom linked to three hydrogen atoms). Those molecules include DNA, proteins, and phospholipids. This reaction, known by the general name of methylatin, can prevent the expression of certain genes. In other words, it can cause the same result as a genetic mutation, even though its mechanism is not genetic, but instead, epigenetic.

SAMe synthesis is markedly reduced in the livers of cirrhotic patients (Duce et al. 1988) and treatment with SAMe increases the survival of patients with alcoholic cirrhosis (Mato et al. 1999), which confirms the important role that an alteration of the methionine metabolism has in the progression of liver disease. Consequently, mice with deficient hepatic synthesis of SAMe, though normal sized and not obese, develop steatosis, NASH, and HCC (Lu et al. 2001). In mice with deficiencies of the enzyme glycine N-methyltransferase—the main enzyme that metabolizes SAMe in the liver—the hepatic concentration of SAMe is around 40 times higher than in normal mice (Martínez Chantar et al. 2008). Surprisingly, even though those "super-SAMe" mice are normal-sized and not obese, they also develop steatosis, NASH, and HCC. Such results indicate that both a deficiency and an excess of SAMe in the liver induce NAFLD, and even the apparition of HCC, in the absence of obesity. This brings out the importance of the metabolism of methyl groups in the regulation of the hepatic function and complicates the therapeutic use of this molecule. CUP2E1 liver activity is increased in patients with NASH, diabetics, and individuals who have fasted for long periods of time. It is also increased in patients with alcoholic steatohepatitis, a disease very similar to NASH. Moreover, hepatic CYP2E1 activity is increased in animals that have been fed a diet deficient in methionine and choline, and in mice with deficient hepatic synthesis of SAMe. These and other results have shown the importance of oxidative stress generated by the peroxidation of lipids via CYP2E1 in the pathogenesis of NASH, that is, in the progression from steatosis to NASH. Surprisingly, a CYP2E1 deficiency in mice did not prevent the development of NASH induced by a diet lacking in methionine and choline, nor did it prevent the peroxidation of lipids, which indicates the existence of an alternative system of lipid peroxidation that acts in the absence of CYP2E1 (Leclercq et al. 2000). Those authors also observed that in CYP2E1-deficient mice treated with a diet lacking in methionine and choline, the hepatic expression of CYP4A10 and CYP4A14 is induced, and that these two enzymes are responsible for the lipid peroxidation and generation of ROS in those animals. CYP4A10 and CYP4A14 belong to the family of microsomal enzymes known by the generic name of CYP 450, of which CYP2E1 is also a member. This means that other members of the CYP 450 family that are not very active in normal conditions can substitute for CYP2E1 in the peroxidation of lipids when the activity of that enzyme is inhibited or mutated. That is what happens with "super-SAMe" mice. SAMe is an inhibitor of the hepatic expression of CYP2E1 and, as a result, its expression is inhibited in "super-SAMe" mice even when they have developed NAFLD. In those mice, as in CYP2E1-deficient animals fed with a diet lacking methionine and choline, the expression of CYP4A10 and CYP4A14 is stimulated and catalyzes the peroxidation of lipids and the formation of ROS.

An important conclusion of these studies is that therapeutic approaches targeting a single enzyme from the CYP 450 microsomal system are not efficient in preventing the generation of ROS and the peroxidation of lipids, and thus fail to block the initiation and progression of NASH. One of the main characteristics of biology is the redundancy of biochemical routes that control essential biological functions, such as cellular proliferation or defense against external cytotoxic agents. The evolutionary advantages to having developed a complex system such as CYP 450, which contains tens of enzymes whose mission is to protect the liver from the cytotoxic action of innumerable xenobiotics, is obvious. On the other hand, the redundancy of enzymes from the CYP 450 complex is a disadvantage when seeking to neutralize that system in order to avoid its side effects, such as the progression of NASH in individuals with steatosis.

Bibliography

Abdelmalek, M. F. and A. M. Diehl. "Nonalcoholic fatty liver disease as a complication of insulin resistance." *Journal Med. Clin. North Am* 91, November 2007, 1125.

Adams, L. A. and K. D. Lindor. "Nonalcoholic fatty liver disease." *Journal Ann. Epidemiol* 17, November 2007, 863.

Ahmed, M. H. and C. D. Byrne. "Modulation of sterol regulatory element binding proteins (SREBPs) as potencial treatments for non-alcoholic fatty liver disease (NAFLD)." *Journal Drug Discovery Today* 12, September 2007, 740.

Alzheimer. A. *Psych. genchtl Med* 64, 1907, 146.

Amadoro, G., M. T. Ciotti, M. Costanzi, V. Cestari, P. Calissano, and N. Canu. "NMDA receptor mediates tau-induced neurotoxicity by calpain and ERK/MAPK activation." *JournalProc Natl Acad Sci USA* 103, February 2006, 2892.

Angulo P. "Obesity and nonalcoholic fatty liver disease." *Journal Nutr. Rev.* 65, June 2007: S57.

Arriagada, P. V., J. H. Growdon, E. T. Hedley-Whyte, and B. T. Hyman. "Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease." *Journal Neurology* 42, March 1992, 631.

Avila, J., J. J. Lucas, M. Pérez, and F. Hernández. "Role of tau protein in both physiological and pathological conditions." *Journal Physiol Rev* 84, April 2004, 361.

Baki, L., J. Shioi, P. Wen, Z. Shao, A. Schwarzman, M. Gama-Sosa, R. Neve, and N. K. Robakis. "PS1 activates PI3K thus inhibiting GSK-3 activity and tau overphosphorylation: effects of FAD mutations." *Journal Embo* J 23, July 2004, 2586.

Blennow, K. and H. Zetterberg. "Pinpointing plaques with PIB." *Journal Nat Med* 12, July 2006, 753.

Braak, H. and E. Braak. "Neuropathological stageing of Alzheimer-related changes." *Journal Acta Neuropathol* 82, 1991, 239.

Brittberg, M., A. Lindahl, A. Nilsson, C. Ohlsson, O. Isaksson, and L. Peterson. "Treatment of deep cartilage defects in the knee with autologous chondrocyte transplantation." *Journal N Engl J Med* 331, October 1994, 889.

Cantoni GL "Biochemical methylations: selected aspects." *Journal Ann. Rev. Biochem* 1975. 45: 285-306.

Chang, BH-J., L. Li, A. Paul, S. Taniguchi, V. Nannegari, W. C. Herid, and L. Chan. "Protection against fatty liver but normal adipogenesis in mice lacking adipose differentiation-related protein." *Journal Mol. Cell. Biol.* 26, February 2006, 1063.

Duce, A. M., P. Ortiz , C. Cabrero , and J. M. Mato . "S-Adenosy-L-methionine synthetase and phospholipd methytransferase are inhibited in human cirrhosis." *Journal Hepatology* 8, January-February 1988, 1530.

Engel, T., J. J. Lucas, P. Gómez-Ramos, M. A. Morán, J. Avila, and F. Hernández. "Cooexpression of FTDP-17 tau and GSK-3beta in transgenic mice induce tau polymerization and neurodegeneration." *Journal Neurobiol Aging* 27, September 2006, 1258.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." *Journal Nature* 391, February 1998, 806.

Glenner, G. G. and C. W. Wong. "Alzheimer's disease and Down's syndrome: sharing of a unique cerebrovascular amyloid fibril protein." *Journal Biochem Biophys Res Commun* 122, August 1984, 1131.

Gómez-Ramos, A., M. Díaz-Hernández, A. Rubio, M. T. Miras-Portugal, and J. Avila. "Extracellular tau promotes intracellular calcium increase through M1 and M3 muscarinic receptors in neuronal cells." *Journal Mol Cell Neurosci* 37, April 2008, 673.

Grundke-Iqbal, I., K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. "Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology." *Journal Proc Natl Acad Sci USA* 83, July 1986, 4913.

Hardy, J. and D. J. Selkoe. "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics." *Journal Science 297*, July 2002, 353.

Harkany, T., I. Abraham, W. Timmerman, G. Laskay, B. Toth, M. Sasvari, C. Konya, et al. "Beta-amyloid neurotoxicity is mediated by a glutamate-triggered excitotoxic cascade in rat nucleus basalis." *Journal Eur J Neurosci* 12, August 2000, 2735.

Hernández, F., J. Borrell, C. Guaza, J. Avila, and J. J. Lucas. "Spatial learning deficit in transgenic mice that conditionally over-express GSK-3beta in the brain but do not form tau filaments." *Journal J Neurochem* 83, December 2002, 1529.

Kennedy, D. "Breakthrough of the year." *Journal Science* 318, December 2007, 1833.

Kerr, J. N. and W. Denk. "Imaging in vivo: watching the brain in action." *Journal Nat Rev Neurosci* 9, March 2008, 195.

Kukar, T. L., T. B. Ladd, M. A. Bann, P. C. Fraering, R. Narlawar, G. M. Maharvi, B. Healy, et al. "Substrate-targeting gamma-secretase modulators." *Journal Nature* 453, June 2008, 925.

Kurtzberg, J., M. Laughlin, M. L. Graham, C. Smith, J. F. Olson, E. C. Halperin, G. Ciocci, et al. "Placental blood as a source of hematopoietic stem cells for transplantation into unrelated recipients." *Journal N Engl J Med* 335, July 1996, 157.

Leclercq, I. A., G. C. Farell, J. Field , D.R. Bell, F. J. González , and G. H. Robertson. "CYP2E1 and CYP4A as microsomal catalysts of lipid peroxides in murine nonalcoholic steatohepatitis." Journal J. Clin. Invest. 105, April 2000, 1067.

Logothetis, N. K. "What we can do and what we cannot do with fMRI." Journal Nature 453, June 2008, 869.

Lu S. C., L. Álvarez, Z. Z. Huang, L. Chen, W. An, F. J. Corrales , M. A. Avila , et al. "Methionine adenosyltransferase 1A knockout mice are predisposed to liver injury and exhibit increased expression of genes involved in proliferation." *Journal Proc. Natl. Acad. Sci. USA* 98, May 2001, 5560.

Lucas, J. J., F. Hernández, P. Gómez-Ramos, M. A. Moran, R. Hen, and J. Avila. "Decreased nuclear beta-catenin, tau hyperphosphorylation and neurodegeneration in GSK-3beta conditional transgenic mice." *Journal Embo* J 20, January 2001, 27.

Mackman, N. "Triggers, targets and treatments for thrombosis." *Journal Nature* 451, February 2008, 914.

Mani, S. A., W. Guo, M. J. Liao, E. N. Eaton, A. Ayyanan, A. Y. Zhou, M. Brooks, et al. "The epithelial-mesenchymal transition generates cells with properties of stem cells." *Journal Cell* 133, May 2008, 704.

Martínez-Chantar M. L., M. Vázquez-Chantada, U. Ariz, N. Martínez, M. Varela, Z. Luka, A. Capdevila, et al. "Loss of glycine N-methyltransferase gene leads to steatosis and hepatocellular carcinoma in mice." *Journal Hepatology* 47, April 2008, 1191.

Masters, C. L., G. Simms, N. A. Weinman, G. Multhaup, B. L. McDonald and K. Beyreuther. "Amyloid plaque core protein in Alzheimer disease and Down syndrome." *Journal Proc Natl Acad Sci USA* 82, June 1985, 4245.

Mato J. M., J. Camara , J. Fernández de Paz, L. Caballería , S. Coll , A. Caballero , L. García-Buey, et al. "S-adenosylmethionine in alcoholic liver cirrosis: a randomized placebo-controlled, double-blind, multicenter clinical trial." *Journal J. Hepatol.* 30, June 1999,1081.

Mato, J. M., M.L. Martínez-Chantar, and S. C. Lu. "Methionine metabolism and liver disease." *Journal Annu. Rev. Nutr.* 28, August 2008, 273.

Montejo de Garcini, E., L. Serrano, and J. Avila. "Self assembly of microtubule associated protein tau into filaments resembling those found in Alzheimer disease." *Journal Biochem Biophys Res Commun* 141, December 1986, 790.

Mount, C. and C. Downton. "Alzheimer disease: progress or profit?" *Journal Nat Med* 12, July 2006, 780.

O´Connor, N. E., J. B. Mulliken, S. Banks-Schlegel, O. Kehinde, and H. Green. "Grafting of burns with cultured epithelium prepared from autologus epidermal cells." *Journal Lancet* 8211, January 1981, 75.

Parsons, C. G., W. Danysz, and G. Quack. "Memantine is a clinically well tolerated N-methyl-D-aspartate (NMDA) receptor antagonist--a review of preclinical data." *Journal Neuropharmacology* 38, June 1999, 735.

Pedersen, W. A., P. J. McMillan, J. J. Kulstad, J. B. Leverenz, S. Craft, and G. R. Haynatzki. "Rosiglitazone attenuates learning and memory deficits in Tg2576 Alzheimer mice." *Journal Exp Neurol* 199, June 2006, 265.

Postic, C. and J. Girard. "Contribution of de novo fatty acid síntesis to hepatic steatosis and insulin resistance: lessons from genetically engineered mice." *Journal J. Clin. Invest.* 118, March 2008, 829.

Price, D. L., R. E. Tanzi, D. R. Borchelt, and S. S. Sisodia. "Alzheimer's disease: genetic studies and transgenic models." *Journal Annu Rev Genet* 32, 1998, 461.

Rader, D. J. and A. Daugherty. "Translating molecular discoveries into new therapies for atherosclerosis." *Journal Nature* 451, February 2008, 904.

Ray, S., M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, et al. "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins." *Journal Nat Med* 13, November 2007, 1359.

Reisberg, B., R. Doody, A. Stoffler, F. Schmitt, S. Ferris, and H. J. Mobius. "Memantine in moderate-to-severe Alzheimer's disease." *Journal N Engl J Med* 348, April 2003, 1333.

Segers, V. F. and R. T. Lee. "Stem-cell therapy for cardiac disease." *Journal Nature* 451, February 2008, 937.

Stephen Salloway, J. M., F. W. Myron, and J. L. Cummings. "Disease-modifying therapies in Alzheimer´s disease." *Journal/Alzheimer´s & Dementia* 4, 2008, 65.

Sturm, R. "The effects of obesity, smoking and drinking on medical problems and costs." *Journal Health Affaire* 21, March-April 2002, 245.

Takahashi, K. and S. Yamanaka. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." *Journal Cell* 126, August 2006, 663.

Thomson, J. A., J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. "Embryonic stem cell lines derived from human blastocysts." *Journal Science* 282, November 1998, 1145.

VAS, C. J. "Alzheimer´s disease: The brain killer." World Health Organization, 2001.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith et al. "The sequence of the human genome." *Journal Science* 291, February 2001, 1304.

Wang, J., L. Ho, W. Zhao, K. Ono, C. Rosensweig, L. Chen, N. Humala, D. B. Teplow, and G. M. Pasinetti. "Grape-derived polyphenolics prevent Abeta oligomerization and attenuate cognitive deterioration in a mouse model of Alzheimer's disease." *Journal J Neurosci* 28, June 2008, 6388.

Wernig, M., J. P. Zhao, J. Pruszak, E. Hedlund, D. Fu, F. Soldner, V. Broccoli et al. "Neurons derived from reprogrammed fibroblasts functionally integrate into the fetal brain and improve symptoms of rats with Parkinson's disease." *Journal Proc Natl Acad Sci USA* 105, April 2008, 5856.

Wold, B. and R. M. Myers. "Sequence census methods for functional genomics." *Journal Nat Methods* 5, January 2008, 19.

Yahagi, N., H. Shimano, A. H. Hasty, T. Matsuzaka, T. Ide , T. Yoshikawa, M. Amemiya-Kudo et al. "Absence of sterol regulatory element-binding protein-1 (SREBP-1) ameliorates fatty livers but not obesity or insulin resistance in Lep(ob/Lep(ob) mice." *Journal J. Biol. Chem.* 277, May 2002, 19353.

Yamaguchi, K., L. Yang, S. McCall, J. Huang, X.X. Yu, S .K. Pandey, S. Bhanot et al. "Inhibiting triglyceride synthesis improves hepatic steatosis but exacerbates liver damage and fibrosis in obese mice with nonalcoholic steatohepatitis." *Journal Hepatology* 45, June 2007, 1366.

Zhang, Y., M. Proenca, M. Maffei, M. Barone, L. Leopold, and J. M. Friedman. "Positional cloning of the mouse gene obese and its human homologue." *Journal Nature* 372 December 1, 1994, 425.

cloning mammals: more than just another sheep

ALEXANDER KIND & ANGELIKA SCHNIEKE

Since the announcement of Dolly in 1997, cloning by nuclear transfer has received considerable media attention. But commentators have often been preoccupied with speculations about armies of cloned dictators or sportsmen, and minor applications such as replacement pets. Here we wish to place nuclear transfer in a broader context and outline the more subtle, but profound consequences of the work.

So, why was such an elaborate way of producing animals actually developed? Visitors to Scotland will notice there is hardly a shortage of sheep. There were in fact two motivations behind the Dolly experiment. One was strictly commercial, to develop a tool for rapid production of identical animals for biotechnology. The second and more powerful impulse was basic scientific curiosity and an opportunity to address a long-standing biological question. As complex animals, frogs, mice, sheep, and people arise from a single cell, and many different cell types are formed; how do they adopt their different fates and how do they maintain or change their identity?

Early investigations and founding principles

Scholars have pondered the question of animal development since ancient times. In the third century BC, Aristotle recognized the importance of sexual reproduction and proposed two alternative models. Either the structure of the whole animal is already preformed in miniature within the egg or embryo, or new structures arise progressively. Aristotle favored the second idea, but without suitable technology the question remained the subject of philosophical debate for centuries. Preformationism became the favored view in seventeenth- and eighteenth-century Europe, as illustrated by the seventeenth-century engraving in Figure 1. Stimulated by the discovery of sperm, or as they were termed at the time "animalcules" in seminal fluid, the physicist and early microscopist Nicholas Hartsoeker speculated about the possible structure of a tiny fetus within. Hartsoeker conjectured that the head of the sperm grew to form the fetus and the tail of the sperm formed the umbilical chord, while the function of the egg was merely to provide a nest supporting its development.

Reliable observations however only became possible after 1830 with the invention of the compound microscope by the British amateur naturalist Joseph Jackson Lister. Multiple lens instruments provided sufficient resolution to make out the detailed structure of living material for the first time. Modern biology arguably began in 1839 when Theodor Schwann and Matthias Schleiden demonstrated that living things are composed of cells. Shortly after, Albrecht von Kölliker showed that sperm and eggs (oocytes) are also cells, but how they interact to form a new organism was a mystery. The eminent chemist Justus von Liebig proposed that sperm transmit their male qualities to the oocyte through the vigorous vibrations of their tails. Then in 1854, George Newport described his observations of fertilization in frogs and suggested that sperm make their contribution by actually penetrating the egg. Around the same time, microscopic investigations were revealing that new cells arose by division of the fertilized egg, making it unlikely that development occurs by preformation.

Oskar Hertwig is credited with beginning the study of fertilization and early embryo development in the sea urchin, a very productive field that provided much of the information subsequently applied to other species. Sea urchins are ideal for microscopy because the eggs are very clear. In 1876, Hertwig described his observations of events following the addition of sperm to eggs. In particular, he noted the presence of two nuclei in the egg, one of which came from the sperm, and reported how they fused together. This provided the first explanation for the role of the two parents. It also focused attention on the importance of the nucleus, and the colored bodies within that could be revealed using the newly developed aniline dyes, and named "chromosomes" in the 1880s.

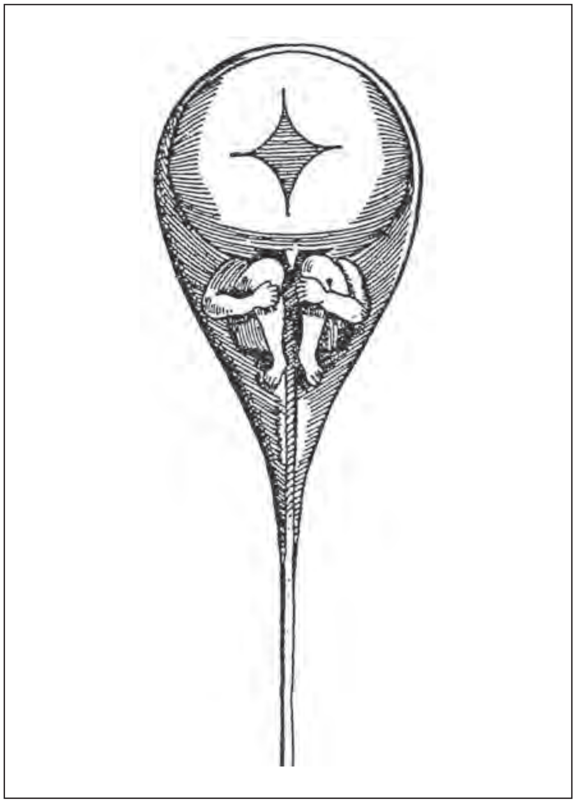


Figure 1. Preformed fetus within a sperm head. Hartsoeker, N. (1694). Essai de dioptrique, Paris.

The German biologist August Weismann ranks perhaps second only to Charles Darwin in his contributions to theoretical biology. In 1892 Weismann made the bold proposal that the nuclei of eggs and sperm contain a hereditary substance, and that this constitutes the only organic continuity between generations (Weismann 1892). This principle laid the foundations for all of genetics and evolutionary biology. Weismann's "germ-plasm theory" states that the germ cells are a lineage quite distinct from the other cells of the body, the somatic cells, and that characteristics acquired by the body during life are not transmitted to the germ cells. This was an explicit rejection of the views of Jean-Baptiste Lamarck that were widely held at the time, including by Darwin himself. Over the next twenty years, the strand of thought commenced by Weismann developed into the modern sciences of genetics and development. In 1900, Gregor Mendel's work on pea hybrids was rediscovered and with it his concept of discrete segregating traits. Within two years Theodor Boveri and Walter Sutton had both proposed that elements specifying Mendelian traits are located in the chromosomes. In 1907, Boveri demonstrated that a normal set of chromosomes is necessary to support embryonic development in the sea urchin. In 1915, Thomas Morgan described the physical location of genes on the chromosomes of the fruit fly in his masterwork *The Mechanism of Mendelian Heredity* (Morgan et al. 1915).

These ideas are now the stuff of basic biology courses, but during the twentieth century they were seen by some as "degenerate and fascist." Germ-plasm theory and all that followed were violently rejected by the ideologues of Soviet Russia. From the early 1930s until as late as 1964, official Soviet policy denied Weismann's ideas and the whole of genetics. Stalin is not primarily remembered for his interest in developmental biology, and it seems likely that this was just a political convenience. The inheritance of acquired characteristics allowed that the human race could be perfected through "progressive materialist" politics. Soviet leaders could therefore justify the hardships endured by ordinary people as worthwhile and necessary for the production of future generations of ideal Communists. This political atmosphere also favored the rise of the notorious Soviet agronomist Trofim Lysenko. In a ranting address to the Lenin All-Union Academy of Agricultural Sciences in August 1948, Lysenko denounced Weismann at length and mocked the "bourgeois pseudoscience" of his followers. In his ridicule, Lysenko inadvertently provided a reasonable description of the modern concept of the genome and a substance that would become known as DNA:

> Weismann denied the inheritability of acquired characters and elaborated the idea of a special hereditary substance to be sought for in the nucleus. The sought for bearer of heredity, he stated, is contained in the chromosome material. [...] An immortal hereditary substance, independent of the qualitative features attending the development of the living body, directing the mortal body, [...] that is Weismann's frankly idealistic, essentially mystical conception (Lysenko 1948).

Nature is however brutally indifferent to political theory. Lysenko's methods were responsible for repeated crop failures in the USSR and, when similar agricultural policies were adopted in China in 1958 under the "The Great Leap Forward," they contributed to the greatest famine in recorded history, between 1959–61.

Cloning and cell determination

Tied in with his concept of the germ plasm, Weismann offered the first testable theory of animal development, a process termed mosaic development. He proposed that the single cell embryo, the zygote, contains factors or determinants localized in discrete regions. As it cleaves, the determinants are distributed unequally between daughter cells and control their future development. The process continues as the various cell types form by "differentiation", as the embryo develops. This model clearly predicts that individual cells of the developing embryo should not share the same potential. However in 1892, Hans Driesch provided the first evidence against Weismannís theory (Driesch 1892). Cells of early sea urchin embryos could be separated and each form a whole larva. Division at the two-cell stage led to two normal larvae and individual cells from the four-cell stage produced four normal larvae. These were in fact the first experimentally cloned animals.

In a lecture presented at London University in October 1913, Driesch stated that the embryo is a "harmonious equipotential system [...] each element of which is capable of playing a number of different roles. The actual role it plays in any given case being a function of its position." Since then, there have been many demonstrations that the embryos of many vertebrates, including mammals, can be reorganized by changing the arrangement or the number of cells, and then recover to form a normal whole animal.

Cloning by nuclear transfer was first proposed as a further means of testing whether nuclei from early and late embryonic cells had equivalent developmental potential, and is a rather older idea than often supposed. Yves Delage, a now obscure French marine biologist, made the first recorded reference to the procedure in 1895, claiming "if, without any deterioration, the egg nucleus could be replaced by the nucleus of an ordinary embryonic cell, we should probably see this egg developing without changes." (Beetschen and Fischer 2004) However Delage is not known to have carried out such an experiment. The honor usually goes to Hans Spemann, a former student of Boveri. In 1928 Spemann performed the first nuclear transfer with a remarkable piece of microsurgery (Spemann 1928). Spemann's own drawings are shown in Figure 2. Using micro-tweezers and a loop of hair from his baby daughter, he constricted a single cell salamander embryo into two parts, one of which contained the cell nucleus (Figure 2CA). Left to develop, the portion with the nucleus divided and

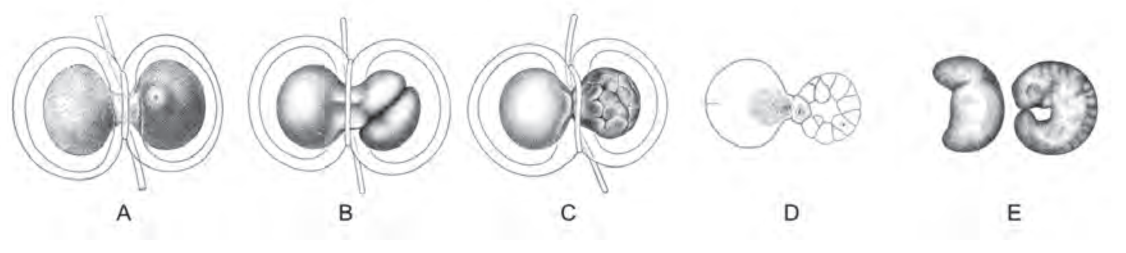


Figure 2. Hans Spemann's nuclear transfer experiment with salamander eggs and a baby hair. A) A loop was used to constrict a single cell embryo into two halves connected by a bridge of cytoplasm, the nucleus lies in the right half. B) First cell division in the nucleated right half. C) Later developmental stage in the nucleated half. D) During cell division one daughter nucleus passes through to the empty cytoplasm. E) Development now proceeds in both halves, but with a time delay, the embryo at left is younger than its twin. From Spemann, H. (1936).

formed an embryo, while the other side remained a pouch of clear cytoplasm (Figures 2CB and 2C). The embryo was allowed to develop to later stages, such as 16-cell, at which time a single nucleus was allowed to pass back into the empty cytoplasm (Figure 2D). The single cell then developed into a normal salamander embryo at a slightly earlier stage, (Figure 2E). This clearly proved that later embryonic cell nuclei are capable of forming a complete animal.

It is not known whether Spemann knew of Delage's earlier proposal, but in 1936 after his retirement, Spemann proposed what he termed "a fantastical cloning experiment." (Spemann 1936) If one could transfer nuclei from cells at yet later stages of development back into fertilized eggs it would be possible to systematically track when cells retained or lost their ability to form a whole organism, a quality now termed "totipotency."

The next decades continued to see many developmental biologists focus on sea urchins and amphibians, because they were relatively easy to culture and manipulate. Frog oocytes are very large cells, around 1–2 millimeters in diameter, and quite prominent as dark grey or brown spheres within their protective jelly coat. In the early 1950s, Robert Briggs and Thomas King carried out Spemann's fantastical experiment with frogs (Briggs and King 1952). They removed the nucleus from an activated oocyte, using a glass needle. A single cell dissected from a later stage embryo was then drawn up into a fine glass pipette connected by rubber tubing to an ordinary syringe. The cell broke open as it was squeezed within the pipette and the free nucleus was injected into the enucleated egg. Culturing the reconstructed embryos further, they found that cell nuclei from blastula stage embryos could direct normal development to feeding-stage larvae. But nuclei from later stage embryos, in which the major embryonic cell lineages such as mesoderm or endoderm were already established, were unable to do so.

John Gurdon and Ron Laskey later extended the work using nuclei from juvenile and adult tissues, such as the skin of the foot web, and generally found that these animals survived up to tadpole stage, but not much further. Gurdon did derive some adult frogs from tadpole intestinal tissue in 1962 (Gurdon 1962), but the possibility that germ cells were present in his tissue left the results in doubt. The overwhelming view at the time was that the developmental capacity of transplanted nuclei decreased with the age and extent of differentiation of the donor cell. Nuclei of the very early embryo may be equivalent, but at some stage their fate becomes determined, "hard wired" by some concrete change, such as the loss or irreversible modification of DNA in the nucleus.

This view was however difficult to reconcile with some well-known phenomena, notably the regenerative capabilities of most fish, and amphibians such as newts and salamanders. If a newt loses a limb, cells from surrounding tissues such as the skin migrate into the wound and undergo a process of "reverse development" dedifferentiating to form a blastema. This is a mass of undifferentiated cells that divide rapidly. Cells within the blastema then differentiate and re-organise to form a replacement limb. This was good evidence that some adult differentiated cells are not determined in their fate and can radically change their identity. Was limb regeneration fundamentally different to generating a whole animal by nuclear transfer? Or was the failure of nuclear transfer a result of technical rather than biological limitations? These open questions provided sufficient motivation for some researchers to continue probing cell determination.

Sheep lead the way

Most biologists are mammals, and naturally keen to investigate species closer to themselves than sea urchins and amphibians, but for many years this was just too difficult technically. Mammalian embryos grow within the controlled conditions of the female reproductive tract rather than pond or seawater and, although large compared to other cells at about one-tenth of a millimeter across, are barely visible to the naked eye. It took until the 1970s and 80s, when embryo culture and micromanipulation techniques had improved sufficiently, for transfer of mammalian nuclei to become practical. The basic idea remained as Spemann had conceived it, the genetic material is removed from an egg, and then replaced with the nucleus of another cell, often by fusing the whole cell with the oocyte.

The natural focus was on the favorite laboratory mammal, the mouse. However, attempts to repeat Briggs and Kings work in mice were repeatedly unsuccessful. In 1981, Karl Illmensee and Peter Hoppe claimed that they had cloned mice by transfer of nuclei from blastocyst stage embryos (Illmensee and Hoppe 1981). However, their work was later investigated and determined as false, although deliberate fraud was never proven. Then in 1984, James McGrath and Davor Solter seemed to put an end to mammalian nuclear transfer. They systematically transferred nuclei from 1-, 2-, 4-, 8-cell and blastocyst stage embryos into enucleated zygotes, 1-cell stage embryos. Nuclei from 1-cell embryos supported development to blastocysts, success dropped off sharply using 2-cell stage nuclei and failed entirely with later stages. This they reasonably interpreted as a rapid loss of totipotency during development. Their paper concludes with the categorical statement that "the cloning of mammals by simple nuclear transfer is biologically impossible." (McGrath and Solter 1984)

In retrospect, it was unfortunate that so many early efforts focused on mice. It has since become clear that they are one of the more difficult species to clone by nuclear transfer. This meant that, somewhat unusually, major breakthroughs were made using livestock. The first nuclear transfer mammals were three Suffolk sheep produced by Steen Willadsen, by merging single cells from 8-cell embryos with enucleated unfertilised eggs (Willadsen 1986). Ironically these lambs were born in 1984, just a few months before McGrath and Solter declared mammalian cloning impossible. The reason for this discrepancy was technical. McGrath and Solter had used enucleated zygotes for nuclear transfer, because mouse oocytes are too fragile to survive nuclear transfer. Willadsen had been able to use unfertilized oocytes, which are more robust in sheep. Years of work have since shown that unfertilized oocytes are successful recipients for nuclear transfer in numerous species, while zygotes can only be used at a very particular stage. Only this year has a model been proposed to explain this difference, as we discuss later (Egli, Birkhoff, and Eggan 2008).

During the next decade nuclear transfer was carried out in a variety of mammals, but, as in the frog, it was only successful using cells obtained directly from very early embryos, or cultured for very short periods.

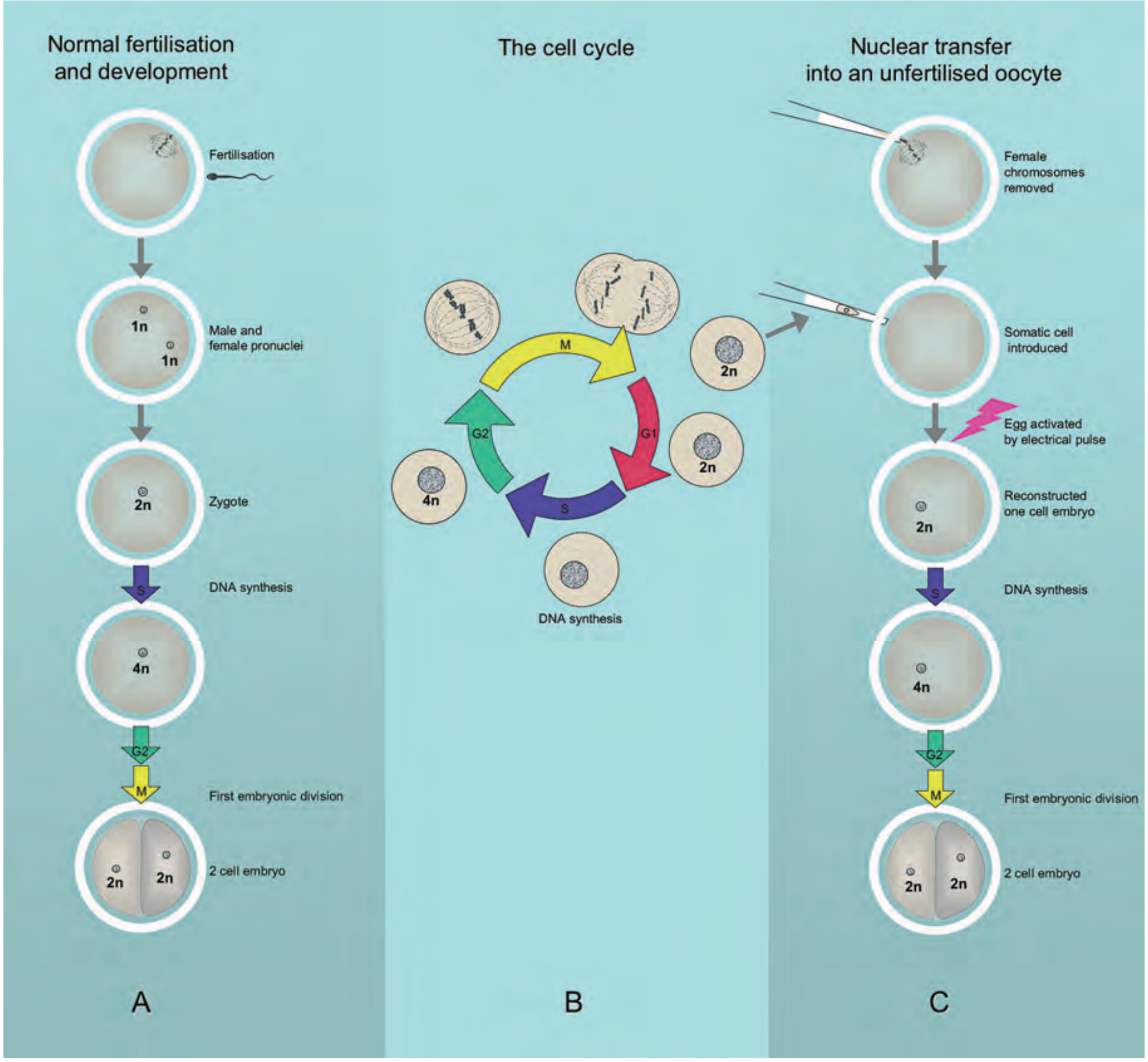


Figure 3. Nuclear transfer and the cell cycle. A) Normal fertilization and development to the 2-cell stage embryo. B) The cell cycle. G1 phase is followed by S phase where the cell duplicates each chromosome, then G2 phase, and mitosis (M) in which the nucleus breaks down, the duplicated chromosomes condense, align on the spindle and are distributed into two new daughter cells. C) Nuclear transfer using a donor cell in G1 phase and development to 2-cell stage embryo. 1n, 2n, 4n = copies of each chromosome.

In the early nineties, Keith Campbell and Ian Wilmut of the Roslin Institute near Edinburgh began to study how the choice of oocyte recipient and the cell cycle stage of the nuclear donor cell affected the outcome of nuclear transfer. This made a very significant contribution to the eventual success of nuclear transfer in mammals, so here we outline the main points.

Mammalian oocytes are formed from the germ cells by a process termed meiosis, which leaves each oocyte with only one copy of each chromosome, usually abbreviated as "1n." When the head of the sperm enters, it causes the oocyte to complete meiosis and initiates the first events of development, illustrated in Figure 3A. The two clusters of male and female chromosomes first form into separate pro-nuclei, which then join to create a single nucleus, in what is now the one cell embryo or zygote. All chromosomes are then replicated by DNA synthesis ready for the first embryonic cell division.

This first division and all subsequent cell divisions that form and maintain the body of the animal is by a process termed mitosis. Mitosis is part of a cycle that ensures that dividing cells maintain the correct number of chromosomes. This "cell cycle" is conventionally divided into four phases, as outlined in Figure 3B. The first phase is termed gap1 (G1), during which the cell has two copies of each chromosome (2n). In the next phase, synthesis (S), the cell replicates all its DNA. Then follows gap2 (G2) phase, when each chromosome is present as four copies (4n). At mitosis (M), the nucleus breaks down, the duplicated chromosomes condense, align on a structure termed the spindle, and are then pulled apart into two new daughter cells. Each new cell contains 2n chromosomes and the process repeats. In rapidly dividing cells the whole cycle takes about one day.

This has profound implications for nuclear transfer. When a cell nucleus is transferred into an unfertilized oocyte it has to be artificially activated, e.g. by an electrical pulse, to kick start development. This initiates DNA synthesis in readiness for the first cell division. However, this occurs regardless of the cell cycle stage of the donor nucleus. If the incoming nucleus was in S or G2 phase, when DNA has already been partially or completely replicated, its DNA will be re-replicated, leading to incorrectly high chromosome numbers, or serious chromosomal damage. The key insight by Campbell and Wilmut was that only donor cells in G1 phase (prior to DNA replication) would support normal development in unfertilized oocytes, as shown in Figure 3C.

The method they developed, and still widely used, is to starve the donor cells of growth factors by reducing the amount of serum in the culture medium for a few days. This blocks the cell cycle before S phase, exactly what is required. Importantly, effective serum starvation requires that cells are cultured for several days.

1995 saw the birth of Megan and Morag at the Roslin Institute, two lambs made by transfer of nuclei from cells grown by Jim McWhir from a day 9 sheep embryo and cultured for 6 to 13 passages (Campbell et al. 1996). These sheep prompted debate amongst the coauthors about what the key factor was that had led to success. Campbell and Wilmut's view was that serum starvation before nuclear transfer not only coordinated the cell cycle, but also induced a quiescent state in the nucleus making it particularly amenable to reprogramming by the oocyte. McWhir contended that the key lay in some special property of the cells he had derived.

Sheep nuclear transfer is dictated by the natural breeding season and so the question could not be resolved until the next year. The original plan for 1996 was to use embryonic cells again and also test whether the technique could be extended to cells at a later developmental stage, fibroblasts from a day 26 fetus. At that time, we were working with PPL Therapeutics, a biotechnology company in the business of producing pharmaceutical proteins in the milk of transgenic sheep, a short walk away from the Roslin Institute. In a discussion over lunch we suggested a bolder experiment and proposed including adult cells in the 1996 nuclear transfer season. This met with skepticism and the opinion that it was too soon, and anyway no finance was available to extend the experiment. There was however the possibility that if the extra work could be justified in commercial terms, our company might provide funds. As it happened, we were then investigating introducing milk transgenes into sheep mammary epithelial cells derived by Colin Wilde of the Hannah Research Institute in Ayr, as a means of testing their expression. Combining the two projects offered an ideal opportunity. If the mammary cells could be converted into live animals, PPL would have a potential means of producing "instant flocks" of sheep known to express a particular transgene. And, most excitingly, using adult cells for nuclear transfer would address the long-standing question of cell determination. The case was made to the managing and research directors of PPL, Ron James and Alan Colman, and the company took the risk of releasing funds for the experiment. In February 1996, cultures of sheep mammary cells and also cultured embryonic cells were serum-starved and transported over to the Roslin Institute. Bill Ritchie, Ian Wilmut's skilled technician, then transferred them into Scottish Blackface enucleated oocytes.

A single female lamb was born on July 5, 1996, and named Dolly by the animal caretaker John Bracken, in honor of Dolly Parton and her great singing talent. Two sheep were also born from fetal fibroblasts and four from embryonic cells. This clearly showed that success was clearly not due to any special cell type; the idea that quiescence played a role was also later discarded. What did emerge was the importance of cell synchronization.

A description of the experiment was published on February 27, 1997 (Wilmut et al 1997). More than a decade later, one can calmly state that nuclear transfer from an adult cell caused the concept of irreversible cell determination to be discarded. However, this would be to overlook the sometimes heated contention that raged for 17 months following publication. So fixed was the view of cell determination that several prominent scientists in the US and Europe dismissed the paper outright as a fraud, perhaps recalling the Illmensee controversy. An article in the New York Times from July 29, 1997, gives a sense of the mood: "How do we know the whole thing wasn't a hoax? Why, some ask, is the rest of the world so willing to accept the world-shattering claim that an adult animal was cloned?"

Other commentators interpreted the inefficiency of adult nuclear transfer as an indication that Dolly was a one-off, an experimental aberration. Several eminent scientists suggested that she was not actually cloned from an adult cell, but had instead arisen from some contaminating embryonic or fetal material. One proposition was that fetal cells present in the blood circulation of the sheep used to provide the mammary cells had somehow carried through into the mammary cell cultures. It seemed that any alternative explanation, no matter how unlikely, was preferable to overturning the doctrine of cell determination. *Time* magazine ran an article on March 2, 1998, headed "Was Dolly a mistake?" that concluded: "Dolly, in other words, may turn out to be a fluke, not a fake. No matter what she is, it's looking less and less likely that we're going to see clones of Bill Gates or Michael Jordan anytime soon."

Meanwhile, we—along with others—had reported more cloned animals (Schnieke et al. 1997), but these were from cultured fetal cells and so did not confirm adult cloning.

The accusations and speculations thankfully ceased on July 23, 1998. That day's edition of the journal *Nature* contained two relevant articles. One gave the results of an independent DNA fingerprinting analysis, confirming that Dolly's nuclear DNA was identical to the cultured mammary cells (Signer et al. 1998). The second was a report by Ryuzo Yanagimachi and Teruhiko Wakayama of the University of Hawaii describing another animal cloned from adult cells, a mouse named "Cumulina" after the cumulus (ovarian follicle) cells used as donors (Wakayama et al. 1998). The reality of adult cloning was finally accepted. More than a decade has now past and it is timely to revue the developments that followed.

Reproductive cloning

Inevitably, most public, political, and ethical discussions have centered on human reproductive cloning and as a result new laws and regulations are in place around the world. It is perhaps worth emphasizing that, despite announcements from odd cult groups and publicity seekers, none of the scientists actually working in the field ever considered carrying out human reproductive clonin.

CLONED MAMMALS
Cattle, deer, domestic cat, dog, ferret, goat, gaur, horse, mouse, mouflon, mule, pig, rabbit, rat, rhesus monkey, sheep, water buffalo, wildcat, wolf.

As is often the case, once a method is well established, it is difficult to see why it was once viewed as impossible. A variety of different cell types, both fetal and adult, have now been successfully used as nuclear donors and over 20 species cloned, including fish, frogs, fruit flies, and the mammals listed in the table. The efficiency is however low in most species, with only 1–5% of reconstructed embryos proceeding to birth. Nuclear transfer animals can suffer ill health, but their offspring, such as Dolly's lamb Bonny, do not.

The majority of nuclear transfer experiments are still carried out on livestock. The improvements made have been incremental rather than dramatic, but have nevertheless resulted in success rates of ~15% in cattle. One important factor has been the advance of oocyte maturation techniques. All livestock oocytes are now obtained from ovaries collected from slaughterhouses rather than flushed from the reproductive tract of live animals. This offers an abundant supply of oocytes and greatly reduces the number of animals required. It also makes commercial nuclear transfer viable despite its inefficiency, especially for the reproduction of elite animals with highly desirable characteristics such as racehorses or prize bulls. In the US, companies such as ViaGen offer livestock cloning as part of their assisted reproduction services. Their website (www.viagen.com) states: "ViaGen enables the owners of cattle, horses and pigs to preserve and multiply their best genetics through gene banking and cloning services, and to protect their brands through genomic services." Devoted (and wealthy) dog owners might also be interested to know that "by calling the toll-free number 888-8VIAGEN you can discuss options for cloning your pet dog."

Nuclear transfer has been used where normal sexual reproduction is impossible as a result of accident, disease, or natural infertility, as demonstrated by the cloning of a mule (Woods et al. 2003). It has also been applied to reproduce endangered species such as the European wildcat or rare cattle breeds. However, only those species with domesticated relatives available to provide suitable oocytes are likely to benefit. Cloning also cannot improve the genetic diversity of small animal populations, vital for long term survival.

Animals for biomedicine

The future will show whether or not nuclear transfer will become just one more, albeit expensive, routine technique for animal reproduction. But it has already become one of the best methods of founding lines of genetically modified "transgenic" large animals. To produce transgenic animals, there are two choices. Transgene DNA can be introduced directly into the zygote, hoping that it becomes incorporated into the genome. Alternatively, genetic modifications can first be engineered into cultured cells that are then used to produce whole animals. The first method is basically a lottery, animals must be produced before analyzing whether a transgene is present. The second method allows much more control and involves fewer animals, because cells can be thoroughly analyzed in the laboratory before any animals are produced.

In mice, a cell-based method has been available since the early eighties. Mouse embryonic stem (ES) cells can be isolated from early embryos, grown indefinitely in culture, undergo manipulations such as the addition of a transgene or the precise alteration of a particular gene (gene targeting), and then be incorporated back into a developing embryo. The phenomenal power of gene targeting technology in ES cells has provided most of our knowledge about the function of genes in whole animals. This was recognized by the 2007 Nobel Prize for Medicine, awarded jointly to Mario Capecchi, Martin Evans, and Oliver Smithies. It has long been clear to many researchers that extending this to large animals would have many useful applications. But despite considerable efforts, functional ES cells had (and still have) not been derived from livestock. Livestock nuclear transfer using ordinary somatic cells that could be grown and manipulated in culture neatly sidestepped this.

In the experiments directly following Dolly, we demonstrated that both transgenic and gene targeted sheep can be generated by nuclear transfer (Schnieke et al. 1997a; Schnieke et al. 1997b). Since then, many others have followed, e.g. gene targeted cattle resistant to mad cow disease (bovine spongiform encephalopathy) (Kuroiwa et al. 2004). Most applications have however been in the area of biomedicine, including rather later than anticipated, transgenic animals producing pharmaceutical proteins in milk. ATryn, an anticoagulant drug used to treat patients with a hereditary deficiency in the blood protein antithrombin, is being produced in transgenic goats from a cloned founder animal, and was launched on the market in November 2007 by GTC biotherapeutics. Nuclear transfer is also being used to engineer multiple genetic modifications into pigs to provide cells or organs for transplantation into humans, termed xenotransplantation.

A number of large animal models of serious human diseases such as cystic fibrosis (Rogers et al. 2008), diabetes, and various cancers are also being developed. These are often extensions of work first carried out in mice, where gene targeting has provided a huge amount of information regarding diseases such as cancers. Many strains of mice with defined genetic defects have been produced and these have been very valuable in understanding mechanisms such as tumor initiation and progression (Frese and Tuveson 2007). Mice are also useful for proof-of-principle studies for novel diagnostic and treatment strategies, as we discuss later. However the major differences in body size, general physiology, anatomy, diet, and lifespan restrict the usefulness of mice closer to the clinic. For example, radiation and thermal therapy cannot easily be scaled down to treat mouse-sized tumors. Nuclear transfer offers the opportunity to extend the range of genetically-defined disease models to other species, such as pigs, which more closely resemble humans in their size, anatomy, and physiology.

Nuclear transfer, embryonic stem cells, and regenerative medicine

As mentioned above, much of the interest in nuclear transfer in the late eighties and early nineties was prompted by the possibilities offered by ES cell technology. Since then, the two fields have been closely intertwined.

ES cells are usually isolated from blastocyst stage embryos. A blastocyst is a tiny fluid filled ball of about one hundred cells containing within it a clump of cells termed the inner cell mass (ICM) that gives rise to all tissues of the body. Blastocysts, or isolated ICMs, are placed in culture and over a period of several days or a week, colonies of tightly packed small cells emerge and continue to grow indefinitely; these are ES cells. For reasons that are unclear, deriving ES cells is however difficult in many species and has only been achieved in mouse, human, and rhesus monkey. Rat ES cells have been announced to the press, but not yet published in a scientific journal.

ES cells are often used as a convenient surrogate for the study of the early embryo. But what they actually are is still not clear. They may be an artifact of tissue culture, something aberrant created in response to artificial growth conditions. Recent evidence however suggests they are a cell type normally present for a short time in the embryo, which can be captured and maintained by the proper culture conditions (Silva and Smith 2008).

The defining characteristic of ES cells is that they can grow indefinitely as undifferentiated cells and then differentiate to many other cell types. When introduced into a blastocyst they can integrate into the ICM and participate in forming all tissues of the body. When given appropriate stimuli they can also form a wide variety of cell types in culture, so called *in vitro* differentiation. Since human ES cells were first derived by Jamie Thomson ten years ago (Thomson et al. 1998) there has been intense interest and enthusiasm surrounding *in vitro* differentiation as a possible source of replacement human tissue, such as nerve cells, insulin producing cells, or heart muscle. Many articles have been written on the subject, so we will not go into detail here. The basic scheme is shown in panel A of Figure 4. The promise of ES based therapy is undoubtedly real, but a few words of caution are perhaps in order. Persuading human ES cells to form useful amounts of an appropriate, properly characterized, and pure therapeutic cell-type remains a very difficult challenge. Rigorous methods also need to be established to ensure that ES derived preparations are free of potentially tumor-forming cells. Research scientists and the biotech industry should be realistic and avoid the tendency to hype their claims.

The Californian pharmaceutical company Geron is perhaps farthest advanced in human ES cell therapies. The company website (www.geron.com) reports the development of human ES derived nerve cell progenitors for acute spinal cord injury and cardiomyocytes for the treatment of heart failure. Although Geron have applied for permission to carry out human clinical trials of their nerve cell progenitors, the United States Food and Drug Administration have placed the application on hold. If and when trials are approved, the outcome will be very important for the future of ES cell therapy.

If they can be produced, ES derived tissues would need to be immunologically matched to the patient in the same way as ordinary donated tissue to avoid rejection. Recipients are also likely to require lifelong immune suppression. Tissue matching is a particular problem for patients with unusual tissue types, such as people of mixed race. Shortly after Thomson's report, it was suggested that nuclear transfer could provide a means of producing tailor-made human tissues by "therapeutic cloning." Cells could be taken from a human patient who needed tissue replacement therapy and used to produce cloned embryos. ES cells would be derived and then induced to differentiate in culture. The replacement tissue would be perfectly matched to the patient's own body, see Figure 4B.

Some of the necessary technical steps have been achieved in animals. For example, ES cells have been derived from nuclear transfer mouse embryos and found to be the same as those from normal embryos. Rhesus monkey ES cells have also been produced from cloned embryos, but so far no human "NT ES" cells have been derived. The major practical problem is the supply of unfertilised human oocytes, which is already insufficient to meet the existing needs of people requesting assisted reproduction techniques such as *in vitro* fertilisation (IVF). A recent news item in *Nature* revealed that, despite two years and $100,000 spent on local advertising, stem cell researchers at Harvard University have managed to secure only one egg donor (Maher 2008).

Therapeutic cloning is therefore unlikely to be realized unless an alternative source of recipient oocytes can be found. Animal, particularly cattle, oocytes are plentiful thanks to *in vitro* maturation. The UK Human Fertilization and Embryology Authority (HFEA) recently approved research into whether these could be used, but many people object to the creation of cytoplasmic hybrid embryos. Biological problems may also arise from incompatibility between the components of the animal oocyte and the incoming human nucleus. Reprogramming factors and important cellular organelles such as the mitochondria may not function properly. Perhaps the most promising source of oocytes is *in vitro* maturation of immature human oocytes from donated ovaries. Although less advanced than in cattle, *in vitro* maturation of human oocytes is improving, being used mainly to help women who must undergo ovariectomy. Several normal births from *in vitro* matured oocytes have been reported.

Despite their possible benefits, human embryonic stem cell derivation and therapeutic cloning both face serious ethical and religious opposition, and it is clear that many people will accept neither the artificial generation nor destruction of human embryos, even if only tiny balls of cells. Laws vary around the world, for example the UK HFEA gave approval in 2007, the Spanish Ministry for Health approved the work in 2008, while Germany has no plans to legalize the procedure. Another problem is the high cost and considerable time required. Therapeutic cloning would most probably be restricted to wealthy patients and then only those whose disease condition allowed them to wait several months. This said, recent advances in nuclear reprogramming have probably already made therapeutic cloning obsolete.

Understanding reprogramming

A human body contains several hundred cell types, each of which differ in a multitude of cellular components. The identity of a cell, its shape, how fast it divides, the materials it synthesizes, the receptors on its surface, and the multifarious nanomachines we call RNA and protein molecules, are all ultimately the product of different patterns of gene expression.

Cloning from adult cells showed that these patterns are not due to immutable genetic differences. The nuclei of even the most highly differentiated cells, such as neurons, or mature B-lymphocytes specialized to synthesize a single antibody, retain the potential to form all the cells of the body (Hochedlinger and

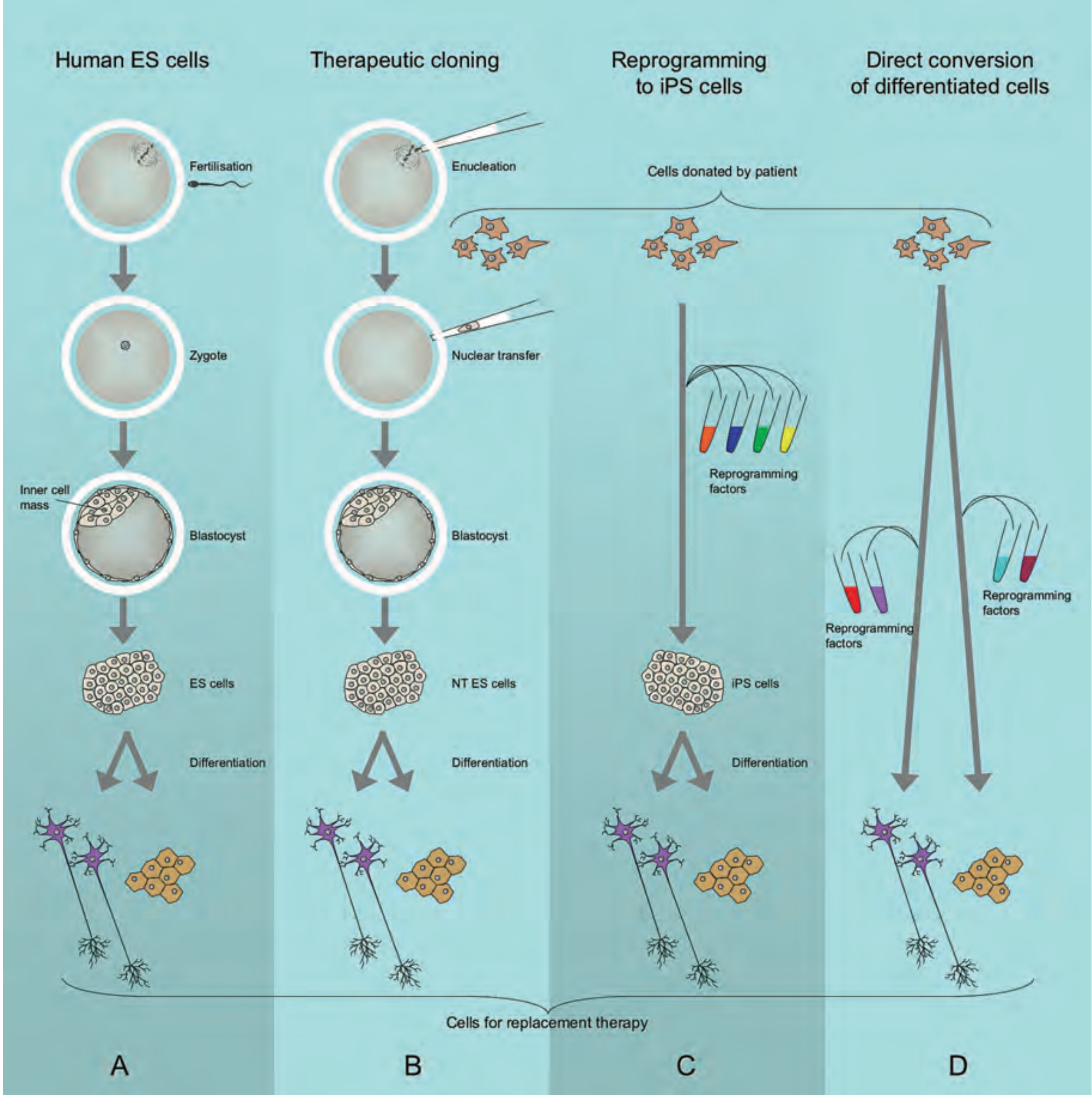


Figure 4. Generation of differentiated cells for disease research and cell therapy. A) Standard human ES cell derivation and differentiation. B–D) Derivation of patient specific cells. B) Therapeutic cloning. C) iPS cells. D) Direct conversion of one differentiated cell type to another. ES, embryonic stem cells; NT-ES, nuclear transfer derived ES cells; iPS, induced pluripotent stem cells.

Jaenisch 2002). Each cell has the same genetic information, in humans around three billion DNA base pairs and an estimated 25,000 genes, but these are expressed in different ways. A parallel may be drawn with computer software and hardware. Different programs dealing with graphics, mathematics, music, or word processing can all run on the same machine without altering the physical components. Of course, unlike computers, cells are self-organizing and no one would suggest that a cell receives a complete set of instructions from some outside agency.

The regulation of gene expression has been studied for over thirty years, and is known to operate at many levels: from the accessibility of the DNA within the nucleus to expression factors; the rate at which genes are transcribed into messenger RNA molecules; the processing and transport of RNA; to the synthesis and degradation of protein products. A glance through the diagrams in a modern molecular or cell biology textbook reveals multitudes of arrows symbolizing the regulatory pathways and feedback loops that govern the workings of our cells. Terms often used are "molecular circuitry" or "gene networks." A cell's identity is the outcome of a complex web of interactions and is a dynamic, not a static condition. Regulatory factors are known to constantly cycle between the nucleus and the cytoplasm, affecting their own expression and other genes. So, a cell can be viewed as constantly updating its own program of gene expression. Where the regulatory cycle promotes continuation of a particular state, it is stable. And in the same way that a physical net can adopt different shapes when pulled or pushed, there are many stable patterns of gene expression and many differentiated states.

In a particular cell, some genes are highly expressed, some less so and others not at all. How that pattern is maintained and reliably passed onto daughter cells is incompletely understood. It is however known that chromatin, the complex of DNA wound around histone proteins, carries marks that relate to whether genes are active or inactive. These marks are "epigenetic" rather than genetic, in that they do not alter the actual DNA sequence. For example, DNA in and around inactive genes often carries methyl groups added onto the nucleotide cytosine. Active and inactive genes also show different chemical modifications to histones. This affects how tightly DNA is bound and how "open" and available it is to transcription factors.

When a nucleus of a differentiated cell is exposed to a foreign environment, for example when two cells are fused together, the regulatory processes are disrupted and the gene expression pattern alters accordingly. For example, a nucleus from a human liver cell can be induced to express muscle genes by fusion with a mouse muscle cell (Blau, Chiu, and Webster 1983) and the nuclei of several different somatic cells express embryonic genes when fused to ES cells (Do, Han, and Schöler 2006).

Nuclear transfer may be regarded as a more complete version of the same phenomenon. When a nucleus is transferred into an enucleated oocyte it undergoes comprehensive removal of the DNA methyl groups and major changes in histone modifications, thoroughly erasing its previous identity. Kevin Eggan and colleagues propose that the key to such successful reprogramming is the free availability of factors regulating gene transcription (Egli, Birkhoff, and Eggan 2008). These are normally associated with the DNA within the nucleus, but are released into the cytoplasm when the nucleus breaks down, ready to be distributed with the chromosomes into the two new nuclei. Unfertilized oocytes have an abundance of such free factors, being primed and ready to reprogram an incoming nucleus—the sperm head.

But what are the factors responsible for reprogramming a nucleus to an embryonic state? Unfortunately mammalian oocytes are tiny, do not propagate, and therefore difficult to analyze with current technology. So, researchers have turned to ES cells.

Direct reprogramming, a radical new approach

Years of intensive study have revealed much about the mechanisms that maintain ES cells in an undifferentiated state and trigger their differentiation. In 2006, this culminated in a major breakthrough. Shinya Yamanaka and colleagues of Kyoto University reasoned that regulatory factors known to be important in keeping ES cells undifferentiated would be good candidates for reprogramming factors. His group identified 24 regulatory genes and constructed viral vectors to transduce them individually into other cells. Different combinations of genes were then introduced into mouse fibroblasts and the cells selected for the expression of a gene characteristically expressed in ES cells. A set of four transcription factors: Sox-2, Oct-4, c-Myc, and Klf4, were found to convert the fibroblasts into something closely resembling ES cells, which they named induced pluripotent stem (iPS) cells (Takahashi and Yamanaka 2006) (Takahashi et al. 2007). Since their original report, Yamanaka and several other research groups have refined the technique and extended it to human cells (see Figure 4C). At the time of writing, the general opinion is that iPS cells and ES cells are essentially the same. However, these are early days and some critics have pointed out differences that may be significant (Liu 2008).

The discovery of a remarkably simple recipe to reprogram differentiated cells into an embryonic state has sparked an almost unprecedented frenzy of research activity around the world. Unlike nuclear transfer, there are no ethical concerns and the techniques are straightforward, opening the study of reprogramming to many laboratories. It has also caused some leading groups previously investigating therapeutic cloning to shift their research focus. The US *Boston Globe* of 1 August 2008 quotes Rudolf Jaenisch saying that the iPS approach "is so much easier, [with] so many fewer restrictions and problems—ethical as well as others, [...] I think we'll probably be moving in this direction."

IPS cells are now such a fast moving area that this account will be out of date almost as soon as it is printed. But some important features are emerging as the story unfolds. At first it seemed that some cells could be reprogrammed and others not. But iPS cells have now been made from many cell types, such as mature B lymphocytes and pancreatic islet beta cells, demonstrating that it is not a quirk of some particular cell type, or an experimental artifact as some skeptics had claimed. Somewhat puzzlingly, different research groups are finding that widely different combinations and numbers of factors are effective. The underlying mechanisms are still obscure, but unlikely to remain a mystery for long. Bioinformatic analysis of whole constellations of genes is revealing the patterns of expression and the details of the regulatory networks that characterize ES and iPS cells and the events involved in direct reprogramming (Mikkelsen et al. 2008; Müller et al. 2008).

An immediate application of iPS cells is the study of degenerative diseases. Human iPS cells have already been isolated from patients with conditions such as motor neuron disease, Parkinson's disease, Duchenne muscular dystrophy, and juvenile onset (type I) diabetes (Dimos et al. 2008; Park et al. 2008), and are being used to generate dishes of the affected cell type in the laboratory. The ready availability of disease-specific iPS cells will have a profound impact on the understanding and treatment of many serious disorders. It will allow the effect of environmental factors such as food additives, toxins, or pathogens on cell degeneration to be thoroughly examined in large-scale studies. Large numbers of drugs can also be screened to identify those that halt, slow, or reverse disease progression.

IPS cells have also raised considerable excitement as a source of replacement tissues without the need for human eggs or ES cells. In a proof-of-concept experiment, Rudolf Jaenisch successfully treated sickle-cell anemia in mice (Hanna et al. 2007). Skin cells from a mouse with sickle-cell anemia were converted to iPS cells and the genetic defect corrected by gene targeting. The iPS cells were induced to differentiate into blood stem cells and then transplanted into the mouse where they reduced anemia and increased survival.

However, it must be stressed that such iPS based cell therapies are still some way from the human clinic. Current methods of producing iPS cells involve potentially dangerous cancer genes, and alternatives must be found. Encouragingly, there are early reports that chemical agents can replace the need for some of the genes in the original recipe, and variations such as adding instructive RNA rather than genes are also being explored.

If iPS cells are identical to ES cells, they necessarily face the same issues regarding therapeutic use. Like ES cells, undifferentiated iPS cells can form tumors and so must be absolutely excluded from any therapeutic preparation. Methods must also be worked out to induce the differentiation of pure populations of therapeutic cell types. Differentiation conditions have been established for some, such as motor neurons, but procedures need to be worked out for many other potentially useful cells.

The two-year history of the iPS revolution has been breathtaking and has almost certainly made therapeutic cloning obsolete. But there are signs of yet another step change. While the oocyte was a black box that did not easily reveal its workings, the derivation of iPS cells has opened wide the study of reprogramming. Knowledge of the normal development of many cell types is also constantly improving and the role of key regulatory molecules becoming clear, allowing them to be used as "knobs and switches" to control cell identity. If the aim is to produce differentiated cells to order, why not do it directly without going through an embryonic intermediate? In a dramatic paper (Zhou et al. 2008) published in August 2008, Doug Melton and colleagues reported treating diabetic mice with three instructive genes carried in viral vectors. This induced some of the exocrine cells in the mouse pancreas, which normally secrete digestive enzymes, to convert directly to insulin producing beta cells without any intermediate formation of iPS, or other ES-like cells. Clearly this work is at a very early stage, but it has opened yet another route to the production of cells on demand (see Figure 4D). Most provocatively, because the work was carried out in mice not in culture there is now the prospect of patients requiring replacement therapy being able to ingest a cocktail of instructive factors designed to generate new cells within their own body with no need for transplantation. Could the regeneration of whole organs also now be on the horizon?

Concluding remarks

In a recent essay (Thomas 2007), John Meurig Thomas outlined the basic unpredictability of scientific progress and the tortuous paths that often lie between original research findings and the development of familiar modern devices and procedures. Famously, following their 1958 paper, Charles Townes and Arthur Schawlow foresaw no practical application for their invention, the optical laser.

Cloning was originally conceived to investigate the determination of cell identity and fate, but is now leading to the ability to change cell fate. Who knows what the most important legacy of Dolly will prove to be? After a little over eleven years, what is most evident is the intense attention she has brought to these questions, and the general sense of wide-open possibility and excitement. Many talented people have been attracted to this research and inspired to embark on projects that would have been inconceivable before 1997. We are optimistic that the current buzz of activity will bring significant advances in medicine and benefit to human health.

Bibliography

Beetschen, J. C., and J. L. Fischer. "Yves Delage (1854–1920) as a forerunner of modern nuclear transfer experiments." *Int. J. Dev. Biol.* 48, 2004, 607–612.

Blau, H. M., C. P. Chiu, and C. Webster. "Cytoplasmic activation of human nuclear genes in stable heterocaryons." *Cell* 32, 1983, 1171–1180.

Briggs, R., and T. J. King. "Transplantation of living nuclei from blastula cells into enucleated frogs' eggs." *Proc. Natl. Acad. Sci. USA.* 38, 1952, 455–463.

Campbell, K. H., J. Mcwhir, W. A. Ritchie, and I. Wilmut. "Sheep cloned by nuclear transfer from a cultured cell line." *Nature* 380, 1996 64–66.

Dimos, J. T., K. T. Rodolfa, K. K. Niakan, L. M. Weisenthal, H. Mitsumoto, W. Chung, G. F. Croft, et al. "Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons." *Science* 321, 2008, 1218–1221.

Do, J. T., D. W. Han, and H. R. Schöler. "Reprogramming somatic gene activity by fusion with pluripotent cells." *Stem Cell Rev.* 2, 2006, 257–264.

Driesch H. "Entwicklungsmechanisme Studien. I. Der Werth der beiden ersten Furchungszellen in der Echinodermentwicklung. Experimentelle Erzeugen von Theil und Doppelbildung." *Zeit. für wiss. Zool* 53: 160–178, 1892, 183–184.

Egli, D., G. Birkhoff, and K. Eggan. "Mediators of reprogramming: transcription factors and transitions through mitosis." *Nat. Rev. Mol. Cell. Biol.* 9, 2008, 505–516.

Frese, K. K., and D. A. Tuveson. "Maximizing mouse cancer models." *Nat. Rev. Cancer* 7, 2007, 645–658.

Gurdon, J. B. "The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles." *J. Embryol. Exp. Morphol.* 10, 1962, 622–640.

Hanna, J. M. Wernig, S. Markoulaki, C. W. Sun, A. Meissner, J. P. Cassady, C. Beard, et al. "Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin." *Science* 318, 2007, 1920–1923.

Hochedlinger, K., and R. Jaenisch. "Monoclonal mice generated by nuclear transfer from mature B and T donor cells." *Nature* 415, 2002, 1035–1038.

Illmensee, K., and P. C. Hoppe. "Nuclear transplantation in Mus musculus: developmental potential of nuclei from preimplantation embryos." *Cell* 23, 1981, 9–18.

Kuroiwa, Y., P. Kasinathan, H. Matsushita, J. Sathiyaselan, E. J. Sullivan, M. Kakitani, K. Tomizuka, I. Ishida, and J. M. Robl. "Sequential targeting of the genes encoding immunoglobulin-mu and prion protein in cattle." *Nat. Genet.* 36, 2004, 775–780.

Liu, S.V. "iPS cells: a more critical review." *Stem Cells Dev.* 17, 2008, 391–397.

Lysenko, T. D. *Soviet Biology: Report and concluding remarks to the 1948 session of the Lenin Academy of Agricultural Sciences.* (English ed.) London: Birch Books, 1948. Online version: www.marxists.org/reference/archive/lysenko/works/1940s/report.htm

Mcgrath, J., and D. Solter. "Inability of mouse blastomere nuclei transferred to enucleated zygotes to support development in vitro." *Science* 226, 1984, 1317–1319.

Maher, B. "Egg shortage hits race to clone human stem cells." *Nature* 453, 2008, 828–829.

Mikkelsen, T. S., J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner. "Dissecting direct reprogramming through integrative genomic analysis." *Nature*, 454, 2008, 49–55.

Morgan, T.H., A. H. Sturtevant, H. J. Muller, and C. B. Bridges. *The Mechanism of Mendelian Heredity.* New York: Henry Holt and Co., 1915.

Müller, F. J., L. C. Laurent, D. Kostka, I. Ulitsky, R. Williams, C. Lu, I. H. Park, et al. "Regulatory networks define phenotypic classes of human stem cell lines." *Nature*, 455, 2008 401–405.

Park, I. H., N. Arora, H. Huo, N. Maherali, T. Ahfeldt, A. Shimamura, M. W. Lensch, C. Cowan, K. Hochedlinger, and G. Q. Daley. "Disease-specific induced pluripotent stem cells." *Cell* 134, 2007, 877–886.

Rogers, C. S., Y. Hao, T. Rokhlina, M. Samuel, D. A. Stoltz, Y. Li, E. Petroff, et al. "Production of CFTR-null and CFTR-DeltaF508 heterozygous pigs by adeno-associated virus-mediated gene targeting and somatic cell nuclear transfer." *J. Clin. Invest.* 118, 2008, 1571–1577.

Schnieke, A. E, A. J. Kind, W. A. Ritchie, K. Mycock, A. R. Scott, M. Ritchie, I. Wilmut, A. Colman, and K. H. S. Campbell. "Human factor IX transgenic sheep produced by transfer of nuclei from transfected fetal fibroblasts." *Science* 278, 1997, 2130–2133.

—. "Human factor IX transgenic sheep produced by transfer of nuclei from transfected fetal fibroblasts." *Science* 278, 1997a, 2130–2133.

—,"Production of gene-targeted sheep by nuclear transfer from cultured somatic cells." *Nature* 405, 1997b, 1066–1069.

Signer, E. N., Y. E. Dubrova, A. J. Jeffreys, C. Wilde, L. M. Finch, M. Wells, and M. Peaker. "DNA fingerprinting Dolly." *Nature* 394, 1998, 329–330.

Silva, J., and A. Smith. "Capturing pluripotency." *Cell* 132, 2008, 532–6.

Spemann, H. "Die Entwicklung seitlicher und dorso-ventraler Keimhälften bei verzögerter Kernversorgung." *Zeit. für wiss Zool* 132, 1928, 105–134.

—, *Experimentelle Beiträge zu einer Theorie der Entwicklung.* Berlin: Springer, 1936. (English ed., *Embryonic development and induction*, 1938.)

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. "Induction of pluripotent stem cells from adult human fibroblasts by defined factors." *Cell* 131, 2007, 861–872.

Takahashi, K. and Yamanaka, S. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." *Cell* 126, 2006, 663-76.

Thomas, J. M. "Unpredictability and chance in scientific progress." *Progress in Informatics* 4, 2007, 1–4.

Thomson, J. A., J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. "Embryonic stem cell lines derived from human blastocysts." *Science* 282, 1998, 1145–1147.

Wakayama, T., A. C. Perry, M. Zuccotti, K. R. Johnson, and R. Yanagimachi. "Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei." *Nature* 394, 1998, 369–374.

Weismann, A. *Das Keimplasma. Eine Theorie der Vererbung.* Jena: Gustav Fischer, 1892.

Willadsen, S. M. "Nuclear transplantation in sheep embryos." *Nature* 320, 1986, 63–65.

Wilmut, I., A. E. Schnieke, J. Mcwhir, A. J. Kind, and K. H. Campbell. "Viable offspring derived from fetal and adult mammalian cells." *Nature* 385, 1997, 810–813.

Woods, G. L., K. L. White, D. K Vanderwall, G. P. Li, K. I. Aston, T. D. Bunch, L. N. Meerdo, and B. J. Pate. "A mule cloned from fetal cells by nuclear transfer." *Science* 301, 2003, 1063.

Zhou, Q. J. Brown, A. Kanarek, J. Rajagopal, and D.A. Melton. "In vivo reprogramming of adult pancreatic exocrine cells to beta-cells." *Nature*, Aug 27, 2008. [Epub ahead of print.]

towards an understanding of cancer

JOAN MASSAGUÉ

The era in which science could conquer cancer has arrived, and the old reputation of cancer as an incurable disease is beginning to fade away. Each year new advances in medical research and clinical care diminish the old myth that cancer is a disease too complicated to comprehend and difficult to cure. The road to understanding and controlling of cancer remains arduous still, but recent progress provides reasons for cautious optimism.

Cancer facts and myths

Cancer is a class of diseases in which cells multiply out of control, invade surrounding tissues, and spread to distant organs in a process called metastasis. Invasion and metastasis are key features that distinguish malignant tumors—cancer proper—from benign growths. Cancer can emerge in essentially any organ of the body, and at any age. In developed countries, cancer is responsible for about one quarter of all deaths, and is beginning to surpass cardiovascular disease as the leading cause of death (American Cancer Society 2008; Jemal et al. 2005). Yet, in spite of these sobering realities, the notion that cancer is an incurable disease should be viewed as an obsolete myth. Most cancers can be treated, many can be successfully managed, and some can be completely cured. Cure rates for some cancers approach 95% of cases, a better success rate than that of some infectious diseases and metabolic disorders.

Fundamentally, cancer is a genetic problem. It emerges from mutations and other pathological changes in the genome of a cell, leading this cell and its descendants to misbehave (Vogelstein and Kinzler 2004). These alterations may be inherited at conception, affecting every cell of the body, but are more commonly acquired by accident in a small number of cells in one or another tissue. In most types of cancer, the transformation of a normal cell into a cancerous one requires multiple mutations that collectively disable key mechanisms for cellular self-control (Figure 1). This accumulation of mutations may take decades, which is one reason that cancer incidence increases with age.

Cancer is also a problem of cell biology. The genetic alterations that give rise to cancer act by disrupting the normal life cycle and social behavior of cells (Gupta and Massagué 2006; Hanahan and Weinberg 2000). Genes whose normal function is to promote cell movement and division may become cancer genes if they suffer alterations that increase these activities (Figure 2). On the other hand, genes whose normal function is to limit cell division, retain cells in place, promote cell differentiation, or eliminate spent and defective cells, lead to cancer when they suffer inactivation. The identification of cancer genes and the cellular functions that they control are at the forefront of contemporary research and anti-cancer drug development.

The identification of cancer genes and their biological functions during the last quarter of the 20th century is leading to better ways to prevent and treat cancer. Improved methods for assessment of cancer risk and more effective cancer prevention campaigns are decreasing cancer incidence and mortality in certain types of cancer. Less invasive surgical procedures, more refined radiotherapy methods, and more sophisticated use of chemotherapeutic drugs are contributing to the growing success of conventional cancer treatments. Moreover, a better understanding of biology and genetics of cancer is allowing the development of better drugs that target cancer cells while sparing healthy ones. And although these new drugs have begun to arrive at a trickle, this trickle is poised to become a flood. The achievement of these goals could be one of the principal scientific feats of the first half of the twenty-first century.

The growing incidence of cancer

Cancer is not a new disease. The Egyptians were surgically treating breast cancer circa 1600 BC (Karpozilos and Pavlidis 2004). By around 400 BC Hippocrates understood the difference between benign and malignant tumors, calling the latter "carcinoma" from the word "carcinos," crab in Greek, referring to the shape that he saw in advanced malignant tumors, and "-oma" meaning swelling. But while cancer is not a new disease, its incidence is on the rise. Current estimates place the worldwide mortality from cancer at nearly 8 million people annually, or about 13% of all deaths (World Health Organization 2008). The World Health Organization forecasts that by 2020 the annual global death toll will rise to about 11.5 million.

Tumors result from the accumulation of multiple mutations in the affected cells. The multiple genetic changes that result in cancer can take many years to accumulate (Vogelstein and Kinzler 2004). This is why cancer in young children and adolescents is relatively rare, and why the risk of developing cancer increases with age. In developed countries an increase in life expectancy and in population median age over the past decades have contributed to an overall increase in cancer incidence. With progress in controlling infectious diseases that are

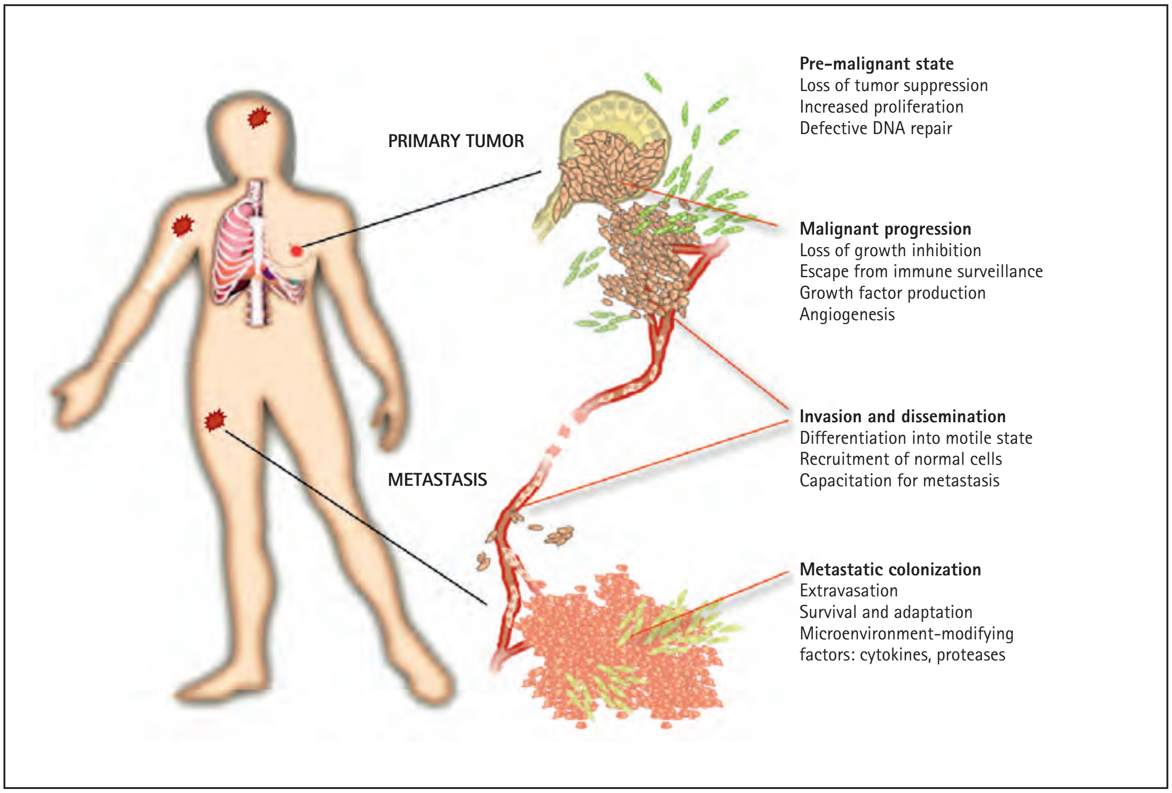

PRIMARY TUMOR
METASTASIS
Pre-malignant state
Loss of tumor suppression
Increased proliferation
Defective DNA repair
Malignant progression
Loss of growth inhibition
Escape from immune surveillance
Growth factor production
Angiogenesis
Invasion and dissemination
Differentiation into motile state
Recruitment of normal cells
Capacitation for metastasis
Metastatic colonization
Extravasation
Survival and adaptation
Microenvironment-modifying
factors: cytokines, proteases

Figure 1. Phases of a solid tumor. Solid tumors such as carcinomas of the lung, colon, breast, or prostate start in epithelial cells that line the surface of the bronchia, the intestinal mucosa, or the alveoli of fluid secretion in the breast and prostate. Mutations that increase the ability of these cells to proliferate generate small pre-malignant tissue masses. These pre-cancerous lesions may progress into malignant tumors by the acquisition of additional mutations that provide freedom from growth-inhibitory controls, protection from destruction by the immune system, capacity to invade the surrounding tissue, and the ability to attract blood capillaries ("angiogenesis"). A further conversion of the malignant tumor leads to the formation of highly motile and invasive cancer cells, and the recruitment of normal cells that act as helpers in tumor dissemination. These changes pave the way for the escape of cancer cells through the lymphatic system and the blood circulation to all parts of the body. Some disseminated cancer cells may have the ability to step out of the circulation ("extravasation") by crossing the blood capillary walls. After they enter distant organs such as the bone marrow, the lungs, the liver, or the brain, cancer cells are able to survive, adapt, and eventually overtake these new environments, leading to the formation of lethal metastases.

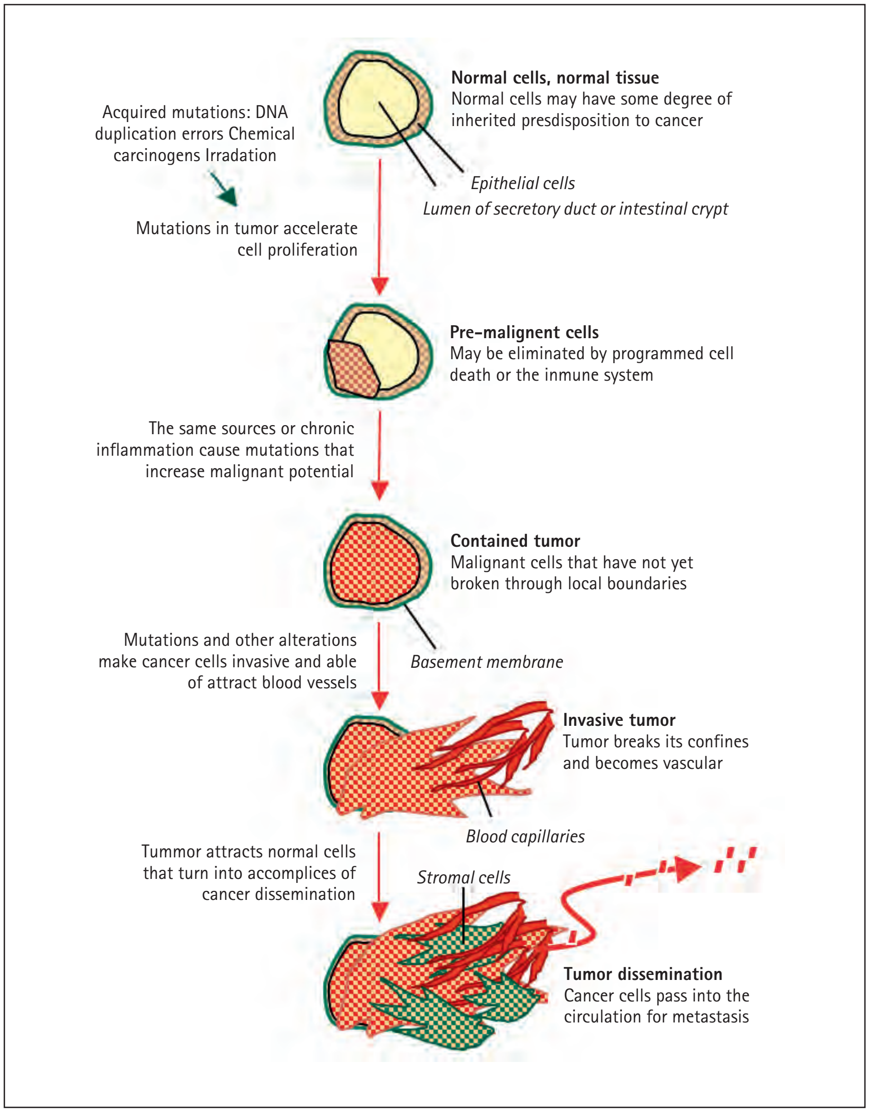

Normal cells, normal tissue
Normal cells may have some degree of inherited presdisposition to cancer
Acquired mutations: DNA duplication errors Chemical carcinogens Irradation
Epithelial cells
Lumen of secretory duct or intestinal crypt
Mutations in tumor accelerate cell proliferation
Pre-malignant cells
May be eliminated by programmed cell death or the inmune system
The same sources or chronic inflammation cause mutations that increase malignant potential
Contained tumor
Malignant cells that have not yet broken through local boundaries
Mutations and other alterations make cancer cells invasive and able of attract blood vessels
Basement membrane
Invasive tumor
Tumor breaks its confines and becomes vascular
Blood capillaries
Tummor attracts normal cells that turn into accomplices of cancer dissemination
Stromal cells
Tumor dissemination
Cancer cells pass into the circulation for metastasis

Figure 2. Sources of cancer mutations. The schematic represents the section of a secretory duct or an intestinal crypt, with a layer of epithelial cells surrounded by a basement membrane lining the cavity. The genetic inheritance of every individual contains a certain level of predisposition—high or low—for different types of cancer. Cancer-predisposing genetic variations that carry a small risk of developing a certain type of cancer are probably common in the human population, and most of these alterations remain to be discovered. Cancer-predisposing inherited mutations that carry a high risk of developing cancer (e.g. BRCA1 and BRCA2 mutations in breast and ovarian cancer, RB mutations in retinoblastoma, and APC mutations in colorectal carcinoma) are rare in the human population. These intrinsic cancer predispositions are present in all cells of the body. However, the initiation of tumor formation requires the acquisition of more mutations in all cases. The sources of cancer mutations may be internal, such as unrepaired errors in DNA replication that normal dividing cells make on their own, or external, such as chemical carcinogens in tobacco smoke or UV radiation in sun exposure. These acquired mutations accelerate cell proliferation and lead to the formation of pre-malignant lesions, such as intestinal polyps or breast tissue hyperplasias. Most of these lesions do not progress any further and are eliminated by cell self-destruction or by the killing action of the immune system. However, some pre-malignant lesions may progress into a contained carcinoma, or "carcinoma in situ" by the accumulation of additional mutations from external sources or from the genomic instability of the pre-cancerous cells. This stage is also facilitated by chronic inflammation syndromes that are triggered by a defective immune system (e.g. ulcerative colitis), an external irritant (e.g. tobacco smoke in the lungs), or an infectious agent (e.g., hepatitis virus in the liver, Helicobacter pilory in the stomach). A tumor becomes an invasive carcinoma when it breaks through the surrounding basement membrane and attracts blood capillary to bring in oxygen and nutrients. Epigenomic alterations in cancer cells, and stress in the surrounding tissue cause the release of factors that recruit normal cells, which end up being turned into helpers in tumor progression. At this stage the cancer cells have access to the circulation and can disseminate throughout the body. Some disseminated cancer cells may go on to reproduce the tumor in distant organs, giving rise to metastasis.

currently ravaging underdeveloped countries, we may expect similar increases in cancer incidence in those countries as well. Other contributing factors include the detection of more early-stage tumors in routine medical examinations, various factors in diet and lifestyle, and the negative impact of cigarette smoking.

The overall incidence of cancer, and the incidence of particular types of cancer vary between different countries (Danaei et al. 2005; World Heatlh Organization 2008). For example, the most common types of cancer by mortality in the United States and Spain are similar, but with one notable exception: lung cancer mortality in females. Lung cancer ranks as the top cause of cancer deaths for males in both countries and for females in the United States. However, until recently at least, lung cancer is in third for position females in Spain (Table 1). This difference is attributed to the delay in tobacco consumption by women in Spain compared to men in both countries and women in the Unites States. Epidemiologic studies demonstrate a tight correlation between tobacco consumption and lung cancer, with a 20-year lag between the two.

Cancer and cancers

"Cancer" includes hundreds of different diseases. Primary tumors arising in different organs or tissues—for, example, breast cancer, lung cancer or leukemia—are obviously different in their appearance, evolution, response to treatment, and mortality. However, tumors arising in the same organ can be further classified into different subtypes that are very distinct from each other. There are at least five different subtypes of breast cancer, and

MALES		
	United States	Spain
	lung (31%)	lung (28%)
	prostate (10%)	colorectal (12%)
	colorectal (8%)	prostate (8%)
	pancreas (6%)	pancreas (6%)
	leukemia (4%)	bladder (6%)
	liver/bile (4%)	stomach (6%)
FEMALES		
	United States	Spain
	lung (26%)	breast (19%)
	mama (15%)	colorectal (15%)
	colorectal (9%)	lung (8%)
	pancreas (6%)	pancreas (6%)
	ovary (6%)	ovary (6%)
	leukemia (3%)	stomach (6%)

Table 1. Cancer incidence in adults in the United States (American Cancer Society 2008) and Spain (Centro Nacional de Epidemiología de España). Numbers in parenthesis represent the percent of all cancer deaths that are due to this particular type of cancer.

even these could be subdivided into different variants. The same could be said of cancer in other organs. With these differences come distinct treatment indications.

Tumors are also classified according to the type of cell that they derive from. Carcinomas are malignant tumors derived from epithelial cells, such as those that form the upper layer of the skin and the digestive mucosa, or the internal structure of organs like the breast, prostate, liver, and pancreas. Sarcomas are derived from cells of connective tissues such as bone, cartilage, and muscle. Lymphomas and leukemias are derived from blood-forming cells, melanomas from melanocytes (skin pigmented cells), and glioblastoma, neuroblastoma, and medulloblastoma from immature neural tissue cells. Carcinomas are the most common type of cancer in adults whereas in the young neuroblastoma, medulloblastoma, and leukemia are the common types.

A third set of parameters in the classification of tumors is based on the extent of spread of the tumor, which is called "stage" of the disease, and the histological appearance under the microscope, which is called the "grade." However, tumors of the same origin, kind, grade, and stage may progress and respond to therapy very differently in different patients. This reality has a major impact on our view of cancer as a disease that we still know too little about. Fortunately, this is about to radically change. The advent of molecular genetics technologies is allowing a better classification of cancers based on their specific origin, molecular alterations, risk of spread to other organs, and treatment of choice.

Causes of cancer

Cancer develops as a consequence of mutations and other abnormalities that alter the genes that control cell behavior (Hanahan and Weinberg 2000; Vogelstein and Kinzler 2004). These mutations may be acquired through the action of external factors—such chemical carcinogens, radiation and infectious agents—or internal errors in DNA replication and repair in small groups of cells throughout life (Figure 2). Cancer mutations may also be inherited, in which case they are in all cells from birth. Current research on the genetic basis of cancer is focusing on the processes that cause these genetic alterations, the types of genes that are affected, and the biological consequences of these effects.

Common examples of chemical carcinogens include tobacco smoking, which causes lung cancer and bladder cancer, and exposure to asbestos fibers, which causes mesothelioma (Danaei et al. 2005). Ultraviolet radiation from the sun can lead to melanoma and other skin cancer. Tobacco smoke carcinogens and radiation are thought to promote the formation of tumors by acting as direct mutagens. Tobacco and asbestos may also cause chronic inflammation that secondarily favors tumor development. Viral infections are the second most important external cause of cancer after tobacco usage (zur Hausen 1999). Viruses associated with human cancers include papilloma virus in cervical cancer, hepatitis B and C viruses in liver cancer, HIV in Kaposi's sarcoma, and Epstein-Barr virus in B-cell lymphomas (Boshoff and Weiss 2002; Parato et al. 2005; Roden et al. 2006; Woodman et al. 2007; Young and Rickinson 2004). Viral infections promote tumor formation by incorporation of the virus genome into the DNA of the host cell, which may increase the activity of neighboring genes that stimulate uncontrolled cell division. Viral infections may also promote tumor growth by causing chronic inflammation and stimulating cell turnover in the host tissues. Liver tissue degeneration, or cirrhosis, caused by alcoholism, is associated with the development of liver cancer. The combination of cirrhosis and viral hepatitis constitutes the highest risk of developing liver cancer, which is one of the most common and deadly cancers worldwide. Certain bacterial infections also favor the development of cancer. The clearest example is gastric cancers tied to chronic inflammation of the stomach mucosa by Helicobacter pylori infection (Cheung et al. 2007; Wang et al. 2007).

Certain types of cancer have a strong hereditary component (Vogelstein and Kinzler 2004). Inherited mutations in the genes BRCA1 and BRCA2 create a high risk of developing breast cancer and ovarian cancer (Walsh and King 2007; Wang 2007; Welcsh and King 2001). Interestingly, BRCA mutations are rare in sporadic breast cancer. In contrast, p53, which is commonly mutated in sporadic cancers, is also the gene affected in the hereditary syndrome of Li-Fraumeni, which includes a predisposition for sarcomas, breast cancer, and brain tumors (Vousden and Lane 2007). Retinoblastoma in children is due to a hereditary mutation in the retinoblastoma (RB) gene, a gene which is also mutated in many sporadic cancers (Classon and Harlow 2002). An inherited mutation of the APC gene gives rise to thousands of polyps in the colon, which leads to early onset of colon carcinoma (Fodde et al. 2001). Another hereditary form of cancer predisposition is caused by mutations in one of several genes (MLH1, MSH2, MSH6, PMS1, PMS2) devoted to the repair of DNA replication errors. This condition, called HNPCC (hereditary non-polyposis colon cancer), include familial cases of colon cancer without a prevalence of colon polyps, uterine cancer, gastric cancer, and ovarian cancer (de la Chapelle 2004). Inherited mutations in the VHL1 gene predispose to kidney cancer (Kaelin 2005).

The inherited mutations that have a strong effect on tumor development are rare in the human population, and account for only a small fraction of cancer. For example, inherited BRCA mutations account for less than 2% of breast cancer in the general population (Welcsh and King 2001). At the other end of the spectrum, certain genetic variations may have a very weak individual impact on the risk of developing cancer but may be highly prevalent in the human population. In certain combinations, these genetic traits could synergize to create a significant risk of cancer. The current view is that cancer arises from complex interactions between external carcinogens and an individual's genome. The identification of these weakly predisposing genetic determinants currently is a topic of intense investigation.

Normal cells and cancer cells

Cells are the basic unit of life. In isolation their basic activities are to resist the environment, incorporate nutrients, faithfully replicate their genome, and divide. However, the cells that form the tissues of a complex organism can no longer perform these tasks autonomously. Single cells evolved to form organized colonies hundreds of millions of years ago because this communal form of life proved advantageous in facing up to harsh environments. But this communal life also meant giving up certain degrees of freedom. For example, it was no longer suitable for a cell in the community to divide or move just as it wished. In our highly organized tissues, such decisions are subject to a complex network of cell-to-cell molecular signals. This form of dialog between cells has been developing and enriching itself over millions of years, and a good portion of our genome is entirely devoted to this task.

Cells communicate with each other by secreting molecules, generally in the form of small proteins known as hormones, growth factors, cytokines, or chemokines. These factors contact receptor proteins on the surface of target cells to activate "pathways," which are sequences of biochemical reactions between signal-transducing proteins inside the cell (Bierie and Moses 2006; Bild et al. 2006; Christofori 2006; Ciardiello and Tortora 2008; Classon and Harlow 2002; Ferrara 2002; Fodde et al. 2001; Hanahan and Weinberg 2000; Karin 2006; Malumbres and Barbacid 2007; Massagué 2004, 2008; Olsson et al. 2006; Pouyssegur et al. 2006; Sweet-Cordero et al. 2005; Vousden and Lane 2007). The end result of this process are positive or negative changes in the ability of the cell to move, metabolize, grow, divide, differentiate, or die. Other proteins inside the cell sense the presence of errors and alterations in the DNA, and prompt their repair or else provoke the death of the cell. Loss of these important signaling and self-controlling functions results in cancer. Cancer cells disobey essential rules of life in community, increasingly misuse proliferative stimuli, and ignore the rules of moderation. Their interaction with their neighbors becomes openly antisocial. They avert the policing action of the immune system. Eventually, they break through the physical barriers that encapsulate the tumor, setting out on the march that will spread cancer cells throughout the body and metastasis.

The mutations that cause cancer precisely affect the genes whose products exert these critical control functions. The progressive accumulation of mutations turn normal cells into pre-malignant and eventually into fully malignant cells (Figure 2). These changes can be observed under the microscope. A malignant process may start with the presence of an excessive number of normal-looking cells, called a hyperplasia, and more specifically with a disordered accumulation of such cells, or dysplasia. As the cells cease to look normal, the lesion is considered a carcinoma in situ, in which the abnormal cells are still confined to the normal limits of the tissue boundaries. When carcinoma cells invade the surrounding tissue by breaking through their underlying barrier (called the "basement membrane"), the lesion is call an invasive carcinoma. Each of these steps is accompanied by, and the result of, the progressively accumulating mutations that lead to cancer.

The specific functions that must be perturbed in order to generate cancer cells include a gain of self-sufficiency in growth-promoting signals; a loss of sensitivity to growth-inhibitory signals; a loss in the ability to undergo cell death (loss of apoptosis); a gain in the ability to perpetually replicate the DNA; and, a gain in the ability to evade surveillance by the immune system (Hanahan and Weinberg 2000). These changes are required for all types of cancer cells, including blood cell cancers such as leukemias. To form a tumor, cancer cells from solid tissues additionally require a gain in the ability to resist hypoxia through the induction of new capillaries that will feed the tumor (angiogenesis); and gain in the ability to detach and invade surrounding tissue (Figure 2). To spread the tumor to distant sites, the cancer cells must also gain the ability to pass into the circulation, enter distant tissues, and adapt to the microenvironment of those tissues eventually overtaking them.

Cancer genes

Cancer genes are divided into two general classes. Genes whose excessive activity contributes to cancer are called "oncogenes" (Hanahan and Weinberg 2000). The genes encode growth factor receptors such as EGFR and HER2, transducers of growth factor signals such as RAS, RAF, and PI3K, cell survival factors such as BCL2, and others. The mutations affecting these genes are activating or "gain-of-function" mutations. Genes whose normal activity prevents the emergence of cancer are called "tumor suppressor" genes. The mutations that affect these genes in cancer are inactivating mutations. Tumor suppressors include sensors of DNA damage such as p53, genes that fix DNA damage such as BRCA1 and BRCA2, inhibitors of the cell division cycle such as RB, receptors and transducers of growth inhibitory signals such as TGFBR and SMAD4, and suppressors of growth signals such as PTEN.

The mutations affecting these genes may be point mutations that alter one single nucleotide in the gene, and a single amino acid in the gene product. Point mutations may either increase or decrease the activity of the gene product, and so point mutations are a cause of oncogene activation as well as tumor suppressor gene inactivation. Small deletions or insertions may similarly cause either oncogene activation or tumor suppressor inactivation. Large-scale mutations involve the deletion or gain of a portion of a chromosome, resulting in the gain of multiple copies of one or more oncogenes, or a loss of tumor suppressor genes. Translocations occur when two separate chromosomal regions become abnormally fused, often at a characteristic location. A well-known example of this is the Philadelphia chromosome, or translocation of chromosomes 9 and 22, which occurs in chronic myelogenous leukemia, and results in production of the BCR-ABL fusion protein (Melo and Barnes 2007). This causes oncogenic activation of the ABL gene. Some oncogenic mutations affect not the protein-coding region of an oncogene but the regulatory or "promoter" region that controls the production of the gene product. Insertion of viral genome near the promoter region may also lead to hyperactivation of an oncogene.

In addition to the various types of mutations that alter the chemical structure of a normal gene turning it into a cancer gene, there is a growing recognition of the impact of epigenomic modifications. These are chemical modifications of the DNA and the proteins that envelop it (Blasco 2007; Esteller 2007). These modifications are known as "epigenetic" changes and can either render a gene silent or make it competent for activation. Epigenetic deregulation can contribute to cancer by failing to keep oncogenes silent, for example, through DNA methylation. Loss of methylation can induce the aberrant expression of oncogenes. Methylation or acetylation of histone proteins that package the chromosomal DNA can also suffer alterations that contribute to cancer. The experimental anti-cancer drug vorinostat acts by restoring histone acetylation and is currently undergoing clinical evaluation.

Ecology of the tumor microenvironment

Each tissue has a characteristic structure, boundaries, vascular supply, and extracellular milieu of hormones, nutrients, and metabolites. Cancer cells that alter this become exposed to environmental stresses including lack of oxygen (hypoxia) and nutrients, acidity, oxidative stress, and inflammatory responses. These stressful conditions select for cells that survive such pressures and become a dominant population in the growing tumor. This phenomenon is known as "clonal selection" (Nowell 1976). The resulting clones of cells are not merely survivors; they are highly effective profiteers that take advantage of the tumor microenvironment.

Tumors are more than just a conglomerate of cancer cells. Tumors also include normal cells that become attracted to, and engulfed by the growing tumor, and may become accomplices in its development (Joyce 2005; Mueller and Fusenig 2004). The collection of non-cancerous cell types that are present in a tumor is called the tumor "stroma," and their importance in cancer is being increasingly recognized. Endothelial cells recruited into the tumor form new blood capillaries that bring nutrients and oxygen into the tumor mass. Macrophages and other immune and inflammatory cells congregate in the tumor in an attempt to respond to the tissue distress. Tumor-associated macrophages, produce growth factors and ECM-degrading enzymes that stimulate the growth and invasion of the cancer cells (Joyce 2005; Lewis and Pollard 2006). Stress-response cells are also recruited into the tumor from the circulation. Several types of blood-derived cells are attracted by signals that emanate from the tumor and proliferate in response to these signals. Stroma-derived factors may in turn stimulate cancer cells to release signals that enhance their ability for form metastases. For example, the stroma-derived cytokine transforming growth factor β (TGFβ) can induce breast cancer cells to release angiopoietin-like 4, which enhances the ability of these cells to seed the lungs after they escape from the primary tumor (Padua et al. 2008). Thus, the stroma of a tumor can provide cancer cells with metastatic advantages.

Metastasis: the deadly spread of tumors

Aggressive tumors may release millions of cancer cells into the circulation before the tumor is detected and surgically removed. Metastasis is the process by which these disseminated cancer cells take over distant organs and ultimately cause organ dysfunction and death (Figure 1). Metastases may be detected at the time of initial diagnosis of cancer or months to years later, when the disease recurs. The disseminated cancer cells may remain dormant in distant organs for a long period, until unknown conditions lead to their reactivation and formation of aggressively growing metastasis.

The administration of chemotherapy to cancer patients after surgical removal of a primary tumor is intended to eliminate all residual tumor cells and avert the eventual emergence of metastasis. Yet, the failure of current therapeutics to control or cure metastasis is responsible for 90% of cancer deaths. If it were not for metastasis, cancer would represent only a small fraction of the problem that it is today. Understanding the many molecular players and processes involved in metastasis may eventually lead to more effective, targeted approaches to treat cancer.

Recent advances in technologies to visualize and track the metastasis process have helped delineate multiple events that lead cancer cells in a primary tumor to reach and colonize a distant site (Fidler 2003; Gupta and Massagué 2006; Weinberg 2007) (Figure 2). Carcinoma cells must first pass through the basement membrane of the tissue compartment in which the tumor occurs. The basement membrane separates the epithelial cell layers in which carcinomas originate, from the subjacent tissue. Basement membranes also envelop the blood vessels. In order to surpass a basement membrane and spread through the surrounding tissue, cancer cells must acquire the ability to detach from their place of origin, adopt a migratory behavior, and release proteolytic enzymes that degrade the protein scaffold of the basement membrane and extracellular matrix.

Once cancer cells form a small tumor mass and create hypoxic conditions, they respond to hypoxia with the secretion of cytokines that stimulate the formation of new capillaries that bring in oxygen and nutrients for tumor growth. As a result of tumor-derived permeability factors, these newly formed capillaries are leaky, providing a route for the escape of the cancer cells into the blood circulation. Lymphatic vessels that drain fluid from the tumor and surrounding tissue provide another route for cancer cell dissemination. Lymph nodes frequently trap traveling tumor cells and document their spread, which is why lymph node status is an important prognostic indicator at initial diagnosis. However, the dissemination of cancer cells to distant organs such as the lungs, brain, bones, and liver occurs mainly through the blood circulation. In the bloodstream, cancer cells associate with each other and blood cells to form emboli that may help withstand mechanical stresses and evade surveillance by the immune system.

Once cancer cells lodge in capillaries at distant organs they must pass through the capillary walls in order to gain access to the organ parenchyma (Figure 3). Extravasation, as this process is known, depends on the ability of the cancer cells to disrupt the tight contacts between endothelial cells of the capillary wall and the enveloping basement membrane. The microenvironment of the infiltrated organ is largely not permissive for the extravasating cancer cells, many of which die. Those that survive form micrometastases that must adapt to the new environment and co-opt its resident cells in order to re-initiate tumor growth and form aggressive metastatic colonies. This process can take months, years, and even decades. Only a small fraction of the cancer cells released by a tumor are capable of fulfilling all these requirements, but the few that do are sufficient for the establishment of lethal metastases.

The ingredients for metastasis

Genetic heterogeneity

Metastasis has many features of a Darwinian evolution process in which selective pressures for the emergence of the fittest individual cells from a tumor cell population. Evolution requires the presence of genetic heterogeneity in a population from which fit individuals can be selected to match particular environmental pressures. In tumors, such heterogeneity is amply provided by the characteristic genomic instability of cancer cells, and it increases the probability that some cells in a tumor will achieve metastatic competence. Thus, the different steps of metastasis do not necessarily represent the acquisition of individual specialized mutations but rather represent the random accumulation of traits that provide the necessary advantage for adaptation to a different organ microenvironent.

Genomic instability and heterogeneity of cancer cells are apparent in the chromosomal gains, losses, and rearrangements found in tumors. DNA integrity can be compromised by aberrant cell cycle progression, telomeric crisis, inactivation of DNA repair genes, and altered epigenetic control mechanisms. For example, one half of all human cancers suffer loss of the tumor suppressor p53, an internal protein that responds to DNA damage by causing elimination of the damaged cell. The loss of p53 allows cancer cells with DNA alterations to survive and accumulate additional mutations (Halazonetis et al. 2008). Inherited mutations in certain DNA repair genes are associated with a higher risk of developing cancer, for example, in the hereditary nonpolyposis colorectal cancer syndrome (HNPCC) (Rustgi 2007), and in familial breast cancer syndromes cause by mutations in BRCA1 or BRCA2 (Martin et al. 2008).

Cancer stem cells

Metastasis requires a robust ability of cancer cells to re-initiate tumor growth after they penetrate a distant tissue in small numbers. Not all the cancer cells in a tumor are capable of dividing indefinitely and so not all cancer cells have the capacity to re-initiate a tumor after they arrive in a metastasis site. Many in fact loose tumorigenic power by falling into a more differentiated state. However, by one mechanism or another a subset of cancer cells in a tumor have the capacity of acting as tumor-propagating cells (Clarke and Fuller 2006). This capacity defines the cells that have it as "cancer stem cells" or "tumor-propagating cells." These functions would support the maintenance of primary tumors, and would be essential for the establishment of metastatic colonies. Additional properties of these cells may include resistance to chemotherapeutic drugs or sensitivity to different drugs compared to the sensitivity of the bulk cancer cell population in the tumor. However, tumor propagating cells need not be a minority of the cancer cells in a tumor; in some types of tumors a majority of the cells may be competent to re-initiate tumor growth if they fulfill the other requirements for metastasis. The extent to which different tumor types may be initiated and sustained by cancer cells that meet these criteria is a subject of intense investigation and one that is likely to meet with different answers in different tumor types.

Metastatic dissemination

In order to become disseminated, cancer cells must be able to break their ties with the cohesive structure of the tissue of origin. Adhesion of cancer cells to each other is reduced by the loss of cell-cell anchoring proteins such as E-cadherin. Loss of E-cadherin in tumors can occur through various mechanisms, including silencing of the gene that encodes E-cadherin, mutations in this gene that result in the production of inactive E-cadherin, or repression of E-cadherin activity by growth factor receptors (Perl et al. 1998; Thiery 2002). Loss of E-cadherin activity also occurs as part of the transformation of cancer cells from an epithelial state into a more motile cell state, a change known as "epithelial-mesenchymal transition" or EMT (Cano et al. 2000; Yang et al. 2004). Normal cells are also kept in place by the extracellular matrix (ECM), a scaffold formed by collagen and other fiber-forming proteins to which cells attach by means of receptors called integrins. These ECM contacts can retain cells in place but can also stimulate cells to form extensions for migratory movement. Various proteins involved in these types of cell shape changes, such as RhoC and NEDD9, have been implicated in cancer cell invasion leading to metastasis (Clark et al. 2000; Kim et al. 2006).

Cancer cells may disseminate from a tumor very early on in a tumor formation process. Cancer cells have been detected in the bone marrow of breast cancer patients with small tumors (Klein et al. 2002; Schmidt-Kittler et al. 2003). This does not necessarily mean that the earliest departing cells are the ones that progress into overt metastasis, but it does indicate that dissemination is not an exclusive property of large, advanced tumors.

Once the disseminated cancer cells reach distant organs, they may remain dormant or die. Dormancy may last years, even decades before disseminated cancer cells burst into aggressive outgrowth, as in the case of breast cancer. Disseminated cancer cells found in the bone marrow of women or transgenic mice with early-stage breast cancer can become activated by transplantation into the bone marrow of mice to cause lethal tumors (Husemann et al. 2008).

Dissemination may also occur from metastatic tumors, which in turn seed new metastases. It is possible that circulating tumor cells can re-infiltrate the same tumors from which they departed. According to this hypothesis, tumors may continuously enrich themselves with their most aggressive progeny, providing a mechanism that couples metastatic ability with tumor growth (Norton and Massagué 2006). This would provide an explanation for the longstanding correlation between metastasis and tumor size (Minn et al. 2007). The timing and mechanisms of cancer cell dissemination are topics of great interest in contemporary cancer research.

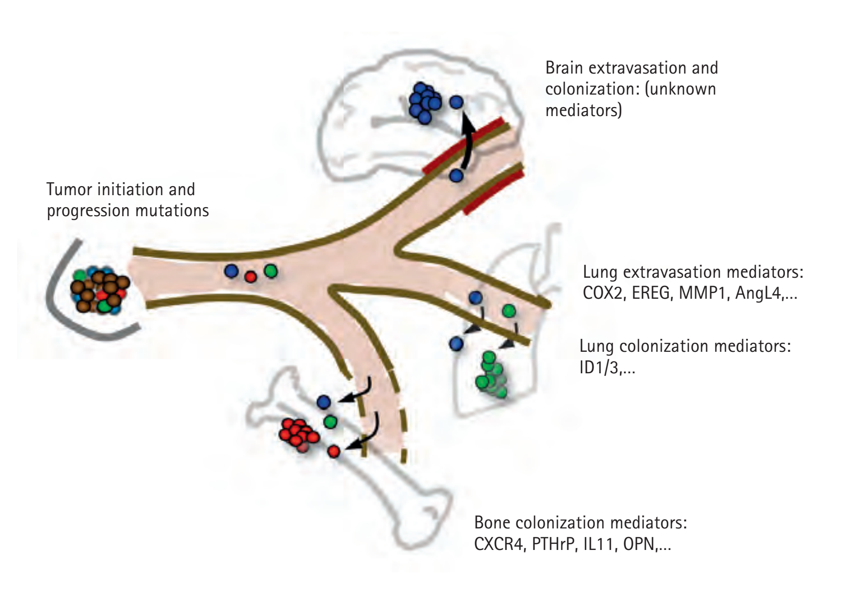

Brain extravasation and colonization: (unknown mediators)
Tumor initiation and progression mutations
Lung extravasation mediators: COX2, EREG, MMP1, AngL4,...
Lung colonization mediators: ID1/3,...
Bone colonization mediators: CXCR4, PTHrP, IL11, OPN,...

Figure 3. Mediators of distant metastasis in breast cancer. Metastasis has been most extensively studied in breast cancer because of the common nature of this disease, the availability of clinical material, and the characteristic set of organs that are affected in this process. Breast tumor can release cancer cells into the circulation as soon as they become locally invasive through the acquisition of tumor initiating and tumor progression mutations (see Figure 2). The disseminated cells that survive the physical stresses of the circulation require additional functions for entry into distant tissues. The passage through the blood capillary walls in those tissues, or "extravasation," is relatively permissive in the bone marrow (and in the liver, not illustrated), because the capillaries in these tissues have natural windows for the constant entry and exit of blood cells. However, on entering the bone marrow, cancer cells must have the ability to survive and productively interact with this microenvironment. The fact that metastasis from breast cancer may take years and even decades to emerge suggests that the disseminated cancer cells originally arrived in this organ unprepared, and they had to slowly evolve the necessary abilities to expand as aggressive colonies. The genes that breast cancer cells misuse in order to survive in the bone marrow include the chemotaxis and survival receptor CXCR4, the osteoclast stimulating factors parathyroid hormone-relaped protein (PTHrP), interleukin-11 (IL11) and osteopontin (OPN), and other genes. In contrast to the bone marrow capillaries, the capillaries in other organs such as the lungs and especially the brain have tight walls that are very restrictive to the passage of circulating cells. For this reason, cancer cells must carry certain activated genes in order to enter these organs. Mediators of breast cancer cell entry into the lungs include the EGFR ligand epiregulin (EREG), the prostaglandin-synthesizing enzyme cycloxygenase-2 (COX2), the collagen-degrading enzyme matrix metalloproteinase-1 (MMP1), and the endothelium disrupting factor angiopoietin-like 4 (AngL4). It is suspected that some of these genes are also engaged by breast cancer cells to enter the brain. The genes that mediate colonization of the lungs and the brain are largely unknown, and the subject of active investigation. ID1 and ID3 have been recently identified as mediators of tumor re-initiation by breast cancer cells entering the lungs. Thus, expression of ID1/3 is a property of tumor-propagating cells, also known as "cancer stem cells".

Different "seeds" for different "soils"

The bones, lungs, liver, and brain are the most frequent sites of metastasis (Figure 3). However, different cancers have different proclivities to spread to these organs. (Billingsley et al. 1999; Gavrilovic and Posner 2005; Hess et al. 2006; Leiter et al. 2004). The compatibility between disseminated cancer cells (the "seed") and certain distant organs (the "soil") was already noted in the nineteenth century by Stephen Paget, who promulgated the "seed" and "soil" hypothesis (Paget 1889). For example, breast cancer can spread to these four sites, with bones and lungs being the most frequently affected. Lung cancer metastasis occurs with preference in the brain, bones, and contralateral lung. In contrast, prostate cancer metastasis principally occurs in the bones, and to a limited extent in the lungs. Furthermore, although these three tumors spread to the bones, they form very different types of lesions: bone metastasis from breast cancer and lung cancer is "osteolytic," meaning that these lesions dissolve the bone matrix causing fractures. In contrast, prostate cancer metastasis is osteoblastic, leading to the generation of abnormal bone tissue that fills the marrow cavity. The predilection for a tumor in a particular organ to metastasize to another location in the same organ also varies. Tumors in one lung can easily metastasize to the other lung, whereas tumors in one breast rarely metastasize to the other breast.

Towards understanding metastasis

The progress achieved since the turn of the twenty-first century is shaping our view of metastasis based on a better understanding of its genetic, molecular, and biological bases. This knowledge is rapidly accumulating based on the identification of genes whose aberrant activity serves the purposes of the metastatic cells. Through these advances, metastasis is transforming from an obscure topic into a problem that is being rationalized and dissected, and may eventually be controlled.

An integrated model of metastasis

Early theories of metastasis proposed competing models of genetic predetermination of an entire tumor mass for metastasis, versus tumor progression giving rise to rare cells capable of metastasis (Vogelstein et al. 1988). With the sequencing of the human genome, powerful microarray technologies have been developed that allow researchers to determine the activation state of every gene in a small tissue sample. Using these techniques, it has been possible to identify patterns of gene activity, or "gene-expression signatures," that can indicate the likelihood that a particular tumor will cause metastasis. If a sample extracted from a primary tumor shows the presence of a particular pro-metastatic gene-expression profile, this would indicate that a substantial proportion of the cells in that tumor are expressing such genes and thus are competent for metastasis. This would support the predetermination theory of metastasis. However, this competence may be quite incomplete. Additional alterations have to occur before the cancer cells become fully equipped to invade and colonize a distant tissue. The acquisition of a complete set of metastatic capacities may occur frequently in a tumor population, as must be the case in tumors that rapidly metastasize to multiple organs, or it may occur slowly in a minority of predisposed cells giving rise to metastases in one or another organ years or decades after departing from the primary tumor. The latter would argue for further progression of a tumor as a necessary step for metastasis.

Recent progress in metastasis research has provided experimental and clinical evidence for both the predetermination and the progression models, leading to a model that integrates features of both. Cancer cells in a tumor with poor prognosis may contain activated genes that provide these cells with some, but not all the functions required for distant metastasis. We call these genes "metastasis progression" genes, because they directly allow the cancer cell population to become competent for metastatic behavior. Metastasis progression genes are necessary but not sufficient for metastatic outgrowth, because a majority of the cancer cells that express these genes are still incapable of forming metastatic tumors. This implies the existence of a complementary set of metastasis genes that provide additional survival and adaptation functions in a given organ. We refer to this class of genes as "metastasis virulence" genes.

Metastasis progression genes

Recent work in our laboratory has identified a set of 18 genes that breast cancer cells use to their advantage both in the primary tumor and in the lungs (Figure 3). This set, termed the "lung metastasis gene-expression signature" (LMS), includes EREG, COX-2, and MMP1, which cooperate in remodeling new blood capillaries in mammary tumors and existing lung capillaries when cancer cells expressing these genes reach the lungs. In mammary tumors the products of these genes support the assembly of leaky capillaries that facilitate the escape of cancer cells; in the lung, the same products facilitate the passage of circulating cancer cells into the parenchyma (Gupta et al. 2007). Another example is the gene that encodes ID1, which inhibits cell differentiation and stabilizes the tumor-propagating ability of cancer cells. In experimental models ID1 is important for the growth of breast tumors and for the re-initiation of the tumor growth after cancer cells reach the lungs. Thus, metastasis progression genes may couple the tissue-specific requirements of the microenvironment in a particular organ to a matching role in primary tumor progression. Breast cancer patients with LMS-positive primary tumors have a higher risk for developing lung metastases, but not metastases in bone or other sites.

Not all metastasis genes that are expressed in primary tumors provide a selective advantage in these tumors. For example, the production of transforming growth factor (TGF) in the stroma of breast primary tumors stimulates the expression of more than one-hundred genes in the breast cancer cells of the same tumor. Among these is the gene encoding the secreted factor ANGPTL4. Unlike EGFR, COX2, MMP1, or ID1, production of ANGPTL4 does not appear to provide an advantage to the cancer cells in the primary tumors—it merely reflects the presence of TGF in the tumor milieu. However, when the stimulated cancer cells reach the lung capillaries, the ANGPTL4 that these cells release causes disruption of the capillary walls and facilitates cancer cell entry into the tissue (Padua et al. 2008).

Specialized contributions of metastasis virulence genes

When cancer cells reach distant organs they are generally faced with a non-permissive microenvironment. To form a metastatic colony cancer must have an ability to resist and exploit this microenvironment. A clear example is provided by osteolytic bone metastasis from breast cancer. Circulating breast cancer cells that enter the bone marrow must find ways to survive in the unique stroma and hormonal milieu of this tissue, and ways to activate the mobilization and action of osteoclasts that mediate bone destruction. Breast cancer cells that are metastatic to bone express high levels of CXCR4. This membrane protein acts as the receptor for the cell survival factor CXCL12, which is abundantly produced in the bone marrow stroma (Wang et al. 2006). Therefore, cancer cells expressing high levels of CXCR4 obtain a specific advantage from the presence of CXCL12 in the bone marrow. In experimental models using mice, breast cancer cells that preferentially colonize the bones show not only high expression of the survival receptor CXCR4 but also an elevated production of the factors PTHrP (parathyroid hormone-related peptide), TNF- , IL-1, IL-6, and IL-11 (Kang et al. 2003). When secreted by bone metastatic cells, these factors stimulate osteoblasts to release RANKL, which activates osteoclast differentiation. Osteoclasts dissolve bone, in turn releasing growth factors such as insulin-like growth factor-I (IGF-1), which favor cancer cell survival, and TGFβ, which stimulates the cancer cells to further release PTHrP. The end result of this process is a vicious cycle of cancer cell-osteoclasts interactions that accelerated the destructive action of bone metastasis.

The ongoing search for genes and functions that mediate metastasis by other tumor types or to other organs is beginning to yield results. Prostate cancer cells secrete factors such as Wnt and bone morphogenetic proteins (BMPs) that stimulate the accumulation of osteoblasts. As a result, prostate cancer gives rise to osteoblasting (bone-forming) metastases, in contrast to the bone-destroying metastases caused by breast cancer. Compared to metastasis in bone and lung, little is known about the genes that cancer cells use to colonize the liver or the brain. However, this topic is under intense investigation and may yield progress in the near future.

Frontiers in cancer prevention, diagnosis, and treatment

Cancer prevention campaigns aiming at reducing high-risk behaviors (tobacco and alcohol abuse, sun exposure, and others) and routine screening for the detection of early-stage tumors are critical to reducing the incidence and mortality of cancer. Early diagnosis leads to therapeutic intervention before a tumor has become disseminated, curing the disease or at least extending the life of the patient. Important benefits have been obtained from screening for breast cancer with mammograms, colorectal cancer with colonoscopy, uterine cancer with cervical cytological testing, and prostate cancer with rectal exam and prostate-specific antigen (PSA) blood testing. Preventive vaccination against certain sexually transmitted strains of human papillomavirus is intended to reduce the incidence of cervical cancer. Genetic testing for certain cancer-related genetic mutations in BRCA1 and BRCA2 (which predispose to breast and ovarian cancers) and DNA repair genes (which predispose to colon cancer and other cancers) is performed in high-risk individuals with a family history of these diseases. Carriers of these mutations are subjected to close surveillance and may elect prophylactic surgery (removal of breasts, ovary, or colon) to reduce the risk of tumor development.

Recent progress is improving the classical approaches for the treatment of cancer (surgery, chemotherapy, radiation therapy) and novel approaches based on targeted therapy and immunotherapy. Surgical methods are gaining in precision and becoming less invasive. The surgical removal of a tumor that has not spread can effectively cure cancer. However, the propensity of cancer cells to invade adjacent tissue and spread to distant sites limits the effectiveness of surgery. Even small, localized tumors have metastatic potential. Therefore, surgery is very frequently complemented with other forms of therapy. Radiation therapy is based on the use of ionizing radiation (X-rays) to shrink tumors before surgery to kill locally disseminated cancer cells. Radiation can cause damage to normal tissue, therefore it can only be applied to a restricted area of the body. Radiation therapy destroys cells by causing extensive damage to their DNA. Most normal cells can recover from radiotherapy more efficiently than cancer cells, providing a window of opportunity for this intervention.

Chemotherapy is the treatment of cancer with drugs more toxic to cancer cells than they are to normal cells. Conventional anticancer drugs poison rapidly dividing cells by disrupting the duplication of DNA or the separation of newly formed chromosomes. As in the case of radiation therapy, normal cells have a higher ability to recoverfrom this damage than cancer cells. For this reason chemotherapy is often used at the maximal tolerated dose, with the consequent side effects on tissues that depend on rapid cell turnover such as the oral and gastrointestinal mucosa, the hair, skin, and nails.

One important aim of current research is to develop drugs that target cancer cells based on the specific dependency of these cells on the oncogenic mutations that they contain (Sawyers 2004) (Figure 4). Such “targeted therapies” are no different from many drugs that are available against other types of diseases. Targeted drugs against cancer aim at achieving higher therapeutic effectiveness with fewer side effects. In combination with these drugs, conventional chemotherapy may be applied at lower levels, also with fewer side effects. Targeted therapies include chemical compounds that generally act by inhibiting the enzymatic activity of oncogene products, and monoclonal antibodies that act by blocking oncogenic receptor on the surface of the cell or by antibody-mediated killing of the destruction of the target cell.

The advent of targeted therapies started in the 1990s as a direct result of the identification of critical cancer genes. The new ability to molecularly analyze tumors is revolutionizing tumor classification, prognosis, and treatment. It remains an area of intense research and high promise. Among the monoclonal antibodies, the anti-HER2 antibody trastuzumab (Herceptin™) is effective against a subtype of breast carcinomas that contain an activated HER2 oncogene (Hudis 2007; Shawver et al. 2002). The anti-CD20 antibody rituximab (Rituxan™) is used to treat B-cell lymphomas that present the CD20 antigen (Cheson and Leonard 2008), and the anti-EGFR antibody cetuximab (Erbitux™) is used against advanced colon cancer (Mendelsohn and Baselga 2006) (Figure 4).

Among the targeted chemical compounds, imatinib (Gleevec™), which is a small-molecule inhibitor of the oncogenic BCR-ABL kinase, is successfully used against leukemias that are caused by this oncogene (Schiffer 2007). The EGFR inhibitor erlotinib (Tarceva™) is used against lung carcinomas that are driven by a mutant EGFR (Ciardiello and Tortora 2008). Moreover, although different cancer subtypes in a given organ may have very different sets of driving mutations, certain cancer subtypes in different organs may surprisingly share common mutations. As a result, the same drug may be effective on molecularly related tumors in different organs. A related class of anti-cancer compounds is the angiogenesis inhibitors, which prevent the formation of blood capillaries that feed tumors. Some, such as the monoclonal antibody bevacizumab (Avastin™), are in clinical use (Ferrara 2002) (Figure 4). However, this drug has met with limited clinical success because cancer cells have multiple ways to stimulate angiogenesis (Meyerhardt and Mayer 2005). Current investigation is focusing on identifying combinations of angiogenic inhibitors that would be effective (Bergers and Hanahan 2008).

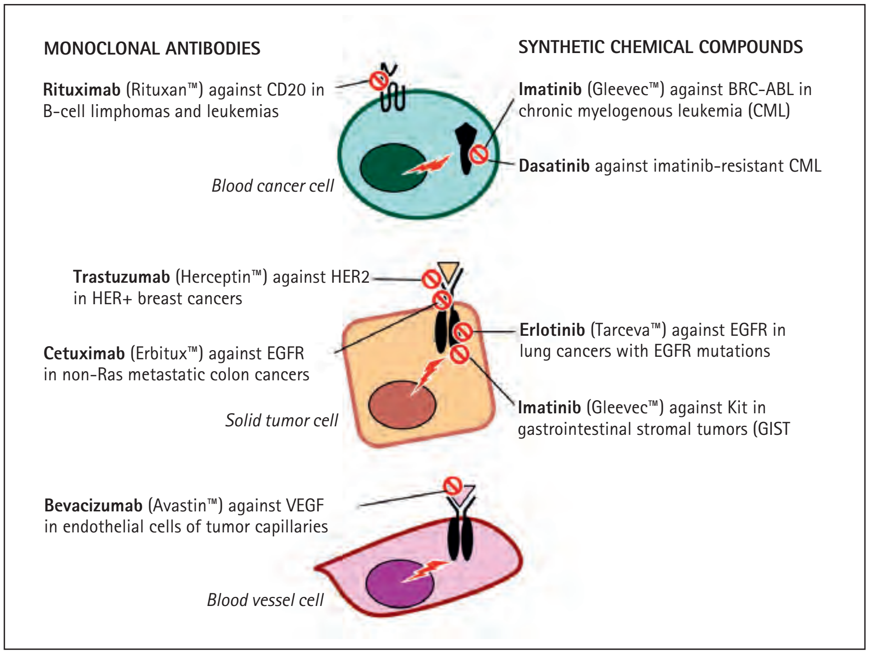

MONOCLONAL ANTIBODIES
SYNTHETIC CHEMICAL COMPOUNDS
Rituximab (Rituxan™) against CD20 in B-cell limphomas and leukemias
Imatinib (Gleevec™) against BRC-ABL in chronic myelogenous leukemia (CML)
Dasatinib against imatinib-resistant CML
Blood cancer cell
Trastuzumab (Herceptin™) against HER2 in HER+ breast cancers
Cetuximab (Erbitux™) against EGFR in non-Ras metastatic colon cancers
Erlotinib (Tarceva™) against EGFR in lung cancers with EGFR mutations
Imatinib (Gleevec™) against Kit in gastrointestinal stromal tumors (GIST
Solid tumor cell
Bevacizumab (Avastin™) against VEGF in endothelial cells of tumor capillaries
Blood vessel cell

Figure 4. Targeted drugs against cancer. Knowledge about the genes and gene protein products that drive certain cancers has allowed the development of drugs that specifically target these proteins blocking their activity and thereby killing cancer cells that depend on these proteins for their survival. These targeted drugs include monoclonal antibodies as well as synthetic chemical compounds. See text for details.

The barriers to improving the treatment of cancer still remain difficult, which underscores the need to add new directions to the future of cancer therapy. What changes in oncology can be envisioned? In the not too distant future, a patient's tumor profile could include not only histopathological grading and information on the status of common oncogenic mutations but also a full molecular portrait of the tumor (Massagué 2007; Nevins and Potti 2007). Recent progress in molecular profiling of tumors has led to the discovery of gene-expression signatures that allow a better classification of tumors into distinct subtypes, a better prediction of the risk and site of metastasis, and better identification of relevant therapeutic targets (Bild et al. 2006; Fan et al. 2006; Minn et al. 2005; Padua et al. 2008; van 't Veer et al. 2002; van de Vijver et al. 2002). A 70-gene "poor-prognosis" signature (MammaPrint) and a non-overlapping set of 21 "recurrence" genes (Oncotype Dx) have been turned into commercial products that assist clinicians in decisions to spare breast cancer patients with good-prognosis tumors from chemotherapy when this treatment is not required. Genes from these signatures can then be directly tested for their ability to mediate metastasis and to serve as targets of drugs that diminish the metastatic activity of cancer cells (Gupta et al. 2007). Drug regimens for cancer patients might include individualized multi-drug combinations targeting specific disease subtypes and metastatic sites. Betters biomarkers of drug response in patients will help better assess the response of individual patients to targeted therapies (Sawyers 2008).

With the recent new knowledge about the molecular, genetic, and cellular bases for cancer development and progression comes new opportunities to improve and expand our ability to prevent, detect, and treat this disease. Working closely together, clinicians and scientists can generate and apply the necessary knowledge to relegate cancer to the status of one more curable or indolent disease in the next few decades.

Bibliography

American Cancer Society. *Cancer facts and figures*, 2008. *http://www.cancer.org/*.

Bergers, G., and D. Hanahan. "Modes of resistance to anti-angiogenic therapy." *Nature Reviews Cancer* 8, 2008, 592–603.

Bierie, B., and H. L. Moses. "Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer." *Nature Reviews Cancer 6,* 2006, 506–520.

Bild, A. H., G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chass, M. B. Joshl, et al. "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." *Nature* 439, 2006, 353–357.

Billingsley, K. G., M. E. Burt, E. Jara, R. J. Ginsberg, J. M. Woodruff, D. H. Leung, and M. F. Brennan. "Pulmonary metastases from soft tissue sarcoma: analysis of patterns of diseases and postmetastasis survival." *Annals of Surgery* 229, 1999, 602–610; discussion 610–602.

Blasco, M. A. "The epigenetic regulation of mammalian telomeres." *Nature Reviews Genetics* 8, 2007, 299–309.

Boshoff, C., and Weiss, R. "AIDS-related malignancies." *Nature Reviews Cancer* 2, 2002, 373–382.

Cano, A., M. A. Pérez–Moreno, I. Rodrigo, A. Locascio, M. J. Blanco, M. G. Del Barrio, F. Portillo, and M. A. Nieto. "The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression." *Nature Cell Biology* 2, 2000, 76–83.

Cheson, B. D., and J. P. Leonard. "Monoclonal antibody therapy for B-cell non-Hodgkin's lymphoma." *The New England Journal of Medicine* 359, 2008, 613–626.

Cheung, T. K., H. H. Xia, and B. C. Wong. "Helicobacter pylori eradication for gastric cancer prevention." *Journal of Gastroenterology* 42, Suppl 17, 2007, 10–15.

Christofori, G. "New signals from the invasive front." *Nature* 441, 2006, 444–450.

Ciardiello, F., and G. TORTORA. "EGFR antagonists in cancer treatment." *The New England Journal of Medicine* 358, 2008, 1160–1174.

Clark, E. A., T. R. Golub, E. S. Lander, and R. O. Hynes. "Genomic analysis of metastasis reveals an essential role for RhoC." *Nature* 406, 2000, 532–535.

Clarke, M.F., and M. Fuller. "Stem cells and cancer: two faces of eve." *Cell* 124, 2006, 1111–1115.

Classon, M., and E. Harlow (2002). "The retinoblastoma tumour suppressor in development and cancer." *Nature Reviews Cancer* 2, 2002, 910–917.

Danaei, G., S. Vander Hoorn, A. D. Lopez, C. J. Murray, and M. Ezzati. "Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors." *Lancet* 366, 2005, 1784–1793.

Chapelle, A. de la. "Genetic predisposition to colorectal cancer." *Nature Reviews Cancer* 4, 204, 769–780.

Esteller, M. "Cancer epigenomics: DNA methylomes and histone-modification maps." *Nature Reviews Genetics* 8, 2007, 286–298.

Fan, C., D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. Van't Veer, and C. M. Perou. "Concordance among gene-expression-based predictors for breast cancer." *The New England Journal of Medicine* 355, 2006, 560–569.

Ferrara, N. "VEGF and the quest for tumour angiogenesis factors." *Nature Reviews Cancer* 2, 2002, 795–803.

Fidler, I. J. "The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited." *Nature Reviews Cancer* 3, 2003, 453–458.

Fodde, R., R. Smits, and H. Clevers. "APC, signal transduction and genetic instability in colorectal cancer." *Nature Reviews Cancer* 1, 2001, 55–67.

Gavrilovic, I. T., and J. B. Posner, J.B. "Brain metastases: epidemiology and pathophysiology." *J Neurooncology* 75, 2005, 5–14.

Gupta, G. P., and J. Massagué. "Cancer metastasis: building a framework." *Cell* 127, 2006, 679–695.

—, D. X. Nguyen, A. C. Chiang, P. D. Bos, J. Y. Kim, C. Nadal, R. R. Gomis, K. Manova-Todorova, and J. Massagué. "Mediators of vascular remodelling co-opted for sequential steps in lung metastasis." *Nature* 446, 2007, 765–770.

Halazonetis, T.D., V. G. Gorgoulis, and J. Bartek. "An oncogeneinduced DNA damage model for cancer development." *Science* 319, 2008, 1352–1355.

Hanahan, D., and R. A. Weinberg. (2000). "The hallmarks of cancer." *Cell* 100, 2000, 57–70.

Hausen, H. zur. "Viruses in human cancers." *European Journal of Cancer* 35, 1999, 1174–1181.

Hess, K.R., G. R. Varadhachary, S. H. Taylor, W. Wei, M. N. Raber, R. Lenzi, and J. L. Abbruzzese. "Metastatic patterns in adenocarcinoma." *Cancer* 106, 2006, 1624–1633.

Hudis, C.A. "Trastuzumab–mechanism of action and use in clinical practice." The *New England Journal of Medicine* 357, 2007, 39–51.

Husemann, Y., J. B. Geigl, F. Schubert, P. Musiani, M. Meyer, E. Burghart, G. Forni, et al. "Systemic spread is an early step in breast cancer." *Cancer Cell* 13, 2008, 58–68.

Jemal, A., T. Murray, E. Ward, A. Samuels, R. C. Tiwari, A. Ghafoor, E. J. Feuer, and M. J. Thun. "Cancer statistics, 2005." *CA: a cancer journal for clinicians* 55, 2005, 10–30.

Joyce, J. A. "Therapeutic targeting of the tumor microenvironment." *Cancer Cell* 7, 2005, 513–520.

Kaelin, W. G. "The von Hippel-Lindau tumor suppressor protein: roles in cancer and oxygen sensing." *Cold Spring Harbor Symposia in Quantitative Biology 70*, 2005, 159–166.

Kang, Y., P. M. Siegel, W. Shu, M. Drobnjak, S. M. Kakonen, C. Cordon-Cardo, T. A. Guise, and J. Massagué. "A multigenic program mediating breast cancer metastasis to bone." *Cancer Cell* 3, 2003, 537–549.

Karin, M. "Nuclear factor-kappaB in cancer development and progression." *Nature* 441, 2006, 431–436.

Karpozilos, A., and N. Pavlidis. "The treatment of cancer in Greek antiquity." *European Journal of Cancer* 40, 2004, 2033–2040.

Kim, M., J. D. Gans, C. Nogueira, A. Wang, J. H. Paik, B. Feng, C. Brennan, et al. "Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene." *Cell* 125, 2006, 1269–1281.

Klein, C. A., T. J. Blankenstein, O. Schmidt-Kittler, M. Petronio, B. Polzer, N. H. Stoecklein, and G. Riethmuller. "Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer." *Lancet* 360, 2002, 683–689.

Leiter, U., F. Meier, B. Schittek, and C. Garbe. "The natural course of cutaneous melanoma." *Journal of Surgical Oncology* 86, 2004, 172–178.

Lewis, C. E., and J. W. Pollard. "Distinct role of macrophages in different tumor microenvironments." *Cancer Research* 66, 2006, 605–612.

Malumbres, M., and M. Barbacid. "Cell cycle kinases in cancer*." Current Opinion in Genetics and Development* 17, 2007, 60–65.

Martin, R. W., P. P. Connell, and D. K. Bishop. "The Yin and Yang of treating BRCA-deficient tumors." *Cell* 132, 2008, 919–920.Massagué, J. "G1 cell-cycle control and cancer." *Nature* 432, 2004, 298–306.

—, "Sorting out breast-cancer gene signatures." *The New England Journal of Medicine* 356, 2007, 294–297.

—, "TGFbeta in Cancer." *Cell* 134, 2008, 215–230.

Melo, J. V., and D. J. Barnes. "Chronic myeloid leukaemia as a model of disease evolution in human cancer." *Nature Reviews Cancer* 7, 2007, 441–453.

Mendelsohn, J., and J. Baselga. "Epidermal growth factor receptor targeting in cancer." *Seminars in Oncology* 33, 2006, 369–385.

Meyerhardt, J. A., and R. J. Mayer. "Systemic therapy for colorectal cancer." *The New England Journal of Medicine* 352, 2005, 476–487.

Minn, A. J., G. P. Gupta, D. Padua, P. D. Bos, D. X. Nguyen, D. Nuyten, B. Kreike, *et al.* "Lung metastasis genes couple breast tumor size and metastatic spread." *Proceedings of the National Academy of Sciences USA* 104, 2007, 6740–6745.

Minn, A. J., G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald, and J. MASSAGUÉ. "Genes that mediate breast cancer metastasis to lung." *Nature* 436, 2005, 518–524.

Mueller, M. M., and N. E. Fusenig. Friends or foes—bipolar effects of the tumour stroma in cancer. *Nature Reviews Cancer* 4, 2004, 839–849.

Nevins, J. R., and A. Potti. "Mining gene expression profiles: expression signatures as cancer phenotypes." *Nature Reviews Genetics* 8, 2007, 601–609.

Norton, L., and J. Massagué "Is cancer a disease of self-seeding?" *Nature Medicine* 12, 2006, 875–878.

Nowell, P.C. "The clonal evolution of tumor cell populations." *Science* 194, 1976, 23–28.

Olsson, A.K., A. Dimberg, J. Kreuger, and L. Claesson-Welsh. "VEGF receptor signalling—in control of vascular function." *Nature Reviews Molecular Cell Biology* 7, 2006, 359–371.

Padua, D., X. H. Zhang, Q. Wang, C. Nadal, W. L. Gerald, R. R. Gomis, and J. Massagué. "TGFbeta primes breast tumors for lung metastasis seeding through angiopoietin-like 4." *Cell* 133, 2008, 66–77.

Paget, S. "The distribution of secondary growths in cancer of the breast." *Lancet* 1, 1889, 571–573.

Parato, K. A., D. Senger, P. A. Forsyth, and J. C. Bell. "Recent progress in the battle between oncolytic viruses and tumours." *Nature Reviews Cancer* 5, 2005, 965–976.

Perl, A. K., P. Wilgenbus, U. Dahl, H. Semb, and G. Christofori. "A causal role for E-cadherin in the transition from adenoma to carcinoma." *Nature* 392, 1998, 190–193.

Pouyssegur, J., F. Dayan, and N. M. Mazure. "Hypoxia signalling in cancer and approaches to enforce tumour regression." *Nature* 441, 2006, 437–443.

Roden, R., A. Monie, and T. C. WU. "The impact of preventive HPV vaccination." *Discovery Medicine* 6, 2006, 175–181.

Rustgi, A. K. "The genetics of hereditary colon cancer." *Genes and Development* 21, 2007, 2525–2538.

Sawyers, C. "Targeted cancer therapy." *Nature* 432, 2004, 294–297.

Sawyers, C. L. "The cancer biomarker problem." *Nature* 452, 2008, 548–552.

Schiffer, C. A. "BCR–ABL tyrosine kinase inhibitors for chronic myelogenous leukemia." *The New England Journal of Medicine* 357, 2007, 258–265.

Schmidt-Kittler, O., T. Ragg, A. Daskalakis, M. Granzow, A. Ahr, T. J. Blankenstein, M. Kaufmann, et al. "From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression." *Proceedings of the National Academy of Sciences USA* 100, 2003, 7737–7742.

Shawver, L.K., D. Slamon, and A. Ullrich. "Smart drugs: tyrosine kinase inhibitors in cancer therapy." *Cancer Cell* 1, 2002, 117–123.

Sweet-Cordero, A., S. Mukherjee, A. Subramanian, H. You, J. J. Roix, C. Ladd–Acosta, J. Mesirov, T. R. Golub, and T. Jacks. "An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis." *Nature Genetics* 37, 2005, 48–55.

Thiery, J.P. "Epithelial-mesenchymal transitions in tumour progression." *Nature Reviews Cancer* 2, 2002, 442–454.

Veer, L. J. van't, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A.T. Witteveen, et al. "Gene expression profiling predicts clinical outcome of breast cancer." *Nature* 415, 2002, 530–536.

Vijver, M. J. van de, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, et al. "A gene-expression signature as a predictor of survival in breast cancer." *The New England Journal of Medicine* 347, 2002, 1999–2009.

Vogelstein, B., E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. Smits, and J. L. Bos. "Genetic alterations during colorectal-tumor development." *The New England Journal of Medicine* 319, 1988, 525–532.

—, K. W. Kinzler. "Cancer genes and the pathways they control." *Nature Medicine* 10, 2004, 789–799.

Vousden, K.H., and D. P. Lane. "p53 in health and disease." *Nature Reviews Molecular Cell Biology* 8, 2007, 275–283.

Walsh, T., and M. C. King. "Ten genes for inherited breast cancer." *Cancer Cell* 11, 2007, 103–105.

Wang, C., Y. Yuan, Y., and R. H. Hunt. "The association between Helicobacter pylori infection and early gastric cancer: a meta-analysis." *The American Journal of Gastroenterology* 102, 2007, 1789–1798.

Wang, J., R. Loberg, and R. S. Taichman, R.S. "The pivotal role of CXCL12 (SDF–1)/CXCR4 axis in bone metastasis." *Cancer Metastasis Reviews* 25, 2006, 573–587.

Wang, W. "Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins." *Nature Reviews Genetics* 8, 2007, 735–748.

Weinberg, R. A. *The biology of cancer*. New York: Garland Science, 2007.

Welcsh, P. L., and M. C. King. "BRCA1 and BRCA2 and the genetics of breast and ovarian cancer." *Human Molecular Genetics* 10, 2001, 705–713.

Woodman, C. B., S. I. Collins, and L. S. Young, L.S. "The natural history of cervical HPV infection: unresolved issues." *Nature Reviews Cancer* 7, 2007, 11–22.

World Heath Organization. *Cancer*. Fact Sheet No 297, 2008. *http://www.who.int/mediacentre/factsheets/fs297/en/index.html*.

Yang, J., S. A. Mani, J. L. Donaher, S. Ramaswamy, R. A. Itzykson, C. Come, P. Savagner, I. Gitelman, A. Richardson, and R. A. Weinberg. "Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis." *Cell* 117, 2004, 927–939.

Young, L.S., and A. B. Rickinson. "Epstein-Barr virus: 40 years on." *Nature Reviews Cancer* 4, 2004, 757–768.

the garden of eden endangered: the ecology and biology of conservation

CARLOS M. DUARTE

The United Nations has declared 2009 the International Year of Biodiversity, in homage to the bicentennial of the birth of Charles Darwin (1809–1882), whose book, *On the Origin of the Species by Natural Selection* (1859) marks the beginning of the science of biodiversity. At last, everything seemed to make sense: the subtle differences among similar species, the colorful plumage of birds and flowers, the numerous adaptations of animals to their surroundings, and the failures, reflected in fossils of fabulous animals that the Church—strapped for a better explanation—condemned as a clever trick by the devil to confuse the faithful. Science had formulated a rational explanation for what had only been explicable until then as the result of supernatural acts. Needless to say, Darwin's revolutionary theses were energetically combated for years. In Spain, the Cantabrian biologist, Augusto G. de L. (1845–1904), lost his post as Senior Professor of Natural History at the University of Santiago de Compostela for teaching Darwin's theories.

Darwin's work, combined with Mendel's Laws—the monk, Gregor J. Mendel (1822–1884), described the basic laws of genetic inheritance—is the seed from which modern biology has grown, triggering an unstoppable and logical sequence of fundamental discoveries such as DNA and the modern genome. The 150 years since the *Origin of the Species* was published are studded with achievements that have shaped a new science, Ecology, which seeks to decipher the keys to the functioning of the biosphere and the role of biodiversity in the balance of nature, expanding the frontiers of knowledge in order, in a moment of severe crisis, to delve deeper into the foundations of the present and future wellbeing of humanity.

In this chapter, I will offer a summary of the achievements and developments that mark the path leading to our understanding of how nature works, and I will point out the challenges we face in the twenty-first century. Rather than following a chronological order, which would offer a disorderly view of progress in this field, I have opted for a thematic organization in which I emphasize the most important achievements and challenges.

The origin and diversification of life

The ocean is the cradle of life on earth. The oldest existing fossils were found in Australia and date from around 3,500 million years ago. They are of groupings of microorganisms with photosynthetic archaea and cyanobacteria that formed carbonate structures similar to the stromatolytes that still survive in different parts of the planet, including Australia (illustration 1).

The oldest existing organisms are microorganisms belonging to the domain of Archaea, which still constitute an important part of the biological communities in the deep oceans. This discovery of Archaea is a recent achievement that has revolutionized our conception of the organization of biological diversity. In 1977, the US microbiologist, Carl R. Woose was the first to use ribosomic RNA to establish relations among microorganisms. He discovered that communities of bottom-dwelling microorganisms included some that represented a new domain, different than both bacteria and eukaryotes. The development of molecular probes capable of distinguishing between bacteria and Archaea, which cannot be told apart under a microscope, has revealed that this group is present all over the planet and that they are particularly prominent in deep parts of the ocean—where there are habitats with conditions similar to those that existed when Archaea first appeared—and also in polar lakes. The discovery of Archaea led to a revision of the domains of life, and the recognition of three: Bacteria, Archaea, and Eukarya, which completely transformed traditional concepts.

Earth's primitive atmosphere lacked oxygen. It was very reductive and lacking in ozone, so ultraviolet radiation penetrated it easily, reaching the Earth's surface with an intensity that is incompatible with life. Only in the ocean, where ultraviolet radiation is strongly attenuated by deep water, was it possible for life to prosper on such a highly irradiated planet. Marine biota deeply and irreversibly altered the Earth's atmosphere, thus altering conditions for life on the continents as well. Specifically, the apparition of oxygen-producing photosynthesis—which produces oxygen by photolysis of water (the photosynthetic process that is characteristic of plants)—in marine microorganisms called cyanobacteria produced a fundamental change in the composition of the Earth's atmosphere: the apparition of oxygen. It now counts for 21% of the atmosphere and when it reacts to ultraviolet radiation in the Stratosphere (around 12,000 meters above the Earth's surface), it generates ozone that absorbs the most harmful ultraviolet radiation, allowing life on land. The concentration of CO_2 in the atmosphere also diminished, as the increase of O_2 is only possible when CO_2 is consumed at a proportional rate by photosynthesis and stored in organic form in seawater, soil, organisms, detritus, and petroleum and gas deposits. The change from a reductive atmosphere to an oxidizing atmosphere is a fundamental change that completely conditions all planetary chemistry including the functioning of the biosphere and the evolution of life.

According to fossil evidence, the origin of the cyanobacteria responsible for this change on Earth was relatively abrupt. It continues to be a mystery and we cannot rule out an extraterrestrial origin. In fact, the apparition of life in the ocean brought about a determinant transformation, not only of the atmosphere, but of the lithosphere as well, because the formation of carbonate and other minerals by marine organisms created the base for many sedimentary rock formations.

There are animal fossils dating back 800 million years, although the first complex animals appeared around 640 million years ago, again in Australia. The first animal occupation of the continents dates from a little over 400 million years ago, and we have found centipede and spider fossils from that time. In fact, the occupation of the continents by life would not have been possible without the alteration of the conditions on planet Earth brought about by primitive marine organisms.

So the evolutionary history of life is much longer in the ocean than on dry land, and this is reflected by the greater diversity of life forms in the ocean. While the ocean contains a modest proportion of the species that inhabit the Earth, it contains an almost complete repertory of all the genomic diversity generated by evolution. Genomic diversity refers to the diversity of genetic machinery made up of genes that codify the proteins that determine different functions. For example, it is sufficient to consider that the genomes of a worm or a fruit fly differ from the human genome in less than half their sequences, so the path of genomic diversity among land animals is relatively short.

The tree of life that reflects the diversification of life forms on Earth has its roots in the ocean. There are 30 phyla—the large branches of that tree—in the ocean, thirteen of which are only found there. In comparison, only 15 phyla have been found on dry land, and only one of them is exclusive to it. In fact, the diversity of life forms in the ocean is often perplexing. For example, many sessile and colored organisms, similar to the flowers that adorn our landscapes, are actually animals—like anemones—or a mixture of animal and plant, like colored tropical coral, whose color is due to the pigments of photosynthetic algae that live among the colonies of polyps that form the coral. In fact, the simple division between animal and plant that is useful on land is frequently misleading in the ocean, as many animals are actually consortiums of photosynthetic species and animals, and many unicellular organisms have capacities belonging to both.

How many species are on this planet?

Ever since the Swedish scientist, Carl Linnaeus established the basis of taxonomy with a system of nomenclature for classifying living beings—in his *Systema Naturae,* published in 1735—the number of described species has never ceased to grow. There are now around 2 million described species. The inventory of species in the biosphere seems endless, although clearly the number of existing species must necessarily be finite. In recent years, there have been significant efforts to arrive at a trustworthy estimate of the number of species the biosphere may contain.

Despite the long evolutionary history of life in the oceans, only about 230,000 known species live there now. That is fifty times less than on land, where some 1.8 million known species currently live (Jauma and Duarte 2006; Bouchet 2006). This has intrigued scientists for many years, leading to diverse hypotheses that attempt to explain such a paradox. There has been talk of the enormous potential for dispersion by marine animals' propagules (eggs and larvae), which would avoid genetic segregation caused by the separation of populations. For example, there are only 58 species of superior marine plants, with seeds and fruit (angiosperms), as opposed to 300,000 on the continents. And there are practically no insects in the ocean, even though arthropods—including insects, crustaceans, arachnids, mites, and other lesser groups—constitute 91% of the inventory of land-based species. A variety of approaches have been taken to estimating what the total number of species might be. There have been extrapolations from the best-known to least-known taxa, assuming a proportionality of species; extrapolations based on the number of new species appearing per unit of examined area, times the total surface area occupied by different habitats; and statistical estimates based on the progression of the rate of discovery of new species. These estimates indicate that the total number of species could be around 12 million. Of these, the largest group would be insects, with almost 10 million species, and nematodes, with around 1 million species. The number of marine species could be slightly over 1 million, making it a little more than 10% of the total number of species (Bouchet 2006).

Illustration 1. Stromalytes at Shark Bay, Western Australia, where living stromalytes were discovered for the first time (photo: Carlos M. Duarte).

Discoveries in the exploration of biodiversity

Each year, around 16,000 new species are described, of which around 1,600 are oceanic (Bouchet 2006). The annual growth of the biodiversity inventory is close to 1%. Given that the current number of described species is thought to be about 10% of the total, at the present rate of discovery, it will take over 200 years to complete the inventory, and possibly longer in the case of marine species, whose inventory progresses more slowly than that of land animals. The Census of Marine Life (www.coml.org) is an international project that coordinates the efforts of thousands of researchers around the world in order to arrive at an inventory of all the existing species in the ocean. Each year, 1,635 new marine species are described—the great majority are crustaceans or mollusks—by close to 2,000 active marine taxonomists (Bouchet 2006). And yet it has been estimated that, at that rate of discovery, we will need from 250 to 1,000 years to complete the inventory of marine biodiversity, which could well have a total number of around a million and a half species—six times that which has been described up to now (Bouchet 2006).

This inventory work involves significant surprises involving not only microscopic organisms, but also relatively large vertebrates such as monkeys (for example, the mangabey monkey, *Lophocebus kipunji,* discovered in Tanzania in 2005) and fish. Although the number of species discovered each year on land is far greater than the number of marine species; discoveries on land are limited to new species from known genera or families, while the taxonomic range of innovations in the ocean is far wider. Our knowledge of the diversity of life in the oceans is still very limited, and the rate of discovery is still surprisingly high.

In the area of microscopic organisms, some of these discoveries also mark significant milestones in our knowledge. For example, the minute photosynthetic cyanobacteria from the genera *Synechococcus* (about 1µm in diameter) and *Prochloroccocus* (about 0.5µm in diameter) were discovered between the late nineteen seventies and the early nineteen eighties. Later studies revealed that those organisms dominate plankton in the large oceanic deserts that represent about 70% of the open seas and are responsible for almost 30% of oceanic photosynthetic production. The magnitude of this discovery and what it tells us about our degree of ignorance of the ocean can be properly understood if we consider that not knowing about these organisms until the late nineteen seventies is equivalent to not knowing there were tropical jungles on land until that time. The ocean continues to amaze us at higher taxonomic levels—even new phyla are being discovered—and that does not occur on land. These surprises include some of the largest animals on the planet, such as the giant squid, *Magnapinnidae,* with enormous fins, sighted various times in the deep ocean (over 2,000 meters deep); the wide-mouth shark, *Megachasma pelagios,* which can be 4 to 5 meters long—discovered in Indian-Pacific waters in 1983—or the small finback whale, *Balaenoptera omurai,* that reaches lengths of 9 meters and was discovered in the same area in 2003.

The greatest opportunities for new discoveries of marine biodiversity are in remote or extreme habitats. On land, the most spectacular discoveries frequently come from tropical jungles in remote and relatively unexplored parts of Asia (e.g. Vietnam), Africa (e.g. Tanzania), and Oceania (e.g. Papua-New Guinea). In the oceans, the remote areas of Southeast Asia and Oceania have the greatest diversity of all marine groups, while extreme habitats—sea trenches, submarine caves, hyper-saline or anoxic environments, hydrothermal springs, and pockets of hyper-saline or anoxic water—have the most surprises (Duarte 2006), along with the insides of organisms, which are home to symbionts. The latter term refers to guests, mutualists, and parasites, and is not limited to small species. For example, the largest known marine worm—up to six meters long—is a whale parasite.

Discoveries of marine biodiversity go far beyond the description of new species—no matter how surprising they may be—including the discovery of ecosystems with previously unknown communities and metabolic systems. In the late nineteen seventies, scientists aboard the US research submarine, *Alvin*, discovered the ecosystems of hydrothermal springs while making geothermal studies in the Galapagos rise (Lonsdale 1977; Corliss et al. 1979). They found an extraordinary seascape of black chimneys that emitted a smoke-like liquid composed of metals and other materials that precipitated as they cooled, creating those chimneys. The latter were colonized by dense masses of previously unknown animals, such as the giant tube worm, *Riftia pachyptila,* albino crabs, fish, and many other organisms, all new to science.

This discovery was not only an important addition to the inventory of marine species, it was also a complete challenge to our belief that solar light was the energy source that permitted the production of organic material—through plant photosynthesis—needed to maintain ecosystems. In the life-filled reefs around these hydrothermal springs, it is not the plants that transform energy into organic matter to feed the ecosystem. That work is carried out by chemoautotrophic bacteria and Archaea, which synthesize organic matter out of reduced inorganic compounds pushed out of the earth by the hydrothermal fluids (Karl, Wirsen, and Jannasch 1980; Jannasch and Mottl 1985). Those new habitats, where life prospers without the need for solar energy, are known as chemosynthetic ecosystems, where microorganisms establish symbiotic relations with invertebrates. Since they were discovered in 1977, around 600 species of organisms living there have been described. And since then, it has been discovered that other reductive habitats on the sea bed, such as the cold seeps of hydrothermal fluids (discovered in 1983 at a depth of 500 meters in the Gulf of Mexico), remains of whales, and zones with a minimum of oxygen, are also home to communities that depend on chemical energy, with communities similar to those of the animals found at hydrothermal springs.

These discoveries were a revolutionary milestone that completely modified our ideas about how ecosystems function and are organized. The microorganisms found in hydrothermal springs have also brought about a small revolution in biology and biotechnology, as many of them have proteins that are stable at 100ºC and that catalyze reactions at a vertiginous speed. *Pyrococcus furiosus* is a species of Archaea discovered in marine trenches off the island of Vulcano (Italy) that stand out because their optimum growth temperature is 100º C. At that temperature, they duplicate themselves every 37 minutes. They also possess enzymes that contain tungsten, which is rarely found in biological molecules. At that temperature, the polymerases of *Pyrococcus furiosus* DNA (Pfu DNA) operate at an enormous velocity, so they are often used in the chain reaction of the polymerase (PCR) that makes it possible to mass produce DNA fragments. It is the fundament of most biotechnology applications that require DNA sequencing.

New discoveries in marine biodiversity also depend on developments in the field of molecular techniques that make it possible to establish the taxonomic position of organisms by analyzing sections of their genome. For example, the use of massive sequencing techniques allowed the American biologist, Craig Venter—leader of the Celera Genomics Project that first sequenced the human genome—to sequence DNA fragments from 1 cubic meter of surface seawater from the Sargassos Sea. That exercise turned up a surprising inventory of 1,214,207 new genes, and close to 1,800 new species of microbes (Venter et al. 2004). Sadly, these techniques do not make it possible to identify the new species, but they are revealing that many anatomically similar marine species are actually different species. Moreover, they are also demonstrating that some species considered different due to their morphological dissimilarities are actually variants of the same species subjected to very different environmental conditions.

The biosphere under pressure: the Anthropocene

The Industrial Revolution, which increased the human capacity to transform the environment, was not only a milestone in the history of our species, but in the history of the planet, which has been transformed by human activity. Any objective study of planet Earth—its climate, the configuration and dynamics of its ecosystems, its basic functional processes—shows that they are affected by human activity. The importance of human activity's impact on the essential processes of the biosphere is reflected in certain indicators, such as the fact that 45% of the Earth's surface has already been transformed by human activity, passing from wild ecosystems to domesticated ones such as farm land, pastures and urban zones. Humanity uses more than half the available flow of fresh water in the world, modifying the amount of water that flows through rivers, and also altering its quality, enriching it with nutrients, nitrogen and phosphorus, organic matter, and contaminants following its use by humans. In fact, human activity notably accelerates the cycles of elements of the biosphere. It has mobilized over 420 gigatons of coal since the Industrial Revolution and—using the Haber Reaction patented by Fritz Haber in 1908—it has fixed 154 megatons per annum of atmospheric nitrogen gas in the form of ammonia for use in fertilizers. That is more atmospheric nitrogen than the processes of nitrogen fixation that occur as a result of nitrogenase activity from plants, terrestrial, and marine microorganisms. Carbon dioxide emissions due to the use of fossil fuels, the production of cement, and fires, along with the release of other greenhouse gasses such as methane, are raising the planet's temperature. When those gasses dissolve in the ocean, they increase its acidity. Those processes have important consequences for the Earth's climate and for the ecosystems it contains. It has also been calculated that human agriculture, forestry and fishing account for approximately 40% of land-based photosynthesis and 20% of costal photosynthesis, worldwide.

These data, to which many others could be added, are sufficient to substantiate the affirmation that our species has become an essential element of change in the basic processes of the biosphere. In 2000, this led the atmospheric chemist and Nobel prizewinner, Paul Crutzen, and his colleague, E. Stoermer, to propose the name *Anthropocene* to designate a new geological era in the history of the planet. An era in which humanity has emerged as a new force capable of controlling the fundamental processes of the biosphere (Crutzen and Stoermer 2000), causing Global Change.

The human capacity to alter the planet begins with the Holocene, at the end of the last ice age, about 10,000 years ago. This was followed by the development and rapid expansion of agriculture, animal husbandry, and the first urban centers. The first indications of this new emerging force are the extinction of large mammals and birds hunted by the first inhabitants of islands and continents. The development of agriculture and animal husbandry led to the transformation of land, converting forests and other ecosystems into farmland and pastures. And those changes were strengthened by the work capacity generated by domesticating beasts of burden (oxen, horse, etc.) and technological developments such as the plow and the wheel. The human capacity to transform the planet experimented a notable push with the Industrial Revolution, which increased the capacity to use energy to transform the planet. It also generated residues such as gasses and synthetic compounds that alter natural processes. Humanity has radically transformed the planet's territory, converting around 45% of the Earth's surface into pastures—they occupy around 30% of the Earth's surface—farmland—another 10%—and urban areas, that occupy approximately 2% of the Earth's surface. Other infrastructures, such as reservoirs, roads, electric lines, railways, etc., occupy another 3% of the planet's surface, approximately. Costal zones are experiencing the highest rates of population growth on the planet. About 40% of the human population lives less than 100 kilometers from the coast, with a population density three times greater than that of continental territories. And the coastal population is growing much more rapidly than the continental one, due to migration, the increased fertility of coastal zones, and increased tourist flow to those areas (Millennium Assessment 2005b). Moreover, the coastline itself is being rapidly occupied by infrastructures (housing, streets and roads, ports, and so on).

Human activity has accelerated the cycles of elements in the biosphere—processes central to the regulation of how this system, and life itself, function. The acceleration of elemental cycles affects practically all chemical elements, but it has more important consequences for those involved in processes essential to the regulation of life—carbon, nitrogen, phosphorus, iron, calcium, and other oligoelements—and of the climate, including carbon—through CO_2 and methane—and nitrogen—through nitrous oxide. The transformation of forests into pastures and farmland accelerates the carbon cycle. No longer trapped in forest biomass, it is rapidly recycled in annual harvests. Agricultural land has less capacity to retain carbon than forested land, and the destruction of wetlands has released carbon retained by those systems, which are important carbon sinks. The extraction of fossil fuels and gasses also mobilizes carbon that had accumulated during epochs in which the biosphere generated an excess of primary production.

The use of fossil fuels, along with the production of CO_2 in cement making, deforestation, and forest fires, has led to emissions of around 450 gigatons of CO_2 into the atmosphere, which has led to a rapid increase in the atmospheric concentration of CO_2, along with other greenhouse gasses such as methane and nitrous oxide. Human activity also generates an excessive mobilization of nitrogen, fundamentally through the production of some 154 million tons of this element every year in the form of fertilizers made from atmospheric nitrogen gas. That nitrogen is mobilized by its transportation in rivers, in the atmosphere, and also as nitrate contamination in the aquifers. Atmospheric transportation allows nitrogen to be carried long distances, so that it also deposits on the open seas. The production of fertilizers requires the extraction from mineral deposits of a quantity of phosphorus proportional to the amount of nitrogen produced in fertilizers. The acceleration of the cycles of those elements has important consequences for the ecosystems, which are altered by a process called eutrophization. That process is caused by an excessive contribution of nutrients to ecosystems and has significant consequences for them.

Humanity currently uses 50% of the fresh water available in the biosphere. In 1995, we extracted over 3,000 cubic kilometers of water for irrigating crops. Food production, including pastures, annually consumes around 14,000 cubic kilometers of water. As a consequence of this agricultural water use, large lakes such as the Aral Sea, in Central Asia, have lost most of their extension and water volume. The Aral Sea's water level drops by 0.6 meters each year, while the surface area of Lake Chad, in Africa, has shrunk by a factor of 20 in just 15 years. Human water use and the transformation of land have resulted in significant changes in the water cycle. Approximately 60% of the marshes existing in Europe in 1800 have disappeared. Construction of reservoirs grew rapidly during the twentieth century, at a rate of 1% per year, and these now rZXetain approximately 10,000 cubic kilometers of water, which is five times as much water as is contained in rivers.

Human activity has synthesized millions of new chemical compounds that were inexistent in the biosphere. They often act as contaminants that harm organisms, including our own species, or they interfere with other processes. For example, Freon and Halon gases used in industry and refrigeration are responsible for the destruction of the ozone layer, which has decayed at an annual rate of around 4% over the last two decades, causing the hole in the ozone layer to expand in the Southern Hemisphere. These compounds have now been controlled—by the Montreal Protocol of 1987—but every year, thousands of new substances are released into the biosphere without any previous testing to determine what impact they may have on human health and the biosphere. Some of them behave like greenhouse gasses and exacerbate the process of global warming. Many such compounds are volatile or semi-volatile and are transported by the atmosphere to areas thousands of kilometers from their sources, so there are no places left in the biosphere that are free of them.

Emissions of greenhouse gasses are causing a strong increase in the planet's temperature, which has already risen by 0.7°C. The temperature is expected to rise another two to seven degrees centigrade over the course of the twenty-first century (Trenberth et al. 2007, Meehl et al. 2007). Besides the temperature increase, other components of the climatic system will also be affected. Important changes in the water cycles are expected, with an increase of precipitation in some parts of the planet and a decrease in others, as well as more frequent and prolonged extreme events such as droughts and flooding (Meehl et al. 2007). The intensity of the wind will increase and extreme events such as tropical cyclones are expected to increase in intensity, reaching areas that have been free of such phenomena until now (Meehl et al. 2007).

Global warming led to an average rise in sea levels of 15 centimeters during the twentieth century, and an additional increase of between 30 and 80 centimeters is projected for the twenty-first century (Bindoff et al. 2007). The increase of partial CO_2 pressure in the atmosphere and its penetration in the ocean has led the latter's pH to drop by approximately 0.15 units. Given that the pH scale is logarithmic, that indicates a 60% increase in oceanic acidity. The increase of partial CO_2 pressure predicted for the twenty-first century will lead to an additional drop of between 0.3 and 0.4 units, which means that the ocean's acidity will have tripled by then.

The impact of Global Change on the ecosystems

The transformation of land by the expansion of pastures, farmland, and urban and industrial areas has been carried out at the expense of ecosystems such as wetlands—many have been drained—tropical forests and other habitats essential to the conservation of biodiversity. Wetlands represent 6% of the Earth's surface, and more than 50% of the wetlands in North America, Europe, Australia, and New Zealand have already been lost. A large part of these and other regions have broken down. In the Mediterranean basin, more than 28% of wetlands were lost in the twentieth century. Forests have suffered important losses as well, as about 40% of the planet's forested area has disappeared in the last three centuries. Forests have completely disappeared in 25 countries, and another 29 have lost over 90% of their forested land. Forested areas are currently expanding in Europe and North America, but they continue to diminish in the tropics at a rate of 10 million hectares per year, which is about 0.5% a year (Millennium Assessment 2005b). The intense occupation of costal zones is causing important losses of coastal ecosystems, which are experiencing the greatest loss rates of all: about 25% of mangrove swamps have been lost, about a third of all coral reefs have been destroyed (Millennium Assessment 2005b), and undersea prairies are shrinking at a rate of two to five percent per annum (Duarte 2002).

Planetary warming is making spectacular changes in the areas of our planet occupied by frozen surfaces, such as the sea ice in the Arctic, which suffered a catastrophic decrease in 2007, and the extensions of Alpine glaciers, which are clearly receding as a result of global warming.

The increase of partial CO_2 pressure will increase rates of photosynthesis, especially by aquatic photosynthetic organisms, as the enzyme responsible for fixing CO_2 evolved when the concentration was much greater than it now is, and its activity is relatively inefficient at current CO_2 levels. Photosynthetic activity will also be increased by temperature increases, as the latter accelerate metabolic rates. However, breathing is a process that is much more sensitive to temperature increases and, in the biosphere, which is dominated by microbe processes, breathing is expected to increase by as much as 40% in the current warming scenario, while primary production would increase by around 20% (Harris et al. 2006). This could lead to a net CO_2 production in aquatic ecosystems—including the oceans—that would worsen the greenhouse effect.

The process of eutrophization resulting from human activity's mobilization of large quantities of nitrogen and phosphorus is leading to an increase in primary production on land and in the seas. Eutrophization implies a breakdown in water quality, the loss of submerged vegetation and the development of alga proliferations, some of which are toxic. When other circumstances coincide with it, such as poor ventilation of water, hypoxia can also spread. Eutrophization is not limited to the continents. It can also affect the open seas, where atmospheric nitrogen contributions have doubled, undoubtedly with significant consequences for the functioning of the oceans, although there is not yet enough research to clearly establish this.

The effects of climate change are particularly clear in the phenological patterns of organisms. Behavioral and reproductive patterns are also suffering, and will suffer, alterations, with earlier flowering in temperate zones and alterations in birds' migratory periods. Activities that organisms begin to carry out in spring in temperate zones are already causing changes in the biogeographic ranges of organisms, with a displacement towards higher latitudes. This displacement includes pathogenic organisms, so the range of tropical or subtropical diseases is also expected to move to higher latitudes. Besides these latitudinal displacements, different organisms also change their range of altitudes. The tree line on high mountains is reaching higher elevations and alpine organisms are extending their upper limit by one to four meters per decade. These changes are leading to relatively important alterations in the makeup of communities in almost every ecosystem on the planet.

Global Change and the conjunction of its multiple effects (warming and eutrophization) seem to be leading to an increase in the problem of hypoxia in coastal waters, where affected areas are increasing by 5% annually (Vaquer-Sunyer and Duarte 2008). Hypoxia is when oxygen levels in coastal waters drop below two to four milligrams per liter, leading to the death of many groups of animals and plants and the release of sedimentary phosphorus. Three circumstances must coincide in order for hypoxia to occur: a) an excess of photosynthetic production that sediments waters in contact with the sea floor; b) stratification through a density gradient due to a thermal gradient, a salinity gradient, or both, between surface water in contact with the atmosphere, and deeper coastal water in contact with marine sediment, so that this stratification creates a barrier that keeps water from ventilating and renewing its oxygen content; and c) increased respiration in the deepest layer of water. Those three processes are affected by Global Change: global eutrophization is increasing coastal production on the basis of increased nitrogen and phosphorous; rising temperatures increase the stratification of the water column, reducing the ventilation of underlying gasses and increasing the breathing rate. Thus, Global Change is expected to considerably increase the breadth and intensity of hypoxia problems and the mortality of marine organisms affected by it in coastal regions (Vaquer-Sunyer and Duarte 2008).

The acidification of the ocean mainly affects organisms with carbonate skeletons. Those in cold oceans are particularly vulnerable to this process, so the polar oceans will be the first to be affected by this oceanic acidification, with problems for the development of organisms with calcified structures. These difficulties will later affect organisms in temperate seas as well, and will eventually reach the tropics.

Coral reefs are particularly vulnerable to temperature increases, as the photosynthetic symbionts that live there and depend on them for adequate growth, die when water temperatures surpass 29ºC. This will be more frequent in the future. In fact, the coral reefs in South East Asia have recently experience massive episodes of whitening (i.e. loss of zooxanthellae symbionts). Coral reefs also suffer the consequences of global eutrophization and the acidification of seawater, and are thus thought to be among the ecosystems most gravely affected by Global Change.

Finally, the accelerated loss of surface ice during the Arctic summer is seriously endangering species that depend on ice for their habitat, including polar bears, seals, and walruses.

The ecosystems' responses to these simultaneous pressures are frequently manifested as abrupt changes of communities. These are known as regime changes and constitute brusque transitions between two states (e.g. shallow lakes dominated by vegetation rooted on the bottom becoming lakes dominated by phytoplankton due to eutrophization, and sea floors with vegetation and fauna that become sea beds dominated by microbe carpets due to hypoxia). These transitions occur following a small increase of pressure that pushes them over a threshold, triggering the change. The first theoretical speculation about these abrupt regime changes in the state of ecosystems dates from the nineteen seventies (May 1977). Since then, it has been shown that these changes are not the exception, but rather the most frequent response by ecosystems subjected to pressure (Scheffer and Carpenter 2003; Andersen et al. 2008). It has also been shown that once the threshold that triggers the regime change is crossed, it is very difficult to return the system to its previous state. That is why it is so important to determine the position of those thresholds. Sadly, at present we are only able to identify those thresholds when they have been crossed (Strange 2008).

Toward the sixth extinction? Extinctions and the biodiversity crisis

The extinction of species is as natural as the emergence of new ones resulting from the slow process of evolution. Fossil evidence indicates there were five great extinctions in our planet's turbulent past. The first took place about 440 million years ago and was apparently due to a climate change that led to the loss of 25% of existing families. The second great extinction, with a loss of 19% of species, took place 370 million years ago, possibly due to global climate change. The third and greatest extinction took place 245 million years ago, possibly due to climate change caused by the impact of a large meteorite. It led to the loss of 54% of existing families. The fourth great extinction, 210 million years ago, caused the loss of 23% of existing families, and its causes are the source of speculation, including a possible increase in ultraviolet radiation due to a supernova. The fifth, and most famous, of the great extinctions took place 65 million years ago. It was caused by the impact of a large meteorite, followed by a series of large volcanic eruptions that caused the loss of 17% of living families, including the dinosaurs.

The database of the International Union for the Conservation of Nature (IUCN, www.iucnredlist.org) reports 850 species already extinct, most on land (583 species) or in fresh water (228 species), with just 16 marine species extinct. The number of species that the IUCN has classified as critical is 3,124, and another 4,564 species are in danger of extinction.

These estimates of the number of endangered species are conservative because only known species can be considered, and we only know about ten percent of the existing species. Moreover, in order for a species to be considered extinct, more than ten years has to have passed since the last time the organism was observed. So some species may well have been extinct for some years now, but have not yet been cataloged as such. Every year, the disappearance of close to 200 species is documented worldwide, although this number is thought to be much greater, if we include species that have disappeared before they were ever described. Some authorities, including biologist, E. O. Wilson, the father of conservation biology, consider that several tens of thousand of species grow extinct every year. That means that, by the end of the twenty-first century, between a third and half of the total number of species on the planet will have disappeared. Unquestionably, we are experiencing—and causing—a grave crisis of biodiversity (Eldredge 1998). The extinction of species due to human activity is not, however, a recent phenomenon. Fossil evidence offers abundant information about many species, especially large mammals and birds, that became extinct following the arrival of humans, especially in America and Australia, as well as the extinction of fauna in the Pleistocene due to hunting.

The transformation of land is one of the leading causes of extinction, as it constitutes an enormous loss of habitat that has led to the extinction of many species. The loss of wetlands, in particular, has had a devastating effect on numerous species of trees, plants, birds, fish, amphibians, and insects living there. Many of the contemporary extinctions affect species in island settings, where the processes of speciation have been particularly important, leading to a high number of endimisms that are always more vulnerable to human action. The human introduction of species that behave as invaders has also led to a significant loss of species. Thus, the introduction of foxes and cats to the Australian continent decimated small marsupials, many of which are now extinct. Others are gravely endangered. Invading species affect local biodiversity, displacing indigenous species. Their aggressive behavior is frequently attributable to the absence of predators or parasite in the areas to which they have been newly introduced. Human activity has introduced, for example, over 2,000 plant species to the US and Australia and some 800 in Europe (Vitousek et al. 2003). In some cases, the invading species can have positive effects on the ecosystem. Thus, or example, the zebra mussel that invades rivers and estuaries in Europe and North America can attenuate the effects of eutrophization on those ecosystems.

Human activity has also significantly affected marine diversity. Over-fishing has particularly reduced the biomass of fish in the ocean, which is a tenth of what it was at the beginning of the twentieth century (Millennium Assessment 2005). Growing pressure on coastal ecosystems is generating a biodiversity crisis of global dimensions, with a loss of habitats of great ecological value (coral reefs, wetlands, mangrove swamps, and undersea prairies), along with the biodiversity living there.

Available analyses indicate that a temperature increase of over 2ºC would cause extinctions of amphibians and corals and that an increase of more than 4ºC—which is within the predictions of climate scenarios for this century—could cause massive mortality that would affect one of every three species (Fischlin et al. 2007), making it comparable to the great extinctions of the past. A recent analysis (Mayhew et al. 2007) compared the rate of extinctions with the average rate of global temperature change, revealing the existence of a correlation between climate change and four of the five great extinctions of the past. This correlation reinforces predictions indicating that current climate change could cause a new massive extinction (Thomas 2004).

The synergic action of the different forces responsible for Global Change is the force that drives the notable erosion of biodiversity. For example, amphibians seem to be declining on a global scale for as yet unclear reasons that seem to have to do with a group of causes: loss of habitat, acid rain, environmental pollution, increasing ultraviolet radiation, and climate change. In fact, the current rate of species extinctions has reached sufficiently high levels for some researchers to postulate that we are already in the sixth great extinction.

The ecology and biology of conservation: the keys to our future

Awareness of the loss of species and ecosystems on scales reaching from local to global has sparked intense research activity over the last twenty years. Scientists seek to evaluate the consequences of extinctions, the role of biodiversity in the functioning of ecosystems, and the benefits biodiversity may have for society. At the same time, a greater knowledge of the biology of species has permitted improvements in the possibility of conserving them. During this period, ecology and the biology of conservation were born.

Large-scale experiments have shown that, in general, greater biodiversity corresponds with greater biological production, a more efficient recycling of nutrients, and a greater capacity by ecosystems to resist perturbations (Schwartz et al. 2000). The goods and services that ecosystems bring to society have been evaluated, including their added value (e.g. food supplies, water purification, regulation of atmospheric gases and of the climate, pollination, control of pathogens and their vectors, and so on), which is more than twice the combined gross national product of all nations (Costanza et al. 1988). The loss of those functions due to the deterioration of ecosystems and the loss of biodiversity would constitute a loss of natural capital with grave economic consequences, and a loss of our quality of life.

The Convention on Biological Diversity (www.cbd.int) signed by most nations—with notable exceptions—in Rio de Janeiro in 1992, is a reaction to this crisis of biodiversity. It is based on the recognition of the intrinsic value of biodiversity, its importance for the maintenance of life-support systems on which society depends, and evidence that biodiversity is being eroded by human activity. The Convention seeks to insure the conservation of biodiversity on the planet and a fair distribution of the wealth generated by its use. One of its objectives is to achieve a protected status for 10% of the Earth's surface. With this impetus, the number of protected areas has proliferated. On land, the objective is slowly drawing closer, but the ocean is still very far from the 10% goal.

Territorial protection is complemented with special measures to protect endangered species. Many are charismatic species whose conservation is energetically pursued with increasingly sophisticated and costly reproductive plans, including the consideration of advances in cloning techniques for their conservation. Cloning is a technique first developed through experimentation with amphibians, and it has been proposed as a possible contribution to the conservation of these species, that are in grave danger of extinction (Holt et al. 2004). A recent initiative was the inauguration on a Norwegian Arctic island of the Svalvard Global Seed Dome, a world bank that preserves seeds of agricultural interest from all over the world as protection against possible catastrophes (see: www.nordgen.org/sgsv/). Both this infrastructure and the risk it addresses were the stuff of apocalyptic science fiction until very recently.

The rate of extinctions and loss of ecosystems grows unstoppably, despite advances in the protection of natural areas and the conservation of specific species. It is increasingly clear that protected areas and efforts to protect individual species can only be understood as partial solutions in the face of impacts responsible for the loss of ecosystems and biodiversity—they must be completed with other strategies and techniques. It is necessary to better understand why species are being lost, the relations between different pressures that lead to their extinction, the possibilities of a domino effect in species extinctions (Rezende et al. 2007), and the relations between the deterioration of ecosystems and the loss of biodiversity. Without such understanding, it will be impossible to formulate more effective conservation strategies. Greater knowledge of the bases on which ecosystems resist pressures is essential to direct actions designed to reinforce that capacity to resist or, when the impact has already occurred, to catalyze and reinforce ecosystems' capacity to recover.

The promotion of scientific knowledge is essential to the generation of new conservation strategies, but it is not enough. The success of any strategy requires the reduction of pressure derived from human activity. Our society is behaving in a seriously irresponsible fashion, eroding and wearing down the natural capital base on which our quality of life, and the future of our species rest. Scientific knowledge must reach beyond the scientific community to inform society, contributing to the creation of better-informed and more responsible citizens. We must cross the frontiers of knowledge, and those that separate it from our society. Our future will be largely determined by the success or failure of our efforts.

Bibliography

Andersen, T., J. Carstensen, E. Hernández-García and C.M. Duarte. *Ecological Thresholds and Regime Shifts: Approaches to Identification. Trends In Ecology and the Environment.* 2008.

Bindoff, N. L., J. Willebrand, V. Artale, A, Cazenave, J. Gregory, S. Gulev, K. Hanawa, et al. "Observations: Oceanic Climate Change and Sea Level." In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H. L. Miller, eds., *Climate Change 2007: The Physical Science Basis.* Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK, and New York: Cambridge University Press, 2007.

Bouchet, P. "La magnitud de la biodiversidad marina." In C.M. Duarte, ed., *La exploración de la biodiversidad marina: Desafíos científicos y tecnológicos.* Madrid: Fundación BBVA, 2006, 32–64.

Corliss, J. B., J. Dymond, L. I. Gordon, J. M. Edmond, R. P. Von Herzen, R. Ballard, K. Green, et al. "Submarine thermal springs on the Galapagos Rift." *Science* 203, 1979, 1073–1083.

Costanza, R., R. D'arge, R. De Groot, S. Farber, M. Grasso, B. Hannon, K. Limburg, et al. "The value of the world's ecosystem services and natural capital." *Nature* 387, 1988, 253–260.

Crutzen, P. J., and E. F. Stoermer. "The 'Anthropocene.'" *Global Change Newsletter* 41, 2000, 12–13.

Darwin, C. *On the Origin of the Species by Natural Selection.* London: John Murray, 1859.

Duarte, C. M. "The future of seagrass meadows." *Environmental Conservation* 29, 2002, 192–206.

Eldredge, N. *Life in the Balance. Humanity and the Biodiversity Crisis.* Princeton: Princeton University Press, 1998.

Fiala, G., and K. O. Stetter. (1986). "Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C." *Archives of Microbiology* 145, 1998, 56–61.

Fischlin, A., G. F. Midgley, J. T. Price, R. Leemans, B. Gopal, C. Turley, M. D. A. Rounsevell, O. P. Dube, J. Tarazona, and A. A. Velichko. "Ecosystems, their properties, goods, and services." In M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. Van Der Linden and C .E. Hanson, eds., *Climate Change 2007: Impacts, Adaptation and Vulnerability.* Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press, 2007, 211–272.

Harris, L. A., C. M. Duarte, and S. W. Nixon. "Allometric laws and prediction in estuarine and coastal ecology." *Estuaries and Coasts* 29, 2006, 340-344.

Holt, W. V., A. R. Pickard, and R. S. Prather. "Wildlife conservation and reproductive cloning." *Reproduction* 127, 2004, 317–324.

Jannasch, H. W., and M. J. Mottl. "Geomicrobiology of deep-sea hydrothermal vents." *Science* 229, 1985, 717–725.

Jaume, D., and C. M. Duarte. "Aspectos generales de la biodiversidad en los ecosistemas marinos y terrestres." In C.M. Duarte, ed., *La exploración de la biodiversidad marina: Desafíos científicos y tecnológicos.* Madrid: Fundación BBVA, 2006, 17–30.

Karl, D. M., C. O. Wirsen, and H. W. Jannasch. "Deep-sea primary production at the Galapagos hydrothermal vents." *Science* 207, 1980, 1345–1347.

Lonsdale, P. "Clustering of suspension-feeding macrobenthos near abyssal hidrotermal vents at oceanic spreading centers." *Deep-Sea Research* 24, 1977, 857–863.

May, R. M. "Thresholds and breakpoints in ecosystems with multiplicity of stable states." *Nature* 269, 1977, 471–477.

Mayhew, P. J., G.B. Jenkins, and T.G. Benton. 2007. "A long-term association between global temperature and biodiversity, origination and extinction in the fossil record." *Philosop. Transc. of the Royal Society*10, 1098/rspb, 2007, 1302.

Meehl, G. A., T. F. Stocker, W. D. Collins, P. Friedlingstein, A. T. Gaye, J. M. Gregory, and A. Kitoh, et al. "Global Climate Projections." In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, eds., *Climate Change 2007: The Physical Science Basis.* Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK, and New York: Cambridge University Press, 2007.

Millennium Ecosystem Assessment 2005a. "Ecosystems & Human Well-Being: Wetlands and water Synthesis." Washington, DC: World Resources Institute, 2005.

Millennium Ecosystem Assessment 2005b. "Ecosystems & Human Well-Being: Volume 1." Island Press, 2005.

Rezende, E. L., J. E. Lavabre, P. R. Guimarães, P. Jordano, and J. Bascompte. "Non-random coextinctions in phylogenetically structured mutualistic Networks." *Nature* 448, 2007, 925–928.

Scheffer, M. and S. R. Carpenter. "Catastrophic regime shifts in ecosystems: linking theory to observation." *Trends in Ecology and Evolution* 18, 2003, 648–656.

Schwartz, M. W., C. A. Brigham, J. D. Hoeksema, K. G. Lyons, M. H. Mills, and P. J. Van Mantgem. "Linking biodiversity to ecosystem function: implications for conservation ecology." *Oecologia* 122, 2000, 297–305.

Strange, C. J. "Facing the brink without crossing it." *Bioscience* 57, 2007, 920–926.

Thomas, C. D. et al. "Extinction risk from climate change." *Nature* 427, 2004, 145–148.

Trenberth, K. E., P. D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker et al. 2007: "Observations: Surface and Atmospheric Climate Change." In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor And H. L. Miller, eds., *Climate Change 2007: The Physical Science Basis.* Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK, and New York: Cambridge University Press, 2007.

Vaquer-Sunyer, R., and C.M. Duarte. "Thresholds of hypoxia for marine biodiversity." *Proceedings of the National Academy of Sciences* 105, 2008, 15,452–15,457.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Ruchs, J. A. Eisen, D. Wu, et al. "Environmental genome shotgun sequencing of the Sargasso Sea." *Science* 304, 2004, 66–74.

mobility in a climate constrained world

JOHN B. HEYWOOD

The context and the challenge[1]

Mobility has always been valued. For most of history, it has meant moving people and goods at the speed a person could walk, a horse could move, an ox could pull a cart, or a boat propelled by sails or oars could travel through the water. Only in the nineteenth century when we harnessed the energy in fossil fuels were we able to move people and goods at a much faster pace. The invention of the petroleum-fueled motor vehicle at the end of the nineteenth century and the airplane at the beginning of the twentieth opened up opportunities for greatly increased speed and travel choice. Roads provided choice that railroads could not, and airplanes only needed runways on which to arrive and depart.

As a result of these innovations, the twentieth century became a "golden age" for mobility. The volume of personal travel and goods moved grew at unprecedented rates. By the end of the century, individuals who in earlier centuries would have spent their lives within a hundred kilometers of where they were born thought nothing of traveling to distant continents on business or for pleasure. Raw materials, manufactured goods, and food from half a world away became widely available. The world's various populations and geographic regions did not participate evenly in this twentieth-century expansion of mobility. At the start of the twenty-first century, the average citizen of one of the wealthier nations could act as though distances were almost irrelevant. But average citizens in many of the world's poorer countries still transported themselves and their goods as their ancestors did.

People everywhere desire ever-increasing mobility, both for its own sake and because it enables them to overcome the distances that separate their homes from the places where they work, shop, go to school, do business, visit friends and relatives, and explore different places. Businesses desire mobility because it helps them overcome the distances that separate them from their sources of raw materials, from their suppliers and their markets, and avoid the impacts of congestion. A growing concern, however, is that today's mobility systems rely on one source of energy—petroleum. And the tension between humankind's desire for mobility and its concerns about the negative impacts associated with exercising that mobility raises fundamental questions about its future.

During the latter half of the twentieth century the negative consequences of enhanced mobility became evident on a regional and even global scale. Pollution

[1] Source: WBCSD 2004.

produced by the internal combustion engines that powered hundreds of millions of motor vehicles began to degrade air quality in more and more cities. The exploration, extraction, transportation, and refining of oil to power transportation vehicles began to damage the environment on an increasing scale. Noise from vehicles on land and in the air, carrying people and goods, disturbed the peace of tens of millions of people. And it is now generally acknowledged that emissions of carbon dioxide from the burning of fossil fuels, a large share of which is transportation-related, is affecting the climate of our planet.

We are now being forced to question whether the extraordinary trends in mobility that have characterized the past fifty years are "sustainable." The World Business Council for Sustainable Development defines "sustainable mobility" as "the ability to meet the needs of society to move freely, gain access, communicate, trade, and establish relationships without sacrificing other essential human or ecological values today or in the future." (WBCSD 2001). With that definition, our current mobility trends are unsustainable.

Put simply, there are too many of us, we use far too much of our available resources, and we use it in ways that are irreversibly damaging our environment. There is too much consumption for our planet's health. And this high consumption is growing year by year due to population growth, increasing affluence, increasing urbanization and suburbanization, and ever-expanding expectations. Yet, mobility is almost universally acknowledged to be one of the most important elements in a desirable standard of living.

Most of us in the world's richer countries like our transportation systems, and much of the rest of the world aspires to have what we have. But people are increasingly aware that their enhanced mobility has come at a price. This price includes the financial outlay that mobility users must make to mobility providers to permit them to supply such systems and the services. But it goes well beyond this. Enhanced mobility has brought with it congestion, risk of death and serious injury, noise, disruption of communities and ecosystems, increased air and water pollution, and emission of climate-changing greenhouse gases.

The World Business Council for Sustainable Development carried out a major project, "Mobility 2030: Meeting the Challenges to Sustainability" (WBCSD 2004), which identified seven major goals where significant progress is needed to make transportation more sustainable:

1. Ensure that the emissions of transport-related conventional pollutants do not constitute a significant public health concern anywhere in the world.
2. Limit transport-related GHG emissions to sustainable levels.
3. Significantly reduce the total number of road vehicle-related deaths and serious injuries from current levels in both the developed and the developing worlds.
4. Reduce transport-related noise.
5. Mitigate congestion.
6. Narrow the "mobility opportunity divides" that inhibit the inhabitants of the poorest countries and members of economically and socially disadvantaged groups within nearly all countries from achieving better lives for themselves and their families.
7. Preserve and enhance mobility opportunities for the general population of both developed and developing-world countries.

This is an extremely demanding agenda. Our challenge is to make progress on individual pieces of this agenda, and at the same time track how well we are doing in the context of this broad set of goals. As we confront these challenges, it is useful to ask: What are the truly fundamental transportation "unsustainables"? A decade ago I was involved in a US National Academies study of this issue (NRC 1997), which concluded there were two such unsustainables. One was the climate change risk from CO_2 emissions, to which transportation is an important contributor. The other was the degradation of ecosystems and the reduction in biodiversity that result from transportation's emissions and infrastructure impacts. These are fundamental because ever-increasing mobility is inevitably making these two risk areas worse. They both link strongly to transportation's vast demand for energy. In this essay, I review how we can reduce transportation's energy consumption and greenhouse gas emissions, and thus its impact on climate change. This involves far more than just a focus on the WBCSD study's second goal of reducing GHG emissions. In parallel, we must pursue goals six and seven because enhanced mobility is essential to continued economic growth in all parts of the world. And progress must be made on the other goals if improving levels of mobility are to continue to be a major enabler for economic and social progress.

Size, growth, and complexity

Our transportation systems in the developed world move people by cars, buses, trains, and airplanes. In developing countries, bicycles and two and three wheelers are widely used, also. Freight is shipped primarily by road and rail, about equally by weight; air freight is growing rapidly. Large (heavy-duty) trucks dominate road freight transport. Transportation systems can be thought of as urban, national, or regional in scale. Figures 1 and 2, show the current status and future projections of key statistics by region and mode for personal transportation and freight. Steady growth that compounds year after year at rates of a few percent per year is evident. Currently, the developed and developing parts of the world are comparable in scale but growth rates in developing countries are higher. These projections (WBCSD 2004) are largely driven by growth in population and per capita income. By 2050 these measures of transportation activity are projected to be double what they are today.

These numbers indicate just how "big" transportation has become: the number of vehicles now in use, the mileage they travel, the weight of goods shipped. Currently, with some 6.8 billion people on the earth and 800 million vehicles, the average distance traveled is 5,000 km per year per person (with a range from 20,000 km/yr/person in the US to 3,000 in Africa). At present, the developed world countries dominate vehicle use but large parts of the developing world are catching up. Freight transport corresponds to 8 tonne-kilometers per person per day. Transportation fuel use is close to 3,500 liters or 1,000 gallons per person per year of which almost half is gasoline, one-third is diesel, and one-sixth jet fuel summing to two-thirds of total world petroleum production of about 82 million barrels per day (a barrel contains 42 US gallons). These consumption rates are so large they are unimaginable. Not only is the current scale vast but, growth rates of a couple of percent per year over several decades will make the scale even larger.

Why worry about the future, and especially about how the energy that drives our transportation might be affecting our environment? The reason is the size of these systems, their seemingly inexorable growth, and the environmental damage our transportation systems do. They use petroleum-based fuels (gasoline, diesel, and aviation fuel) on an unimaginable scale. When these fuels are burned inside engines, the carbon in these fuels is oxidized to the greenhouse gas carbon dioxide, and thus the amount of carbon dioxide entering the atmosphere from using these fuels is likewise immense. Transportation accounts for 25 percent of worldwide greenhouse gas emissions. As the countries in the developing world rapidly motorize, the increasing global demand for fuel will pose a

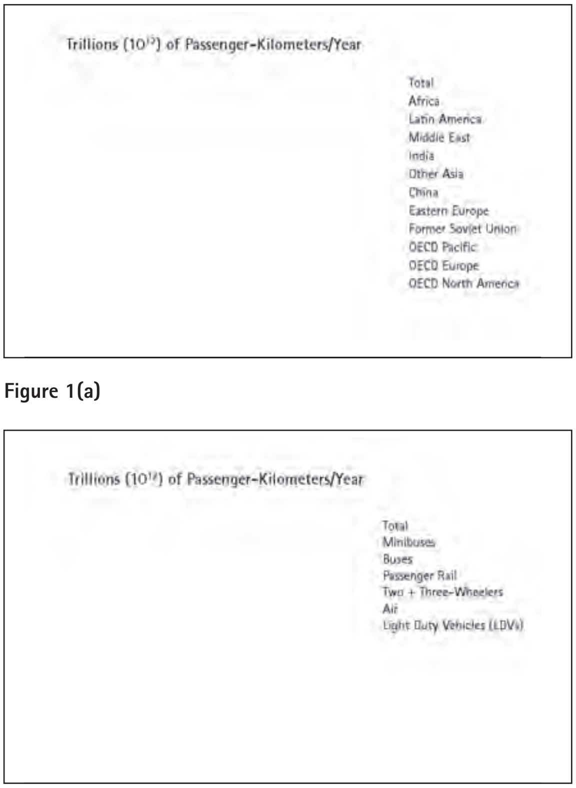


Figure 1(a)

Figure 1(b)

Figure 1. (a) Personal mobility activity by region; (b) Personal mobility activity by mode; out to 2050 (WBCSD 2004).

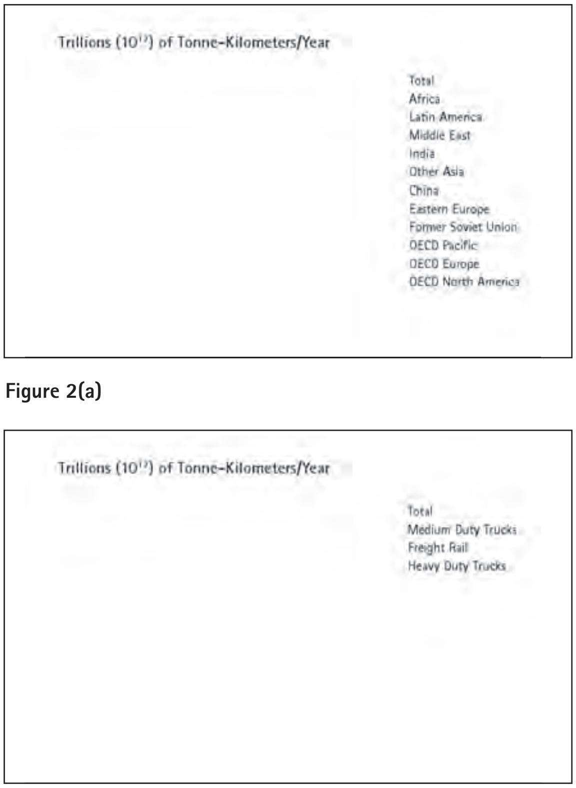


Figure 2(a)

Figure 2(b)

Figure 2. (a) Freight transport activity by region; (b) Freight transport activity by mode; out to 2050 (WBCSD 2004).

major fuel-supply challenge. It will also significantly increase the concentration of greenhouse gases in the atmosphere. The US light-duty vehicle fleet (automobiles, pickup trucks, SUVs, and vans) of some 250 million vehicles currently consumes close to 600 billion liters (150 billion gallons of gasoline) per year. If other nations burned gasoline at the same rate, world consumption would rise by a factor of almost 10.

Several countries have used fuel economy or CO_2 emission targets or standards as a strategy for reducing transportation's energy consumption and greenhouse gas emissions. Figure 3 shows proposed European Union and US fuel economy and GHG requirements. Since hydrocarbon fuels derived from petroleum are 87% carbon by weight, the fuel economy or fuel consumption, and the CO_2 emissions that result from burning that fuel, are linked in a straightforward way: burning 1kg of fuel releases 3.2kg of CO_2. Typically, the reductions in GHG emissions these targets or regulations will require are some 30% by 2020, just over 10 years away. In 25 years (by 2035) a factor of two reduction is thought to be plausible. Looking farther ahead to 2050, estimates indicate that at least a 70% reduction from today's GHG emissions levels would be required to reduce emissions sufficiently to keep CO_2 levels in the atmosphere below 550 ppm (IPCC 2007), a concentration viewed by many as the best we are likely to be able to achieve to hold down global warming. All of these targets, nearer-, mid- and longer-term, are extremely challenging because changes (whether through deployment of better technology or implementing effective conservation) *on this large a scale* takes significant effort, time and money.

Our options for change

As we look ahead, what opportunities do we have for making transportation much more sustainable, at an acceptable cost? Several options could make a substantial difference. We could improve or change vehicle technology to make it much more efficient; we could change how we use our vehicles so we consume less fuel; we could reduce the size and weight of our vehicles; we could use different fuels with lower GHG footprints. We will most likely have to do all of these to achieve the drastic reductions in transportation's energy consumption and greenhouse gas emissions now judged to be necessary.

In examining alternatives, we have to keep in mind these aspects of our existing transportation system. First, it is well suited to its primary context, providing mobility in the developed world. Over decades, it has had time to evolve so that it balances economic costs with users' needs and wants. Second, this vast optimized system relies completely on one very convenient source of energy—petroleum. And it has evolved technologies—internal-combustion engines on land and jet engines (gas turbines) for air—that well match vehicle-operating characteristics with this energy-dense liquid fuel. Finally, vehicles last a long time so changing impacts take a long time. Constraining and then reducing the local and global impacts of transportation energy use will take decades.

Let's look at the efficiency with which we use energy in our vehicles. Efficiency ratings can be misleading: what counts is the fuel consumed in actual driving. Figure 3 shows the energy flows in a typical mid-size car during urban driving. Only about 16% of the fuel energy actually drives the wheels: this overcomes the aerodynamic drag, the tire rolling resistance, and accelerates the vehicle. Vehicle fuel consumption can be improved by reducing losses in both the propulsion system and the rest of these vehicles. (Kasseris and Heywood 2007). Today's automobile gasoline engine is about 20 percent efficient in urban driving though it

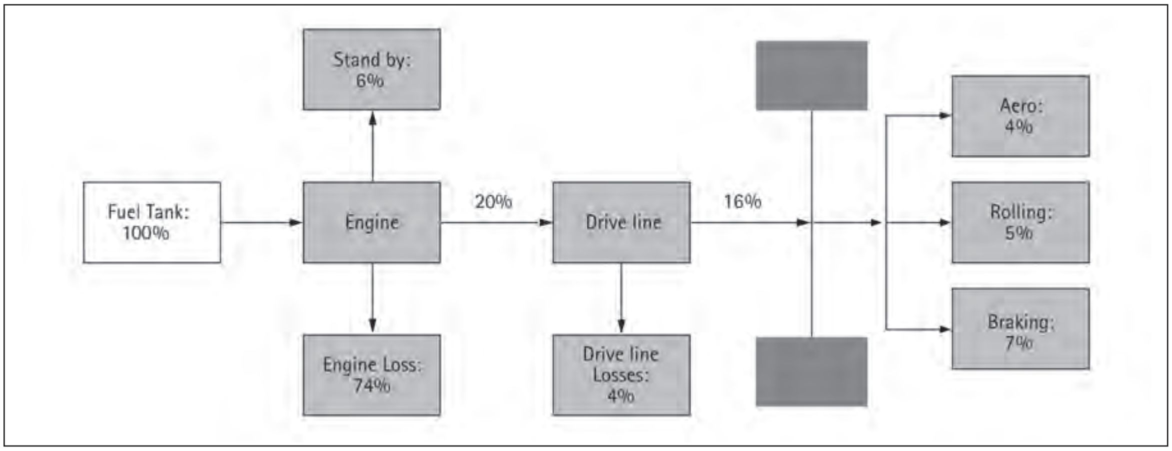


Figure 3. Energy flows from fuel tank to vehicle wheels in a typical current passenger car in urban driving (Bandivadekar et al. 2008).

is 35 percent efficient at its best operating point. But many short trips with a cold engine and transmission, amplified by cold weather impacts and aggressive driving, significantly worsen fuel consumption, as does substantial time spent with the engine idling. These real-world driving phenomena reduce the engine's average efficiency so that only about 10 percent of the chemical energy stored in the fuel actually drives the wheels. Amory Lovins, a strong advocate for much lighter, more efficient vehicles, has stated it this way: with a 10 percent efficient vehicle, and with a driver, a passenger and luggage—a payload of some 300 pounds, about 10 percent of the vehicle weight—"only 1 percent of the fuel energy in the vehicle's tank actually moves the payload!" (Lovins et al. 2005). Surely we can do better!

When we do our energy and GHG accounting as we use vehicles, we must include what it takes to produce the fuel from crude oil and distribute that fuel, to drive the vehicle through its lifetime of 100,000–150,000 miles (150,000–240,000 kilometers), and to manufacture, maintain and dispose of the vehicle. These three phases of vehicle operation are often called well-to-tank (which accounts for about 15 percent of the total lifetime energy use and greenhouse gas emissions), tank-to-wheels (75 percent), and cradle-to-grave (10 percent). We see that the energy required to produce the fuel and the vehicle is not negligible. This total life-cycle accounting becomes especially important as we consider fuels such as biofuels or hydrogen that do not come from petroleum, and new types of vehicle propulsion systems. It is what gets used and emitted in this total sense that matters.

We will explain shortly that improving existing light-duty vehicle technology can do a lot. By investing more money in increasing the efficiency of today's engines and transmissions, decreasing vehicle weight, improving the tires and reducing drag, we can bring down fuel consumption by about one-third over the next 20 years or so years—an improvement of some 3 percent per year. This reduction in fuel consumption would cost about $2,000 per vehicle: at likely future fuel prices, this amount would not increase the overall lifetime cost of ownership. Such incremental improvements have occurred steadily over the past 25 years, but we have purchased larger, heavier, faster cars and light trucks and thus have effectively traded the direct fuel consumption benefits we could have realized for these other attributes. Though most obvious in the US, this shift to larger more powerful vehicles has occurred and continues elsewhere as well.

What engine or propulsion system choices do we have? We can continue to use spark-ignition engines, fueled with gasoline or a biofuel such as ethanol. We can also use diesel engines, which are more efficient, using diesel or biodiesel. Hybrid electric vehicles (HEVs), with an internal combustion engine and a battery and electric motor, are another growing, more efficient option. During the next decade or so plug-in hybrid electric vehicles could become a viable option, using any of these liquid fuels along with electricity to recharge the batteries from the electric grid. Longer-term, after 2030, widespread use of fuel cell vehicles using hydrogen, and battery electric vehicles using electricity, are possible but might well be much more expensive. In addition vehicle weight reduction and reduced tire rolling resistance and drag are likely to occur and augment any propulsion system improvements. Note that a smaller, lighter, more efficient engine and transmission in a lighter-weight vehicle compounds the positive benefits of these improvements in especially advantageous ways.

Standard gasoline spark-ignition engines are continuing to improve their power per unit displaced volume, and their typical operating efficiency, by some 2% per year. Improvements come from friction reduction, variable control of the engine's valves, cylinder deactivation when the engine is lightly loaded, direct-injection of fuel into the engine's cylinder, increasing the engine's compression ratio, and more sophisticated sensing and control of engine operation. Increasingly the gasoline engine is being boosted by raising the intake air pressure with a turbocharger to increase engine output. This allows significant engine downsizing, which improves average engine efficiency. Diesel engines also have power and efficiency improvement potential, though not as great as that available to gasoline engines. Future diesels will need effective emissions treatment technology (traps and catalysts) in their exhausts to control the air pollutants particulates and NO_x. This will add significant cost and some fuel consumption penalty. Thus, future turbocharged gasoline and diesel engines will be much closer in how they operate, their power per unit engine size (or displaced volume) and their average driving efficiency. Importantly, this future gasoline engine will be significantly cheaper—about half the cost—of the competing diesel.

Hybrid electric vehicles (HEV) are now being produced and sold in volumes that are a few percent of the market. Current hybrids comprise a battery pack, electric motor, a generator, electric power controls, and a sophisticated transmission. Most current configurations use a parallel hybrid arrangement where the transmission can decouple either the engine or the motor from the wheels, and a control strategy that switches off the engine at idle and low loads, and recovers some 90% of the braking energy through regeneration. These "charge-sustaining hybrids" improve fuel consumption significantly, the magnitude of the improvement depending on type of driving (e.g., low-speed urban, or high-speed highway) and other key details. In the future, with improved hybrid technology, vehicle fuel consumption reductions relative to gasoline engine vehicles in the 40–50% range appear feasible. "Electric drive" augmented by an on-board internal combustion engine is inherently an attractive propulsion system for the future. It is, however, likely to be $2,000–$3,000 more expensive than its improved conventional counterpart. (Bandivadekar et al. 2008).

A plug-in hybrid vehicle (PHEV) is a hybrid-gasoline electric vehicle with a much larger battery that can be recharged from the electric grid. The vehicle would use an advanced battery pack (e.g., lithium-ion battery technology) in a configuration similar to that of the conventional hybrid. Above a threshold battery state-of-charge (SOC), the PHEV operates in "charge depleting" (CD) mode, where it uses the electrical energy in the onboard battery to meet the vehicle's power demands. When it reaches its minimum SOC threshold, the vehicle switches to "charge sustaining" mode, which is equivalent to vehicle operation in a conventional HEV. Both liquid fuel energy and electricity are used to drive the vehicle. Note that any electricity used on the vehicle consumes about three times as much primary energy when it is produced from fossil fuels. Plug-in hybrid technology is being developed, but at present is much too expensive for broad market appeal, and it will need "green" electricity if it is to provide significant additional greenhouse gas reduction potential beyond what charge-sustaining hybrids can provide.

The battery-electric vehicle sources all of its energy from off-board electricity and is charged from the electric grid. The BEV will require a trade-off between vehicle size, cost, and range. The typical 400–mile vehicle range of today's conventional multipurpose vehicles appears implausible in an all-electric vehicle from a cost and weight perspective; even a 200–mile range is daunting. BEVs do not seem a viable large market contender at present, though they are being pursued as a small city or urban car opportunity.

Fuel cells for vehicle propulsion applications employ the proton-exchange membrane (PEM) fuel-cell system to power an electric motor, which drives the vehicle, usually in a series configuration. A fuel cell operates like a battery in that it transforms chemical energy in the hydrogen fuel into electricity. Its key difference from a battery is that the fuel (hydrogen) and oxidizer (air) are supplied continuously to the cell's electrodes. In a fuel-cell hybrid configuration a battery, which stores electrical energy, improves the overall system performance and allows regenerative braking. This hybrid battery uses the same high-power lithium-ion battery now starting to be used for conventional hybrid vehicles. Fuel-cell vehicles must overcome a number of technological challenges and greatly reduce their cost before they can come to market in significant volumes. In particular, fuel cell performance and durability are limited by the properties of present-day electrolyte membrane materials, by catalyst requirements, and by the complex systems management needed for fuel-cell operation. In addition to this need for improved fuel-cell systems, developing an onboard hydrogen storage system that offers adequate vehicle range, is a major cost, size and weight problem. Of course, producing and distributing hydrogen—creating the hydrogen infrastructure—is a major challenge, too.

Transmissions also matter greatly. Automatic transmissions are popular in the United States primarily due to their ease of use and smooth gear shift, and their sales volume is growing elsewhere. Transmission efficiency is likely to improve in the near to mid term by increasing the number of gears as well as by reduction of losses in bearings, gears, sealing elements, and hydraulic system. While four speed transmissions dominate the current US market, five-speed transmissions are becoming standard. Six-speed automatic as well as automated manual transmissions are already present in some cars and are likely to become standard over the next decade. Luxury vehicles have started deploying seven and eight speed transmissions, which could become standard in the mid-term. Future transmission efficiencies of 90–95% are anticipated. This is a significant improvement over the previous generation of transmissions.

Vehicle weight reduction is another obvious way to improve fuel consumption. Figure 4 shows the dependence of fuel consumption on vehicle weight in the US light-duty vehicle fleet. A commonly used rule of thumb is that a 10% reduction in vehicle weight can reduce fuel consumption by 5-7%, when accompanied by appropriate engine downsizing, at constant performance. Weight reduction in vehicles can be achieved by substituting lighter-weight materials, and by vehicle redesign and downsizing. Downsizing a passenger car by one vehicle size-class can reduce vehicle weight by approximately 10%. However, vehicle size is an attribute that consumers value.

Tire rolling resistance reduction is also a fuel consumption reduction opportunity, and could reduce consumption by a few percent. Note that new tire technologies can be introduced into the much larger market of replacement tires and thus achieve

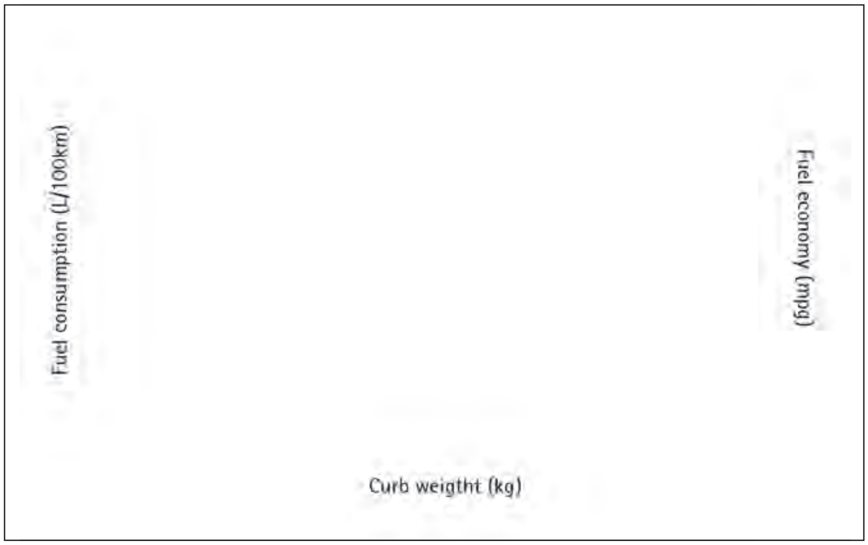


Figure 4. Light-duty vehicle fuel consumption as a function of vehicle weight for US model year 2005 vehicles (Cheah et al. 2008).

benefits faster than if implemented in new cars alone. Tire pressure monitoring and proper inflation levels are useful fuel efficiency opportunities also.

In highway driving, at least half of the energy required to propel the vehicle is used to overcome the aerodynamic drag. Thus, reductions in aerodynamic drag can achieve meaningful reductions in fuel consumption. A 10% reduction in drag can achieve up to a 2% reduction in average vehicle fuel consumption. Significant reductions in drag from current levels are feasible through vehicle steamlining.

GHG emission reductions from gasoline and diesel ICE vehicles and HEVs are proportional to reductions in petroleum consumed. Further reductions in GHG emissions could be achieved if the effective carbon content of fuels can be lowered through the use of low carbon-emitting biofuels. For PHEVs, BEVs and FCVs, the well-to-tank emissions produced during the generation and supply of electricity and hydrogen strongly affect the ultimate GHG emission reduction potential. Electricity production efficiency improvements, as well as increased contributions from nuclear, renewable sources, and fossil fuels with carbon capture and sequestration, could lower the well-to-tank emissions from electricity generation: plug-in hybrids would then be an attractive option for reducing both petroleum and GHG emissions.

Let's explore our fuel options in more detail. Our current transportation systems—land, water, and air—overwhelmingly use petroleum-based hydrocarbon fuels. These fuels dominate because they are liquids, have very high energy density, and fit well with today's engine technologies: spark-ignition engines, diesels, and gas turbines. An illustration of their attractiveness is that when refueling our cars today, fuel energy flows through the nozzle we hold in our hand at the rate of 10 MW providing another 400 miles of driving with a 5 minute refueling time.

Since petroleum-based fuels dominate the transportation sector, they have developed very large-scale refining and distribution systems. More than 300 billion gallons of refinery products are distributed across the US each year, some one-third of world production. The ability of alternative fuel streams to be compatible with and integrated into these refining and distribution systems is obviously a critical aspect of their attractiveness.

What are the possible alternatives? Natural gas use in transportation varies from less than 1% of vehicles almost everywhere, to about 10% in a couple of countries where tax policies have made it an economical option. Natural gas has attractive engine combustion characteristics but it is a gaseous fuel that must be compressed and then stored in high-pressure tanks on the vehicle. The drawbacks of a gaseous fuel (lower specific engine power, reduced driving range, compression work in vehicle fueling, vehicle interior space impacts of fuel storage tanks, extra cost, methane emissions) more than offset the attraction of the lower carbon-to-hydrogen ratio of this fuel. Furthermore, demand for natural gas in other applications is rising rapidly, as is its cost. As a widely used vehicle fuel, it prospects do not seem promising.

Oil sands (e.g., tar-like deposits in Canada) and heavy oils (more dense oils from Venezuela) are already contributing a growing fraction (about 5%) to liquid transportation fuels. Over time, other non-petroleum sources of hydrocarbon fuels, such as natural gas conversion to a liquid, oil shale, and coal, are likely developments. These pathways can produce high-quality transportation fuels, and volumes from such sources are expected to steadily increase. However, the carbon dioxide emissions during the production of these fuels are higher than those from petroleum-based fuel production due to the significant energy required to make them, and their higher carbon content.

Liquid transportation fuels derived from biomass have the potential to contribute significantly to supplying energy for our vehicles. Sources of biomass include corn, prairie grasses, switchgrass, miscanthus, forest, and municipal wastes, and other dedicated fuel crops. End products include ethanol, biodiesel, and, potentially, gasoline- and diesel-like fuels. Critical questions that need to be resolved are the availability of suitable land for these crops, the greenhouse gas releases that occur as land uses change, fertilizer and water requirements, land degradation over time, water pollution issues, and the net energy requirements during production. There is substantial potential for an important biofuel contribution to transportation but the extent of that opportunity still needs extensive evaluation. In the US maybe some 20% of transportation's energy could come from biofuels in about 20 years time.

Biofuels, electricity, and hydrogen, require a different type of life cycle analysis in transportation since the fuel production and distribution cycle is now the dominant component. Biofuel impacts vary from being comparable to the full life-cycle GHG emissions of gasoline-fueled vehicles for corn grain ethanol, to better than gasoline-fueled vehicles (sugar cane ethanol in Brazil), to potentially significantly better when based on cellulosic biomass conversion. Electricity's energy and GHG burdens vary substantially since they depend on how the electricity is generated. When generated from fossil fuels, electricity's burden is substantial, and the much more efficient use of electricity on the vehicle is essentially offset by the inefficiencies in electricity generation at the power plant and in distribution. Important questions are: what are the plausible sources of green—low GHG emissions—electricity for transportation with, say, plug-in hybrids, and when would such green electricity become available? Hydrogen faces similar questions: how could it be made and distributed with low GHG emissions? Any hydrogen produced in the nearer-term, would most likely come from natural gas, and overall has energy consumption and GHG emissions levels that are not much different from those that would result from using petroleum-fueled vehicles.

Performance of these vehicle technologies

We have projected the performance and costs of these various propulsion system and vehicle technologies out some 25 years. These projections for the mainstream powertrain vehicles are shown in figure 5. (Bandivadekar et al. 2008) Substantially better fuel consumption (at constant vehicle performance, and size) is anticipated, but the costs increase. The vehicle weight reduction (20% in these vehicles) costs some $700. Note that in Europe and Asia where average vehicle size and weight is some two-thirds that in the US, the weight reduction potential may well be less. Also, in Europe, about half of the passenger vehicle fleet is diesel so the average fleet fuel efficiency is already higher.

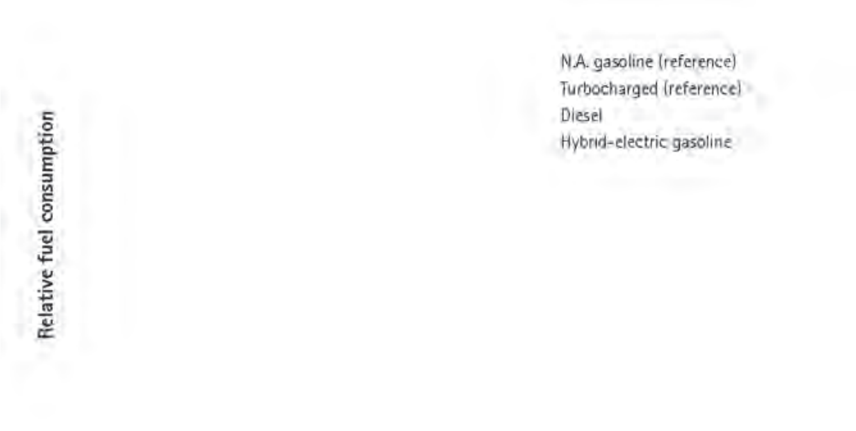


Figure 5. Relative fuel consumption of present and future vehicles with different advanced powertrains for 2006, 2020, and 2035 (Bandivadekar et al. 2008).

Overall, these improved future vehicles with these different powertrain options would cost some $2,000 more for a gasoline engine vehicle, $2,500–3,000 for a turbo-gasoline vehicle, $3,500–4,300 for a diesel and $4,500–5,500 for a hybrid, all relative to current mainstream gasoline-engine equivalents. Plug-in hybrids and fuel cell vehicles would probably cost $6,000–8,000 more; battery electric vehicles $10,000–20,000 more, depending on range. At present vehicle concepts with battery systems with significant on-board electrical storage capacity are clearly not yet ready for the mass market.

For the mainstream technology vehicles, the vehicle's lifetime fuel savings that go with these improved-fuel-consumption propulsion systems, when appropriately discounted, would offset these increases in vehicle cost at current fuel prices. But to date, this positive overall economic outcome has only pulled lower cost technology improvements into vehicles. It has not as yet created a strongly growing market for new and significantly more efficient technologies such as hybrids, though high fuel prices and lower diesel fuel taxes than gasoline, along with better diesel vehicle drivability, have pulled in the diesel to close to 50% of the new vehicle market in Europe.

It is then important to complete the full life cycle analysis by including the energy consumption and GHG emissions of the fuel cycle and vehicle production cycle. The vehicle production cycle currently adds about 10% to the energy and GHG emissions burden. Some 25 years ahead, this will rise to between 15–20% due to increasing use of new and lighter-weight materials that are more energy intensive, and through reductions in vehicle use fuel consumption. The fuel production and distribution cycle with petroleum fuels adds about 20%; hydrocarbon fuels from non-conventional petroleum sources like oil sands are likely to add about twice that.

Figure 6 shows a comparison of the vehicle petroleum consumption and well-to-wheels GHG emissions of these various future propulsion systems, in a mid-size lighter-weight US car 25 years ahead. On the petroleum consumption scale, plug-in hybrids,

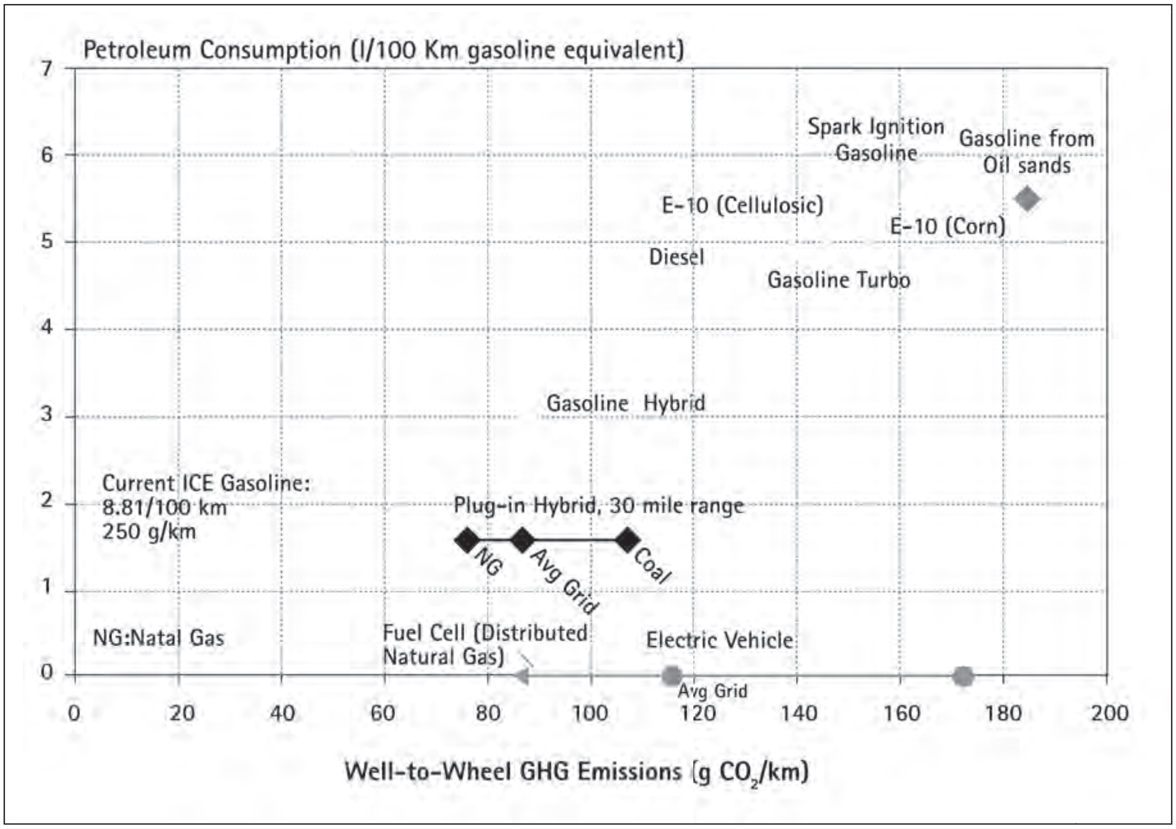


Figure 6. Petroleum consumption and well-to-wheels greenhouse gas emissions of future (2035) cars with various different propulsion systems and fuel sources (Bandivadekar et al. 2008).

fuel cells, and electric vehicles give significantly lower fuel consumption. But of course these vehicles also use electricity from the grid or hydrogen from a new production and distribution infrastructure. These additional energy requirements on the energy supply side reduce the GHG emissions reductions. With electrical supply systems, which are based on coal and natural gas (as in the US), the additional vehicle GHG emissions benefits are offset by the electrical generation emissions. With nuclear and renewable electricity generation the picture is much better. Battery electric vehicle emissions for a standard size vehicle are significantly worse in large part due to the added weight of the substantial battery pack required for adequate driving range. We see that future GHG emissions per vehicle could eventually be reduced to about one-third of the emissions from a current vehicle. Improvements in mainstream engines, transmissions, and reductions in vehicle weight and drag, decrease petroleum consumption by some 30–40%. Plausible quantities of biofuels could provide an additional 10% benefit. Hybrid technology provides a significant additional 40% reduction from these mainstream vehicle levels. While plug-in hybrids and fuel cells with hydrogen significantly reduce or remove petroleum consumption, in any build-up transition phase, their GHG emissions impacts are no better than conventional gasoline charge-sustaining hybrid levels.

Trade-offs and marketability

So far, we have compared vehicle characteristics at constant vehicle performance or acceleration, and fixed interior size: i.e., as we project into the future these vehicle characteristics do not change. Data from the past two decades show that vehicle performance and size have steadily increased, especially performance. In the US, while engines and transmissions have become increasingly more efficient over the past 20 or so years, on-the-road vehicle fuel consumption has remained constant. In Europe, the performance escalation has not been as great, and about half of the engine efficiency improvements have shown up as vehicle fuel consumption reductions. The emphasis placed on reducing actual fuel consumption is critical. In the US, such emphasis has been close to zero since the early 1980s, while about half the potential fuel consumption benefits have been realized in Europe. Vehicle purchasers and users have shown a clear preference for increasing vehicle performance and larger vehicle size, thus providing market "pull" for these attributes. The automobile companies compete amongst themselves by offering ever-increasing performance and size, providing the "push." In the US, the emphasis on enhanced performance has been so strong that, along with some size increases, a fuel consumption gain at constant performance of some 25% has been lost. In Europe, emphasis on performance has not been as strong, and half of the fuel consumption improvements that could have been realized have actually been achieved.

We have indicated that vehicle weight and size reduction could also contribute significantly to reduced petroleum consumption and greenhouse gas emissions. This is an important opportunity, and it adds to what powertrain improvements can do. Direct weight reductions through substitution of lighter materials and basic vehicle design changes (which, for example maximize the interior volume for a given vehicle length and width) enable secondary weight reductions as vehicle components are appropriately downsized. A shift in vehicle size distribution away form larger vehicles also reduces average weight and initially can be accomplished by changes in production volumes with existing models. Our estimates indicate that a 20% reduction in sales-weighted average vehicle weight could be achieved over about 25 years at a cost of about $700. The maximum potential for weight reduction in the US is about 35%, but the additional reduction beyond 20% would cost significantly more. These are substantial weight reductions and will require rethinking vehicle design. Vehicle weight reductions of 20–35% on their own result in some 12–20 reduction in vehicle fuel consumption. (Bandivadekar et al. 2008)

Market penetration: a critical issue

Improved propulsion system and vehicle technologies only have impact when vehicles with these technologies are being used in large numbers. Such improved technologies must therefore have strong market appeal, production volumes must be built up and high production volumes be sustained for 5–10 years to impact a large fraction of the in-use fleet. In-use fleet models when given market penetration rates of the different promising propulsion systems with their different fuel consumption and GHG emissions characteristics, can then examine how the overall fleet fuel consumption and GHG emission rates evolve as new vehicles enter the fleet and older cars are scrapped. The assumptions that are critical but difficult to estimate are the market penetration rates or evolving production volumes of these improved and new vehicle technologies.

What governs the rate of deployment of improved powertrain and vehicle technologies and of alternative fuels into the market? Even if the demand for an emerging vehicle or propulsion system technology is strong, the supply of such systems could be limited. This could be due to constraints in engineering and capital resources, as well as in supply chains. The automobile is a highly complex product, and consumer expectations from a mass-produced vehicle are demanding. The development and design of new powertrains and other sub-systems, often in vehicle architecture, is a major time and resource consuming task, and may take some 15 years to become available across all market segments.

Automobile manufacturing is both a capital- and labor-intensive business, and the established industry players are risk averse. It normally takes two to three years for an auto manufacturer to build a new production facility. Thus, to convert 10% of the US domestic production capacity (1.5 million vehicles per year) to produce hybrids would take a capital investment of approximately $2.2 billion, some 10% of the annual capital expenditure of the US motor vehicle manufacturing sector.

Implementation Stage	Vehicle Technology				
	Gasoline Direct Injection Turbocharged	High Speed Diesel with Particulate Trap, NOx Catalyst	Gasoline Engine/ Battery-Motor Hybrid	Gasoline Engine/ Battery-Motor Plug-In Hybrid	Fuel Cell Hybrid with onboard Hydrogen Storage
Market competitive vehicle	~ 2-3 years	~ 3 years	~ 3 years	~ 8-10 years	~ 12-15 years
Penetration across new vehicle production	~ 10 years	~ 15 years	~ 15 years	~ 15 years	~ 20-25 years
Major fleet penetration	~ 10 years	~ 10-15 years	~ 10-15 years	~ 15 years	~ 20 years
Total time required	**~ 20 years**	**~ 25 years**	**25-30 years**	**~ 30-35 years**	**~ 50 years**

Table 1. Estimated time scales for technology impact (adapted from Schafer et al. 2006)

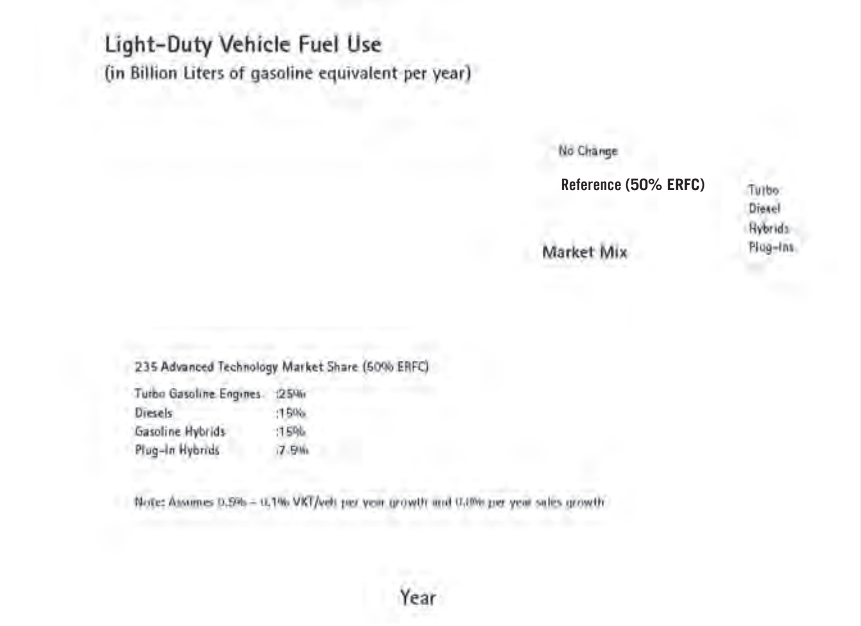


Figure 7. Illustrative scenario that explores the in-use US vehicle fleet fuel consumption impact of increasing numbers of advanced technology vehicles out to 2035. Half the efficiency improvements (50% ERFC) go to reducing vehicle fuel consumption; half offsets performance and size increases. Vehicle weight reduction, 10%. No change shows the impact in fleet size and mileage (Bandivadekar et al. 2008).

As these supply side constraints suggest, the time scales required for new technologies to have a significant impact on fleet fuel use are long. Schafer et al. (2006) split this total time into three stages, as shown in Table 1.

In the first stage, a market-competitive technology needs to be developed. For a technology to be market competitive, it must be available across a range of vehicle categories at a low enough cost premium to enable the technology to become mainstream. Table 1 shows estimates of how long it would take for these different propulsion systems to become available as mainstream alternatives in the market. Of these, only turbocharged gasoline, diesel, and gasoline hybrid powertrain technologies are currently ready to be designed for production. While no concrete product plans have been announced for plug-in hybrid vehicles, several major auto manufacturers have expressed interest in developing a commercial product within the next decade. The situation for a market competitive fuel cell vehicle is more speculative. A survey of announcements from major automakers suggests that a commercial mass-market fuel cell vehicle is at least ten years away.

In the second stage of technology implementation shown in the table, penetration across new vehicle production, represents the time scale for the vehicle technology to attain a market share of some one-third of the total vehicle sales. Broadly, this time scale reflects expectations about large-scale viability of these propulsion systems based on engineering and cost constraints.

The third stage of technology implementation is the build-up in actual use of substantial numbers of these vehicles. A meaningful reduction in fleet fuel use only occurs when a large number of more fuel-efficient vehicles are being driven around. This will happen over a time scale comparable to the median lifetime of vehicles, which is some 15 years.

Overall, we see that the total time scales before significant impacts occur from these new vehicle technologies, are long.

Real world impacts

Figure 7 shows an example of the impact of growing volumes of more efficient vehicles in the US light-duty fleet. This illustrative scenario assumes that production volumes of turbocharged gasoline engine vehicles, diesels, hybrids, and plug-in hybrids all grow steadily from current market shares to the percentages given in the figure by 2035. Many other assumptions are involved, of course. (Bandivadekar et al. 2008) The "no change in technology" line shows the fleet's gasoline consumption rising steadily due to growth in fleet size and mileage. With the "emphasis on reducing fuel consumption" (shown as ERFC) at 50%, half the efficiency improvements are realized as actual fuel consumption reductions. The growing thin wedges for each of the more efficient engine/propulsion system technologies are clear. Overall, this scenario reduces the fuel consumption from 765 to 594 billion liters per year, a 22% reduction. The two inputs that have the greatest effect on the scenario impact calculation are this emphasis on reducing fuel consumption and the percentage of the most efficient technology—hybrids—in the 2035 sales mix. With full, 100%, emphasis on reducing fuel consumption rather than 50%, fleet fuel consumption is reduced by an additional 15% to 505 billion liters per year. If the sales volume of hybrids doubles—i.e., if some 50% of the 2035 new vehicles are hybrids—an additional 10% reduction in fleet fuel consumption to 543 billion liters could be achieved. Combined, these two additions would give an additional 30% reduction relative to the market mix line in figure 7. Note that the impact of these more efficient technology vehicles grows slowly at first, but beyond about 2030 plays an increasingly more substantial role in reducing fleet fuel consumption and GHG emissions as the technology improves and deployment grows.

What we learn from these scenarios is that the inexorable impacts of growth in vehicle-miles traveled can be offset, and fuel consumption and GHG emissions can be leveled off and then pulled downwards.

But it will take a couple of decades to do this. Actions that directly affect the full in-use vehicle fleet, like reduced driving due to higher fuel prices, can impact emissions faster than new vehicle technology. And, as we have explained, focusing strongly on *reducing on-the-road fuel consumption* rather than allowing vehicle performance and size to escalate is really important.

As an illustrative example of what is needed to get onto this path, we have analyzed what it would take to halve the fuel consumption, or double the fuel economy, of the new car sales fleet in 2035. (Cheah et al. 2008) It would require that two-thirds of new vehicle production be hybrids, require 75% of the energy efficiency improvements to go into actual fuel consumption reduction instead of increased performance and size (in the US this percent has been zero; in Europe it has been around 50%), and would require on average a 20% vehicle weight reduction. While feasible, this is a challenging, demanding, and time-consuming task.

We might expect our 2020 targets of a one-third reduction (e.g., the US CAFE fuel economy requirements) to be less challenging, but that is not the case. With the target date only some 10 years away, the fuel consumption improvements in the various powertrain technologies available will be correspondingly less, and the time available to build-up significant production levels of these technologies is less too. Thus, the 2020 task turns out to be at least as demanding than this factor of two improvement in 25 years.

Looking much further ahead to 2050, we are learning that more of the same types of incremental improvements will not get us to where we may well need to be—GHG emissions down to some 20–30% of what they are today. That is where plug-in hybrids or fuel cell technologies may well have to come into large-scale use with their different sources of energy—electricity or hydrogen. If these "energy carriers" are produced with really low greenhouse gas emissions, then these energy carriers and the technologies that use them would bring substantial additional GHG emissions reductions. But the time scales for such radical changes in technology and energy carriers to have significant impact are long, several decades, as Table 1 suggests. And success in developing battery and fuel cell technology, and low GHG emitting energy production and distribution infrastructures, is far from certain. Major efforts to explore these options are in progress, as they should be.

Other options

We do have other options. For decades, transportation system improvement opportunities have been studied and some have been implemented. But many have not because of the difficulty in coordinating businesses, local, regional, and national governments, as well as blending new ideas with existing infrastructures.

In passenger transport, the opportunities could be significant because the current pattern of largely single-occupant vehicle usage is inefficient in terms of energy and money. Many studies demonstrate the potential for reducing energy consumption and emissions through use of what we call Intelligent Transportation Systems (ITS) that electronically link vehicles to one another and to the infrastructure, and thereby reduce vehicle use by densifying and reorganizing land-use patterns, enhancing collective modes of travel, substituting information and communication technologies for travel, and enhancing use of non-motorized travel modes. The opportunities are many and compelling, with various co-benefits, such as less travel delay, reduced road infrastructure investment, and less local air pollution. The energy and climate benefits that would result could be significant in the longer term.

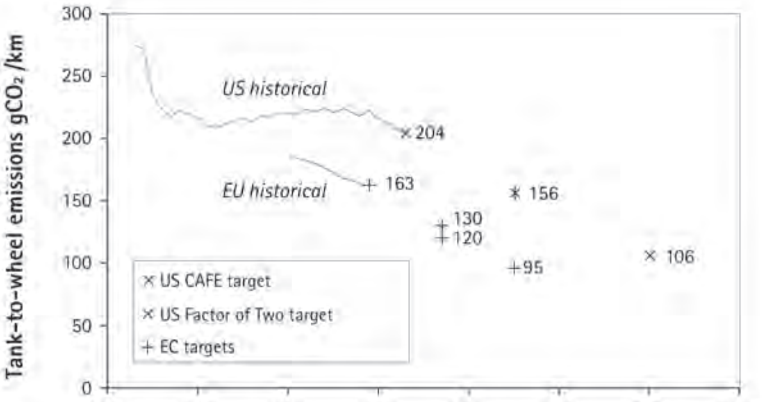
ITS is based on technologies that can sense individual vehicles and their characteristics, such as speed and location within the transportation network. Various technologies are available for doing this: devices that can sense vehicles using specialized roadside infrastructure, the Global Positioning System, or the cellular telephone network. For this information to be of value to more than the drivers of individual vehicles, it must be communicated from the vehicle to the infrastructure or between vehicles to enable the gathering of data about the overall status of the network. Collecting these massive amounts of data and reducing them to a form in which they can be used either to provide information to individual drivers or to manage the transportation network as a whole requires sensing, communicating, computing, analysis, and feedback.

Work on ITS is motivated by the following shortfalls in the performance of our current surface transportation system. Congestion, which reflects insufficient capacity on our highways is a major issue that affects the movement of both travelers and goods. The costs—both financial and environmental—of addressing capacity issues by building or expanding traditional infrastructure can be prohibitive, especially in urban areas where land-use constraints apply. ITS-based concepts can help combat congestion in two ways, through use of traffic information, and dynamic road-use pricing. Highway safety is also a major concern.

In the context of this essay our major issue is the energy and environmental impact of transportation. By smoothing traffic flow, fuel consumption, and emissions per vehicle are reduced. However, if infrastructure use increases as a result of increased capacity, then fuel use and emissions will increase. It is, therefore, important to explore how ITS can be used to improve the transportation system in a synergistic way with the other issues listed above (Sussman 2005).

To achieve a significant reduction in vehicle use and GHG emissions requires a mix of complementary strategies. The changes discussed earlier concerning the efficiency of the vehicle, and these changes in the transportation system to reduce congestion, often assume that current residential and travel patterns will continue. If land use patterns change towards more dense urban occupancy from suburban and rural housing patterns, then vehicle miles would be reduced. If people are willing to live along denser urban corridors, they could be served more efficiently by public transportation. And a significant role for small "city cars" might develop. However, land use patterns for more than 60 years have been moving in the other direction—toward suburban living.

For more diversified transportation systems to evolve, two sets of policy changes would be needed: greater control of land use and greater use of road pricing. These opportunities need to be considered. The net effect of a concerted effort to internalize congestion and environmental externalities, reduce single-occupant vehicle use, and encourage the use of small, efficient neighborhood or city cars for local travel could be large.



Notes:
- For US, unadjusted NHTSA fuel economy numbers used
- US CAFE target = 35 mpg by 2020; US factor of target = Halve 2006 FC by 2035
- Conversion factor = 8,788 gCO2 per gallon gasoline (EPA 2005)
- For EU, historical data for EU-15 passenger cars shown (European Commission 2006)
- Lower EC target in 2012 (120 gCO2/km) includes use of biofuels

Figure 8. Average new vehicle greenhouse gas emissions targets and regulations in the US and Europe.

What can we expect?

Figure 8 shows greenhouse gas emissions targets for light-duty vehicles for Europe and the US. These targets are aggressive in that the time scales proposed for achievement are short, and thus they require faster progress than our factor of two reduction in 25 years. In Europe, achieving these objectives in the near term will be especially difficult because there is less "slack" in the existing in-use vehicle fleet: that fleet is already half diesel (a more efficient engine than the gasoline engine), and average vehicle size and weight are some two-thirds of the US levels. Also, performance escalation in Europe and Japan, which has occurred, has been significantly lower than in the US opportunities for further improvement are, therefore, less.

Figure 9 illustrates our challenge. It shows global and US projections out to 2050 for GHG emissions from light-duty vehicles. Today the US fleet emits about half the global emissions from this mobility sector. Europe and Japan contribute a significant fraction of the rest. Even with aggressive implementation of more efficient technology, emissions are usefully but only modestly reduced from current levels.

In summary we see the potential for a 30 to 50 percent reduction in the fuel consumption of light-duty vehicles—cars, SUVs, light trucks—over the next 10 to 25 years. This will come from improving mainstream powertrains, developing new more efficient propulsion systems, and reducing vehicle weight and size. Whether or not we achieve this potential will depend on how we control ever-increasing vehicle performance expectations, and especially how urgently we pursue these technology improvements. The latter will depend on the context for such actions: the price of petroleum, our sense of its availability, our GHG emissions concerns, the comprehensiveness and effectiveness of the policies and regulations we impose to force such change, and our willingness to moderate our demand for mobility. In the nearer-term, 10 to 20 years, it is clear what path we should be on to reduce transportation's petroleum consumption. Mid-term, as we focus increasingly on GHG emissions reduction, the path becomes less clear and the challenge of continuing to reduce petroleum use and GHG emissions increases. We will then have to resolve the questions of how much of a contribution we can obtain from biofuels, the extent to which we can use electricity as a major energy source for our vehicles, and whether or not hydrogen and fuel cells for vehicle propulsion should be our ultimate objective. We are not yet able to resolve these questions, but we need to develop the knowledge base so that as time evolves we develop ever better answers.

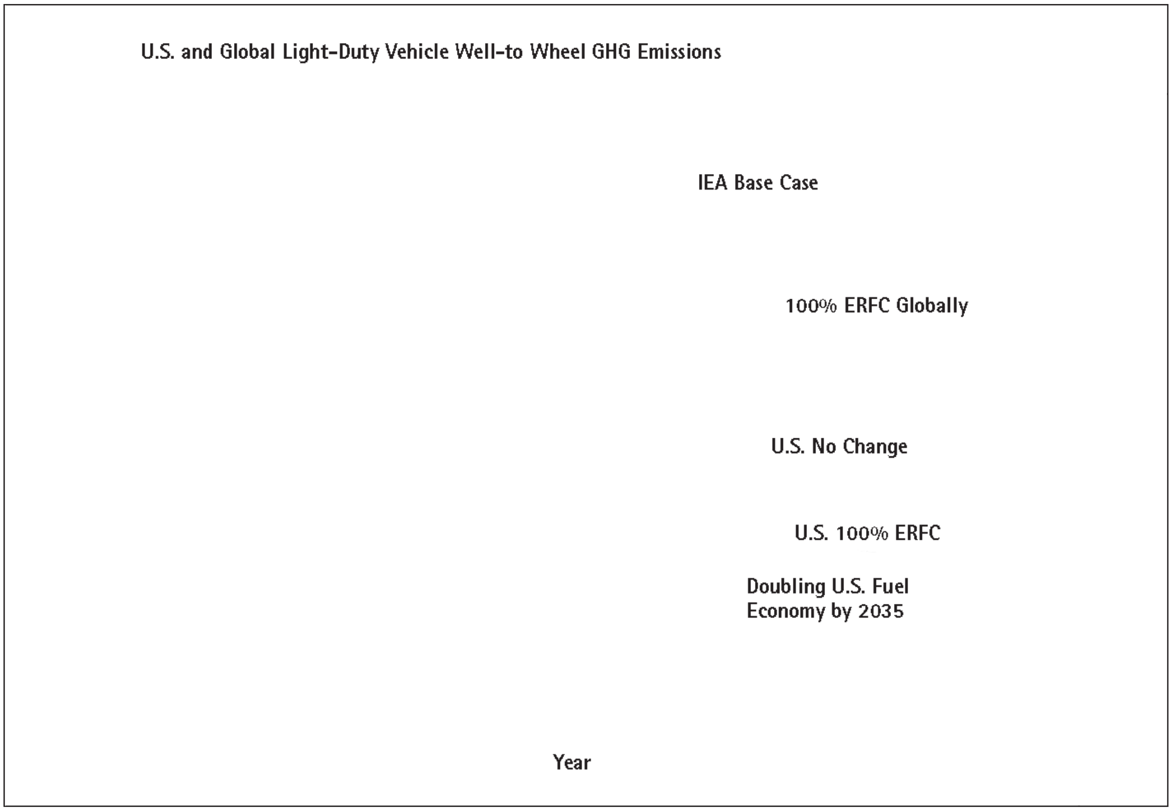


Figure 9. Global and US scenarios out to 2050 showing effects of halving new vehicle fuel consumption by 2035 on in-use vehicle fleet well-to-wheels greenhouse gas emissions. Average vehicle performance and size are held constant (100% emphasis on reducing fuel consumption) (Bandivadekar et al. 2008).

Worldwide demand for transportation services is growing inexorably, and there is no single major development that alone can resolve the growing problems of vehicle fuel consumption and GHG emissions. This essay has explained that progress must come from a comprehensive effort to develop and market more efficient vehicles and more environmentally benign fuels, find more sustainable ways to satisfy demands for transportation services, and prompt all of us who use our vehicles and other transportation options to reduce our travel-related energy consumption. All of these changes will need to be implemented at very large scale to achieve significant reductions in transportation's petroleum and energy use, and GHG emissions. Implementation is likely to increase the cost of transportation to ultimate users, and will require government policies to encourage, even require, moving toward these goals while sharing the burdens more equitably and minimizing total social costs.

Transitioning from our current situation onto a path with declining fuel consumption and emissions, even in the developed world, will take several decades—longer than we hope or realize. We must keep in mind that what matters is effecting changes that will have substantial impact on these issues. We will need much better technology, more appropriate types of vehicles, greener fuel streams, and changes in our behavior that emphasize conservation. We need nearer-term results that get us out of our currently worsening situation. We will need to change to far more sustainable pathways in the longer term. And we will need to pursue all these opportunities with urgency and determination.

Bibliography

Bandivadekar, A., K. Bodek, L. Cheah, C. Evans, T. Groode, J. Heywood, E. Kasseris, M. Kromer, and M. Weiss. "On the Road in 2035: Reducing Transportation's Petroleum Consumption and GHG Emissions." MIT Laboratory for Energy and the Environment, Massachusetts Institute of Technology, July 2008. http://web.mit.edu/sloan-auto-lab/research/beforeh2/otr2035/.

Cheah, L., A. Bandivadekar, K. Bodek, E. Kasseris, and J. Heywood. "The Trade-off between Automobile Acceleration Performance, Weight, and Fuel Consumption." *Society of Automotive Engineers paper 2008-01-1524.* SAE International Powertrains, Fuels and Lubricants Congress, Shanghai, China, June 23–25, 2008.

Intergovernmental Panel On Climate Change. "Climate Change 2007: Synthesis Report." *Summary for Policy Makers*. Cambridge, UK: Cambridge University Press, 2007.

Kasseris, E., and J. B. Heywood. "Comparative Analysis of Automotive Powertrain Choices for the Next 25 Years." *Society of Automotive Engineers, SAE paper 2007-01-1605, 2007*. National Research Council.

—, "Toward a Sustainable Future: Addressing the Long-Term Effects of Motor Vehicle Transportation on Climate and Ecology". *Special Report* 251, Committee for a Study on Transportation and a Sustainable Environment, Transportation Research Board, National Academy Press, Washington, DC, 1997.

Schafer, A., J. Heywood, and M. Weiss. "Future Fuel Cell and Internal Combustion Engine Automobile Technologies: A 25-Year Life Cycle and Fleet Impact Assessment." *Energy*, vol. 31, no. 12, 2006, 2064–2087.

Sussman, J. *Perspectives on Intelligent Transportation Systems (ITS).* New York: Springer, 2005.

World Business Council For Sustainable Development. *Mobility 2001: World Mobility at the End of the Twentieth Century and its Sustainability.* Report prepared for the Sustainable Mobility Working Group, WBCSD, by the Massachusetts Institute of Technology and Charles River Associates, August 2001. Available online at http://www.wbcsdmobility.org.

World Business Council For Sustainable Development. *Mobility 2030: Meeting the Challenges to Sustainability.* World Business Council on Sustainable Development. Sustainable Mobility Project Report, 2004 (180 pages). Available online at http://www.wbcsd.org/DocRoot/fI311MAvneJpUcnLgSeN/mobility-full/pdf.

current challenges in energy

CAYETANO LÓPEZ

Introduction

Growing worldwide demand for energy, and problems of scarcity and environmental impact associated with conventional sources are at the base of a very probable energy crisis in the next two or three decades. Petroleum will become increasingly expensive and scarce, while the climatic effects of massive use of all fossil fuels will by then be clearly felt. At the same time, current nuclear installations will have reached the end of their useful life. And it is not clear, especially in Europe, whether the power they will no longer provide when shut down, will be supplied by new nuclear plants.

At the present time, we cannot abandon any existing energy sources. They must receive the necessary modifications to eliminate or reduce their environmental impact, and new sources must be added, especially renewable ones. Below, I will describe the state of available technologies and the most promising developments in each of them, always on a time scale of the next few decades.

On a longer scale, nuclear fusion will be part of a catalog of more sustainable energy sources, but it will not be ready in the time period under consideration here and will thus be unable to help in resolving the crisis. That is why I will not address nuclear fusion here, although a powerful and interesting program is being developed on an international scale. The goal is to harness the reactions of nuclear fusion as an energy source, but foreseeable progress places it outside the time span we have chosen for the present analysis of energy problems.

Energy and civilization

Energy is a fundamental ingredient in human life. There is no industrial, agricultural, health, domestic, or any other sort of process that doesn't require a degree of external energy. Human beings ingest around 2,500 kilocalories of energy per day as food. But in industrialized countries, the average daily amount of supplementary (exosomatic) energy consumed in combined human activities (industrial, domestic, transportation, and others) is equivalent to 125,000 kilocalories per person. That is fifty times more, and in the case of the United States, the figure is one hundred times more (see, for example, British Petroleum 2008). In fact, there is a strong correlation between individual energy consumption and prosperity among different societies.

In figure 1, each country is represented in a diagram in which the "Y" axis specifies the Human Development

Index (HDI) for that country as determined by the UN using basic data on the wellbeing of its inhabitants. The "X" axis shows annual energy use per capita (in this case, in the form of electricity). Two interesting phenomena are observable here. In the poorest countries, the correlation is very strong, with energy consumption leading to clear improvements in the HDI. But in more developed countries, the very large differences in energy consumption do not significantly affect levels of wellbeing. This indicates that, for the latter countries, energy saving is a possible and desirable policy. In the most prosperous countries, saving is actually the cleanest and most abundant energy source. On the other hand, the necessary economic and social development of the comparatively poor countries that make up the greater part of the world's population will inevitably require greater energy consumption, so it is unrealistic to think that global energy use could diminish in the future. If it did, it would be an absolute catastrophe for the least-developed countries, which lack everything, including energy. Therefore, while energy saving must be a central aspect of active polices in first-world countries, from a global perspective, we must deal with the problem of a growing demand for energy.

Current energy sources

The primary energy sources are identified and it seems unlikely that any will be added in the foreseeable future. From the dawn of humanity to the beginning of the Industrial Revolution in the early nineteenth century, the only available sources of primary energy were wood and other forms of natural biomass, beasts of burden, and wind for maritime or river traffic. With the development of the first steam engines, coal entered use as an energy source and it continues to be an important source of consumed primary energy today. Later, with the widespread use of automobiles with internal combustion engines calling for liquid fuels, petroleum and its by-products became the preeminent source of energy. Finally, over the last half century, natural gas has become an important component in the generation of electricity and the production of heat for industrial and domestic uses.

These fuels—coal, petroleum, and natural gas—are found at different depths in the Earth's crust. They were formed in earlier geological epochs by natural processes in which organic materials—mainly plants and marine organisms—were subjected to high pressure and temperatures. That is why they are known as fossil fuels. Their contribution to the sum of primary energy consumed worldwide at the end of 2007 (British Petroleum 2008) was 35.6% for petroleum, 28.6% for coal and 23.8% for natural gas. Together, they thus represent 88% of the total. As we will see below, there are many reasons why this cannot be sustained, even into the near future. The rest comes from nuclear energy, which provides 5.6% of the total, and renewable energies, mostly hydroelectric.

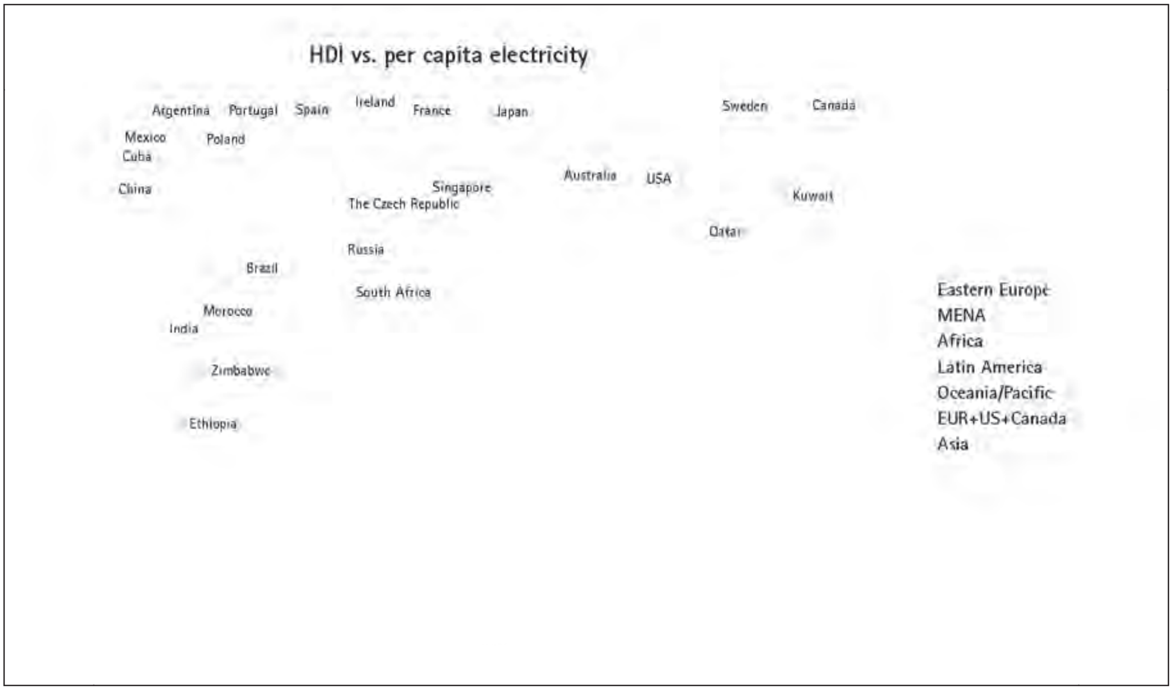


Figure 1. The Human Development Index (HDI) as a function of the amount of electrical energy consumed per person per year. Drawn up by this author on the basis of HDI data from the UN (UN 2006) and data on electricity use by the International Energy Association (IAE 2008).

Energy drawn from wind and the Sun in various ways is a marginal factor from a global perspective, but it is beginning to have a greater presence in some countries, especially Spain. So that is the global perspective; there are no more available sources of primary energy.

Of all this primary energy, an important part is transformed into electricity (about 40% in a country like Spain), while the rest goes to the transportation sector and other industrial and domestic uses.

Fossil fuels

The enormous predominance of fossil fuels as a primary energy source has some important consequences:

First, they are unequally distributed. Two thirds of the known reserves of petroleum, probably the most difficult fuel to replace, are under five or six countries in the Middle East, which implies a degree of dependence that is not especially compatible with a stable supply. Natural gas is also very concentrated in that area, and in the countries of the former USSR, while coal is more evenly distributed in all parts of the planet.

Second, these are non-renewable raw materials. They were formed over the course of dozens or even hundreds of millions of years and are thus irreplaceable. Moreover, they are limited resources. In particular, the use of petroleum as an energy source on which the lifestyle of industrialized nations is based, could be just a brief fluctuation in the history of humanity, limited to a period of about two centuries.

Third, these raw materials are scarce. There is some debate about the amount of available petroleum, but most geologists and petroleum experts agree that, at the current rate of consumption—no less than 85 million barrels of petroleum a day, which means burning a thousand barrels of petroleum per second—we only have enough for a few decades. It can be argued that the amount of petroleum extracted depends on the price and that, if it rises, there will be no practical limit to production. But this argument overlooks the fact that it takes more and more energy (in prospecting, pumping, treatment, and logistics) to extract petroleum from deposits that are increasingly deep or depleted. In the mid twentieth century, the energy required to extract a barrel of petroleum was equivalent to around 1% of the contents of that barrel. Today, that cost has risen to between 10% and 15%. When the energy needed to extract a barrel of crude oil comes close to the energy that same barrel could produce, no matter what its price, then it will have disappeared as a primary energy source, although it may continue to be useful, especially in the petrochemical industry, where it is used to synthesize a multitude of compounds that are fundamental to almost all branches of industry and agriculture.

At the current rate of consumption, proven petroleum reserves will last about 40 more years, while those of natural gas will last around 60 years. Coal reserves will last approximately a century and a half (British Petroleum 2008). There will be new discoveries, of course, and there are also the so-called non-conventional petroleums drawn from hydrocarbons dispersed in sand, bituminous schists, or heavy tars, but we must always remember the growing energy cost, and thus, their decreasing net yield and higher price. At any rate, there will not be a sudden end to supplies, passing from the current levels of use to nothing. There will probably be a progressive rise in price and, at some point, a progressive decrease in consumption and production as well.

Finally, we know that burning fossil fuels generates enormous amounts of atmospheric carbon dioxide (CO_2). This gas is one of those that produces the greenhouse effect and thus contributes to global warming. Given how fast this phenomenon is taking place (in geological terms), it could produce serious climatic disturbances that are potentially harmful for our civilization (not for life, as has frequently been alleged, nor for human life, but certainly for our complex and demanding social organization).

In sum, our social activity is based on fossil fuel use that, due to environmental concerns and limited supplies, must be limited in the future. Nevertheless, coal will continue to be a massive energy source for decades to come, but its use will only be tolerable if the contamination it produces can be palliated.

In consequence, the second energy challenge (the first is reducing consumption in developed countries) is to diminish the primacy of fossil fuels in energy production.

Preparing petroleum substitutes

Transportation depends almost entirely on liquid fuels derived from petroleum. Coal and natural gas are now important for electric production but they could conceivably be replaced by renewable or nuclear energy in the long term. However, it is not easy to imagine alternatives to the use of petroleum by-products for transportation. All of these involve very far-reaching changes.

The first possible alternative is the use of biofuels—bioethanol and biodiesel—to at least partially replace conventional fuels. But we have recently seen the collateral problems that can arise, especially in the area of food production, even when biofuel production is only just beginning. Of course, the influence of bioethanol production—the most controversial case—on food prices is limited and price rises coincide with other, deeper causes, some of which are momentary and others, structural. The only grain that is widely used to make ethanol is corn, while wheat and barley are employed in marginal amounts with regard to total production.

Rice is not used at all. And yet, prices have risen for all these grains, especially rice. Moreover, about half the current production of bioethanol comes from Brazilian sugarcane, and the price of sugar has not risen at all.

In any case, making ethanol from grains is the worst possible solution, not only because of its impact on food production, but mostly because of its poor energy yield. In fact, between fertilizers, seeds, harvesting, transportation, and treatment, the amount of energy contained in a liter of ethanol is barely more than that required to obtain it from cereals (see, for example: Worldwatch 2007; CIEMAT 2005). Therefore, from an energy standpoint, it is unreasonable to use this type of raw material. Environmental concerns associated with the use of water and tillable land also seem to discourage it (Zah 2007). On the other hand, the energy yield of sugar cane is much higher, and the yield of ethanol from what is called lignocellulosic biomass—present in woody or herbaceous plants and organic residues—is even higher. This is called second-generation ethanol. All of these conclusions appear in the interesting graph in figure 2, which is taken from Zah 2007. It offers all the data about fossil fuel consumption in the growing, harvesting, pretreatment, and other processes needed to obtain biofuels from different plant materials, as well as the overall environmental impact, compared to the direct use of petroleum by-products.

The third challenge, then, is to perfect the already existing technology to produce second-generation biofuels on a level useful to industry. This is not far off, and some pilot plants are already experimenting with various processes for generating ethanol from the sort of biomass that has no effect on food, requires less energy cost, and has less environmental drawbacks (see, for example: Ballesteros 2007; Signes 2008).

Thus, cane ethanol and second-generation biofuels could diminish petroleum dependence in the transportation sector, although they could not entirely replace it, due to the limited amount of tillable land and available biomass compared to that sector's gigantic fuel consumption.

It is easier, at least in principle, to replace fossil fuels used to generate electricity—resorting to renewable or nuclear sources—than to find substitutes for every petroleum product. Thus, in the long run, I think we will turn to electric vehicles, first as hybrids and later purely electric. The problem here is how to store the electricity. The batteries used at present are inefficient and very contaminating, but intense research into new devices for storing electricity is currently under way and will allow the construction of electric vehicles with adequate performance.

In general, we should say that energy storage, be it electricity, heat, hydrogen, or any other form,

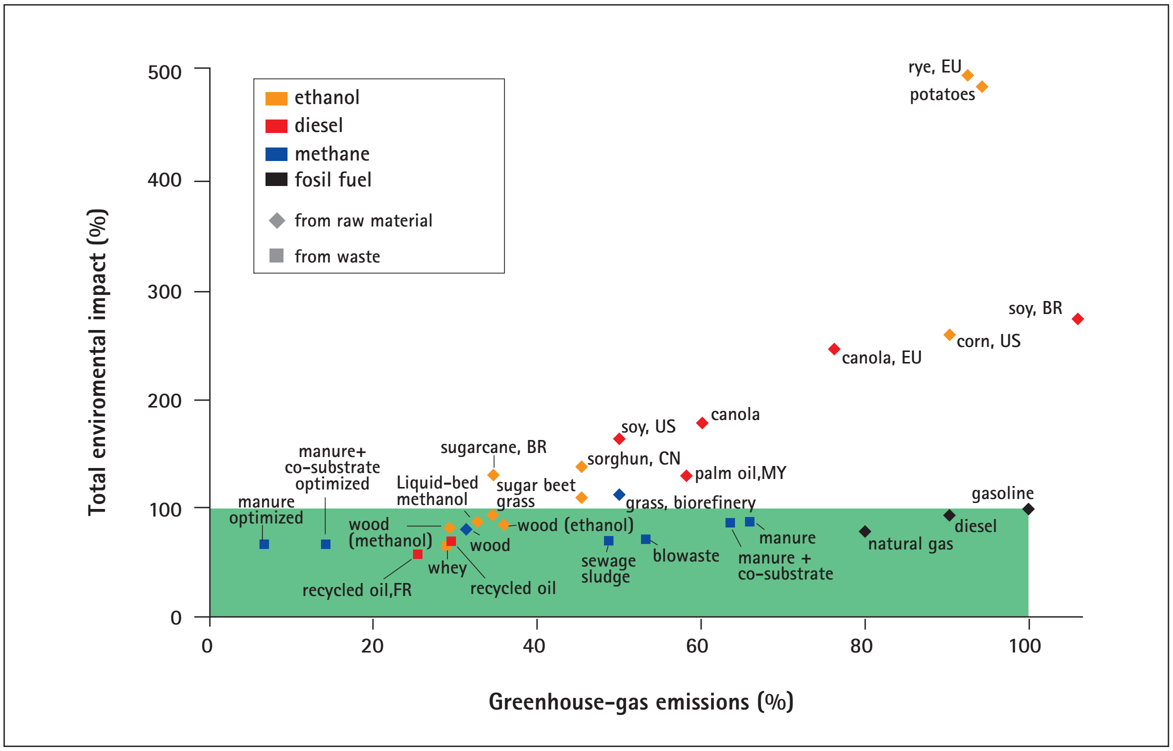


Figure 2. Environmental impact and consumption of fossil fuel in the production of biofuels as compared to the direct use of fuels derived from petroleum. (Zah 2007).

	Coal	Natural Gas	Nuclear	Hidraulic	Other Renewable	Others
World Average (06)	40%	20%	16%	16%	2%	6%
USA (06)	49%	20%	19%	7%	2%	3%
France (06)	4%	4%	78%	11%	1%	2%
China (04)	83%	0%	2%	15%	0%	0%
Spain (07)	24%	34%	18%	10%	11%	3%

Table 1. Percentages of total electricity generation from primary energy sources.

currently occupies a central position in energy research, both because of its importance to the future of the transportation industry and in order to solve problems derived from the intermittence of renewable sources, as we will see below. In other words, if we manage to improve electric storage technology (see, for example, José Luis Mata Vigi-Escalera, Club de la Energía 2008a), which is a formidable challenge if we want to reproduce the performance of a gasoline-based vehicle—then, an important portion of future vehicles will be electric. Therefore, below, I will concentrate on the production of electricity, which is shaping up to be the most flexible and adaptable energy, even for the future of the transportation industry.

Clean coal?

The electricity production scheme varies considerably from one country to another. In table 1, we offer some data about the relative makeup of energy sources used to generate electricity in Spain, some other countries, and the world average (IEA Statistics; Club Español de la Energía 2008a).

It can be seen that, with the exception of France, that relies very heavily on nuclear power, and partially Spain, which has an appreciable use of renewable sources, the basic energy source continues to be coal. And it will continue to be so for a long time, due to its abundance and its distribution on almost all continents. The case of China is particularly notable. According to the International Energy Association, in recent years, it has been opening a new coal-based electric plant every week. But coal is by far the most contaminating fossil fuel of all, spewing almost twice as much carbon dioxide into the atmosphere per energy unit produced as natural gas, and about 40% more than the gasoline used in internal combustion engines, not to mention its sulfur, nitrogen, and heavy metal components.

So, if we want to continue using coal as an energy source, we must develop procedures to eliminate or at least limit atmospheric CO_2 emissions (the other emissions are already controlled right in the power plants). This is known as Coal Cartridge Systems (CCS) and is still in its early stages. In particular, the capture of CO_2 emitted during coal combustion can be carried out with oxicombustion techniques that modify the composition of the air entering the boilers so that the gas emitted is almost entirely CO_2. That way, no separation is necessary. This can also be done by applying separation techniques to air-based combustion. Both methods generate additional energy costs and will need new physical-chemical processes, which have been tested in laboratories but not on the needed industrial scale. As to the CO_2 that is obtained as a result—we must find underground or underwater deposits hermetic enough that CO_2 injected into them will remain trapped there for centuries.

In reality, deposits of this type exist naturally. For example, deposits that have held natural gas for geological periods of time can be used to store carbon dioxide once the natural gas has been exploited. The same is true for exhausted petroleum deposits, sedimentary saline formations, and so on. In fact, most of the experiments with CO_2 storage around the world are associated with oil fields whose production is falling. The carbon dioxide is injected under pressure in order to improve production, obtaining crude oil that would not come out using conventional extraction techniques.

Another interesting experiment is being carried out at Sleipner, a gas production camp on the Norwegian coast of the North Sea. In that field, methane, the principal ingredient of natural gas, comes out mixed with significant amounts of CO_2. Once the two are separated in the extraction plant, the CO_2 is injected back into the seabed at a depth of about a thousand meters, in a bed of porous boulders with water and salts. They have been depositing CO_2 there since 1996, and data about how hermetic it is will be of great value when seeking new locations for massive use. At any rate, we should mention that the processes

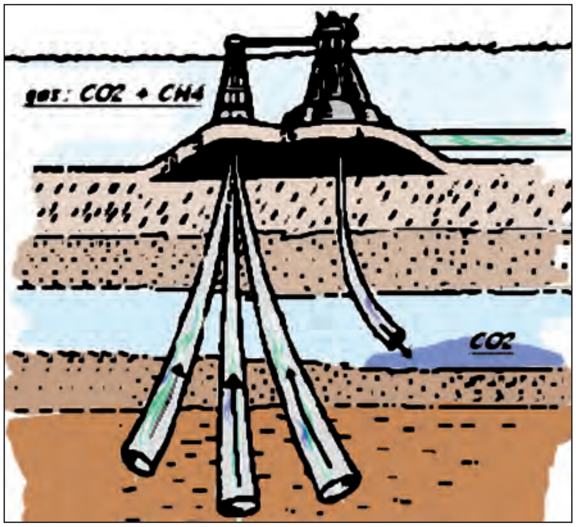


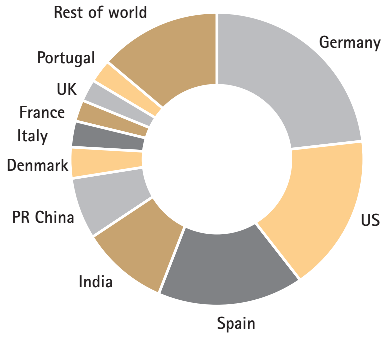
Figure 3. Sleipner Camp on the Norwegian coast of the North Sea.

of capturing and storing carbon dioxide will always signify additional costs, which must be added to the price of energy obtained from the clean use of carbon. Experts estimate that this cost will be between 30% and 100% of the cost associated with non-CCS coal use (Socolow 2005; Fundación para Estudios sobre la Energía 2008). Still, we must view this additional cost in the context of the rising price of both conventional and renewable energies, additional costs for CO_2 emissions, and aid for non-contaminating energy of the sort defined in Spain's Special Tax Code. The conclusion is that humanity will not stop using such an abundant and widespread energy source as coal, but its use has grave environmental consequences that it is extremely important to counteract with techniques such as CCS.

Renewable electricity. The wind

Perhaps the most important challenge for us in the next few decades will be significantly increasing the contribution of renewable energy compared to current levels, which are marginal on a planetary scale. Hydroelectric power has the greatest presence and its resources have been used in the most complete way, but other renewable energies, such as wind and solar power, have advantages and disadvantages. Their advantages are the opposite of the disadvantages to fossil fuels mentioned above—they are sustainable, unlimited, and hardly contaminate at all, even when we consider their complete lifecycle and their territorial distribution. Their disadvantages fall into two categories: high cost and intermittence.

TOP 10 TOTAL INSTALLED CAPACITY



	MW	%
Germany	22,247	23.7
US	16,818	17.9
Sapin	15,145	16.1
India	7,845	8.4
PR China	5,906	6.3
Denmark	3,125	3.3
Italy	2,726	2.9
France	2,454	2.6
UK	2,389	2.5
Portugal	2,150	2.3
Rest of world	13,060	13.9
Total top 10	80,805	86.1
Total	93,864	100.0

Figure 4. Installed wind capacity as of 31 December 2007.

One of the reasons why renewable electricity is so expensive is its degree of dispersion, which is an intrinsic characteristic offset only by its unlimited and sustainable character. However, it is reasonable to think that the expense of conventional energy will continue to increase as supplies diminish and environmental costs are figured in. In that case, its costs would converge with those of renewable energies at some point. The high expense of renewable energies is also due, in part, to the fact that the technology associated with it is still not very advanced. In that sense, the way to diminish costs derived from the lack of technological development is to create a worldwide market. That is the only way to lower the expense of producing the needed components, as it will lead to production of larger series, and to the emergence of more companies, ending the oligopolies currently existing in some key fields. Moreover, it will make it possible to implement improvements in the operation and maintenance of renewable energy plants, following a certain period of operating experience in conditions of industrial exploitation. Indeed, the different systems currently being activated to stimulate the spread of renewable energies are intended to broaden that market through the use of subsidies or other types of aid that compensate for initial difficulties.

As is well known, in Spain and some other countries that are advanced in this field, a special tax code has been enacted for renewable energies (and cogeneration), with the exception of hydroelectric power. This consists of a series of incentives or subsidies per kilowatt hour of renewable origin, intended to compensate for the greater current costs and thus stimulate growth in that sector. Special tariffs are different for each generating technology, reflecting the different costs at present, but they are supposed to diminish over time as costs decrease until they converge with conventional energies. This, and any of the other existing schemes, has already proved efficient in the first of the renewable energies that can be considered widespread on the worldwide market: wind power. In fact, at the end of 2007, the global figures for accumulated wind power were already 93,900 MW (Global Wind Energy Council 2008), which has made it possible to configure a dynamic industrial sector that is growing all over the world.

As can be seen in figure 4, the three countries with the greatest installed capacity are Germany, the United States and Spain, although, due to its lesser overall consumption, Spain is the one that obtains the greatest fraction of its electricity from that energy source—around 9%. In fact, Spain is second in the world, after Denmark, in both the total percentage of electricity from wind power, and the installed capacity per capita (European Wind Energy Association 2008).

With public support, the creation of a global wind-energy market is making headway, not only by creating new industrial activity and jobs, but also by progressively reducing the cost of energy thus produced. At the end of the nineteen seventies, when aerogenerators had a power potential of about 50 kW and a rotor diameter of about 15 meters, the unit price was about 20 to 40 euro cents per kWh. Now, aerogenerators have about 2 MW of power potential and a rotor diameter of nearly 100 meters, making the cost of energy production only slightly higher than conventional sources. The tariff of the special code for wind energy exceeds the conventional one by about 2.9 euro cents per kWh (about 2 cents per kWh in the United States).

Of course, they have gotten bigger, but there have also been many other technological improvements that affect their moving parts, the materials they are made of, their conversion, transformation and evacuation systems, and how they are manufactured and erected. The challenge in this field is achieve market expansion and technological improvements needed to bring the unit cost of electricity down to that of conventionally produced power. It is also a challenge to conquer the marine medium, with so-called off-shore wind power, where the wind itself is better (sustained winds without turbulence), although there are considerable difficulties involved in anchoring and maintaining aerogenerators when the water reaches a certain depth, as well as evacuating the electricity they produce.

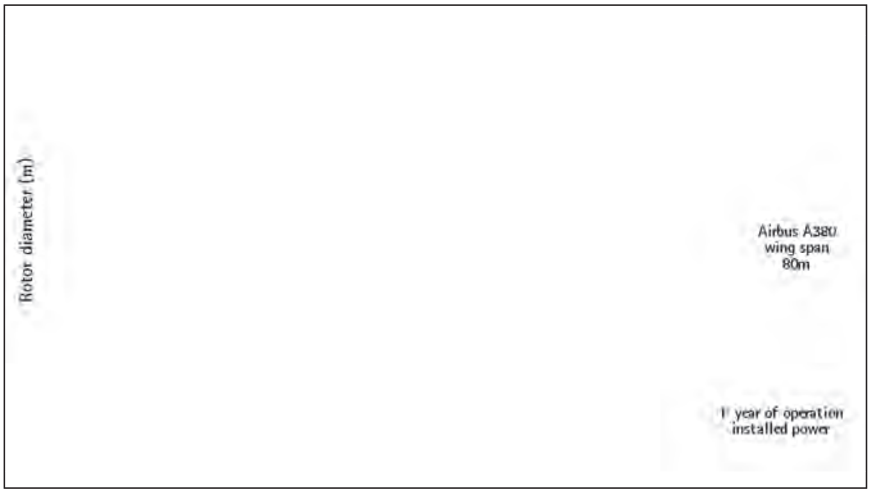


Figure 5. The progressive growth of aerogenerators. Power potential in MW and rotor diameter (twice the length of the blades) are indicated, along with the first year in which aerogenerators of each power level entered service. In Germany, there are now aerogenerators of up to 7 MW. To give an idea of their size, they are compared with an Airbus 380.

Thus, wind energy has a long way to go, both technologically and in terms of its territorial extension to other settings—the sea, of course, but also small-scale wind power, both in urban settings and in settlements that are outside the power network, or have a weak one. As happens with all renewable sources, the problem of intermittence has yet to be solved. Wind is discontinuous. In Spain, for example, wind parks only generate energy for an average of about 2000 hours a year, as can be seen in figure 6. That is something less than a quarter of the time.

Moreover, the time when electricity is being generated does not always coincide with periods of maximum demand. Nevertheless, in the windy month of March 2008, wind power counted for no less than 18.7% of the electricity generated in Spain that month, and for around 18 hours on 22 March, 9,900 MW of wind power was active, some 41% of the overall electricity being generated at that moment. And, during the entire weekend of 21–23 March, wind-powered electricity represented 28% of total production.

Solving the problem of intermittence calls for solving that of storage. The amounts of electricity we are dealing with here can be stored by pumping water into double reservoir dams, very few of which yet exist. Another system is to convert the electricity produced by aerogenerators into hydrogen that can later be converted back into electricity in a fuel cell, as needed. In fact, storing energy from renewable sources could be one of the applications for hydrogen as an energy vector. And, of course, if new devices are invented to store energy directly, such as a new generation of batteries, which we mentioned above in our discussion of transportation, then wind power could contribute to electric supplies in a manageable and still more significant way.

Renewable energy. The Sun

In terms of energy, solar radiation reaching the Earth's surface constitutes an average power of around one kW per square meter. If we average that out for all the hours of the year, in a sunny location like the south of Spain, it would add up to about 1,900 kWh per square meter per year. That is equivalent to the energy contained in 1.2 barrels of petroleum, or a coat of petroleum 20 centimeters deep. Given the enormous expanses of very sunny desert soil, as primary energy, sunshine on the Earth's surface is thousands of times greater than all the energy consumed worldwide.

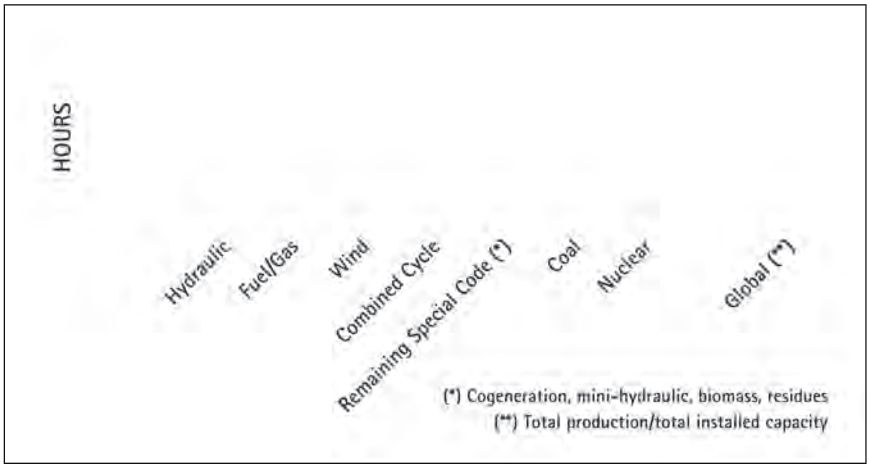


Figure 6. Functioning hours of electricity plants according to their source of primary energy in 2006 (Red Eléctrica Española).

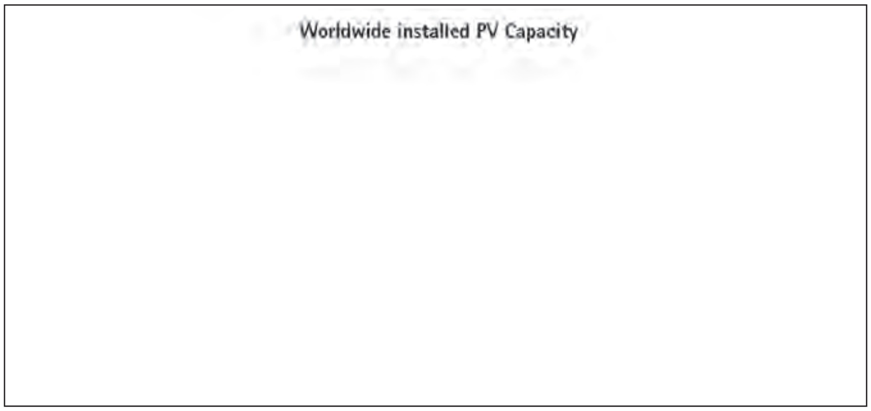


Figure 7. Installed photovoltaic capacity in the world (EPIA 2008).

There are two way of using solar energy. The first is to convert it directly into electricity using photovoltaic cells made from materials that transform the energy of the Sun's photons into electrons in a conductor. The second transforms radiant energy into high-temperature heat, which is then turned into electricity using a conventional turbine. That is known as thermoelectric solar energy.

Photovoltaic solar energy has the same drawbacks as the rest of the renewable energies: price and intermittence. The price comes from the cost of building photovoltaic cells, making it the most expensive of all renewable energies at the present time, requiring considerable public support. In fact, in systems based on special tariffs, photovoltaic energy is the one that receives the highest subsidies. On the other hand, photovoltaic technology is one of the most versatile and adaptable to urban settings due to its distributed character and the fact that it does not require large transformation systems, unlike thermoelectric devices. As to its diffusion, the total installed capacity generated by this method around the world is increasing at a dizzying rate in recent times, as can be seen in figure 7.

Germany is the country with the greatest installed capacity—3,800 MW—though Spain has undergone a very considerable expansion of photovoltaic installations over the last two years, with 630 MW at the end of 2007. That expansion, which would not be sustainable over time, is associated with a bonus of over 40 euro cents per kWh in the Special Tax Code, and the announcement that the amount of that bonus would be decreased in September 2008. Indeed, the level of the photovoltaic bonus is a good example of the importance of determining incentives in an intelligent manner. If they are too low in comparison to the real, foreseeable costs, they will not foster development of that technology, given that, as we saw above, the creation of a broad market is a necessary condition. But if the bonus is too high, it will not encourage technological advances needed to lower costs, which would in turn lower the amounts of bonus money associated with such costs.

Currently, most of the panels installed are composed of cells made from silicon, crystalline or polycrystalline wafers. The average yield of such devices in field conditions, that is, the fraction of solar energy deposited on the surface of the material that actually becomes electricity, is somewhere between 10% and 15%. There are other alternatives for improving that performance or for decreasing the cost of photovoltaic cells. One way is to explore other types of material and deposition techniques. These are known as thin-film systems, and they also use silicon—though less than conventional systems—or other more exotic and less abundant materials that improve photoelectric conversion. There are also multi-layer systems that allow the overlapping of materials sensitive to different frequencies of the solar spectrum, which increases total performance. There, the objectives are to find materials and cell-production procedures that use the smallest amount of materials, and to find materials that are cheap, do not contaminate, work well in different applications—in construction, for example—and seem best adapted to this kind of technology. Nevertheless, conventional solar panels based on silicon wafers are expected to predominate for many years to come.

Still, photovoltaic systems are expected to become more efficient quite soon, thanks to concentration techniques based on optical devices that direct solar radiation from a large area onto a much smaller photovoltaic surface, increasing its yield. At any rate, the fundamental goal of photovoltaic technology is to reduce costs, which are still very high in comparison to other renewable energies.

Another way of using solar radiation to produce electricity is thermoelectric technology. There, sunlight is concentrated on a receiver containing a fluid that heats up and then transfers that heat to a conventional turbine, generating electricity. This technology has been known for years, is straightforward and robust. And it has undergone considerable development in recent years, especially in Spain and the United States. Research into the shape of solar collectors and receivers has led to the design of a variety of devices, but here we will only consider the two most widespread technologies, cylindrical-parabolic collectors, and the tower or central receiver.

In the first case, the heat is concentrated in a tubular receiver containing fluid (normally a mineral oil with adequate thermal properties) that reaches a temperature of 400°C, then passes through a heat exchange, generating high-temperature, high-pressure vapor that drives a turbine. In the nineteen eighties, following the second major petroleum crisis, a group of plants (the SECS complex) was built in California's Mojave desert, with a total power potential of 350 MW. It continues to work today, with no problem at all, furnishing not only electricity, but also valuable information about how such technology works. After they were put into use, and the crisis was over, no more were built. Meanwhile, in the same period, the Almería Solar Platform (PSA) was built. It is now part of the Center for Energy, Environmental and Technological Research (CIEMAT), a world-class laboratory that researches all kinds of thermoelectric technologies, trains personnel and tests all sorts of components and devices. The existence of the PSA is one of the factors that explains our country's leading role in this field.

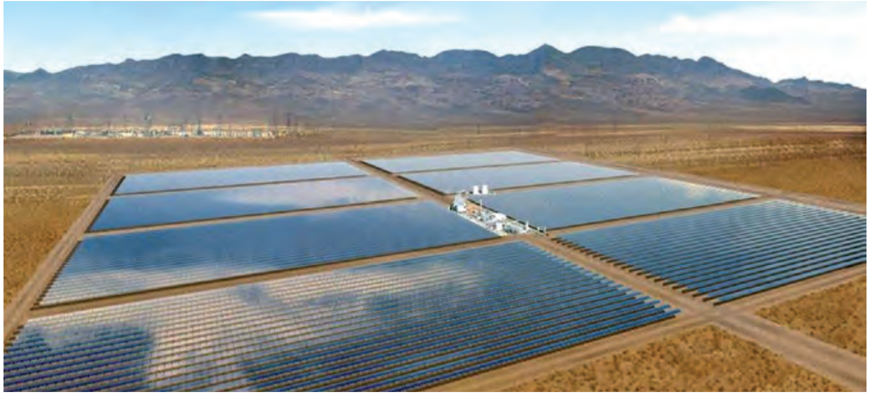


Figure 8. The 64 MW Acciona-Solargenix Plant in Boulder, Nevada.

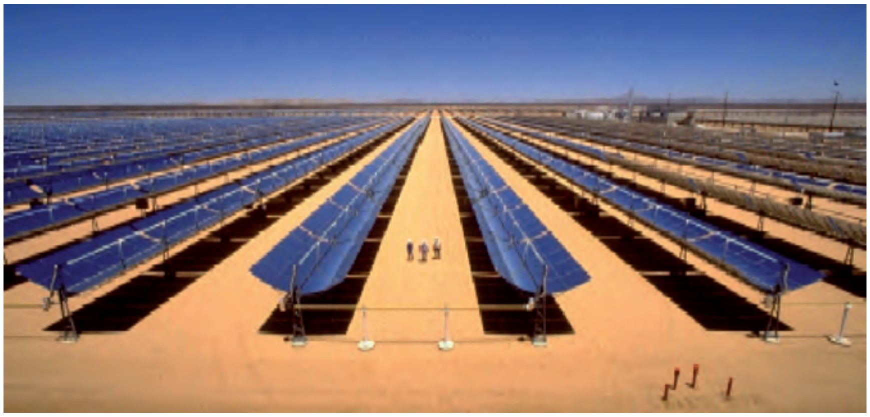
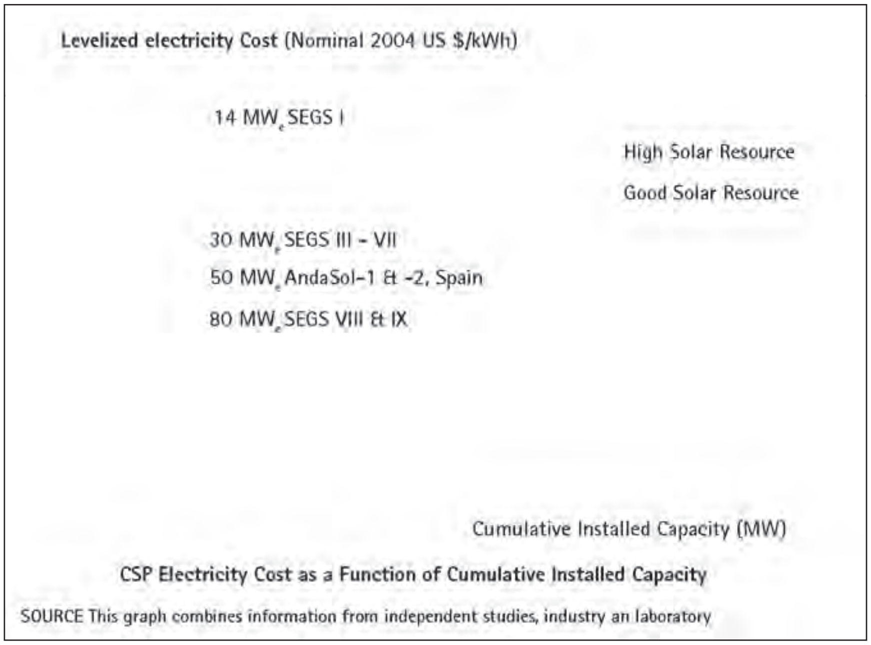


Figure 9. A view of the solar field at the SECS plants in Kramer Junction, California.

The second commercial plant in the world to use cylindrical-parabolic collectors is in the Nevada desert. It was built and is operated by Acciona. There are currently projects to built this type of plant in Spain, reaching a probable power potential of about 2,500 MW in the next four or five years. A considerable number will also be built in the United States, most with Spanish participation. For example, the so-called Solana Project was recently assigned to Abengoa, with two large thermosolar plants totaling 240 MW to be built in Arizona. Figures 8 and 9 give an idea of what this kind of plant looks like, and the sort of spaces in which it can be installed.

In Spain, among the numerous projects under way, is Andasol I. Nearly completed, this is the first 50 MW plant in a set designed by a consortium whose majority member is Cobra, of ACS, and a German firm called Solar Millennium. The Andasol plant, near Guadix in Granada, deals with one of the basic problems mentioned above with regard to the optimum use of renewable energies: storage. There, heat is stored, which has some advantages compared to storing electricity. In a plant with storage, when the sun is shining, part of the solar field feeds the storage device while the rest generates heat to produce electricity in the turbine. Thus, when the demand for electricity remains high after the Sun sets, it is possible to continue to generate electricity with the stored energy. In the case of Andasol I, the storage facility is able to continue generating electricity at maximum capacity for 7.5 hours, which makes the plant perfectly manageable and able to adapt its supply of electricity to meet demand.

The thermal storage employed in this type of plants is based on large quantities of melted salts (nitrates) that store heat by growing hotter, then release it again as they cool. It is a simple and safe system, although the levels of power being handled call for considerable amounts of salts. Specifically, Andasol I uses 28,500 tons of nitrates. There are other ways to store heat, including latent heat in materials that change phase, rather than heat that can be felt and is associated with temperature differences, or devices based on solids. These alternatives will be more clearly defined and improved as we gain more experience in this field.



CSP Electricity Cost as a Function of Cumulative Installed Capacity

SOURCE This graph combines information from independent studies, industry an laboratory

Figure 10. The estimated drop in the cost of thermoelectric electricity as a function of installed capacity according to CSP Global Market Initiative (SolarPaces 2004).

This type of solar energy is more costly than traditional energy, though less so than that of photovoltaic origin. Its bonus in the Special Tax Code is around 20 euro cents per kWh and, as with all renewable power sources, costs are expected to drop as the market expands. According to studies by SolarPaces, its cost will converge with those of conventional energy when around 15,000 MW have been installed, as can be seen in figure 10.

In order for this to happen, certain technological advances will have to be made, especially in the manufacturing of the absorption tubes, and the supply market will have to diversify. Its current narrowness impedes the development of the mechanisms of commercial competition that are essential for the improvement of fabrication processes. Improvements are also expected in heat-bearing fluids. As was mentioned above, a thermal mineral oil is currently being used, but it has the problem that, above a certain temperature (around 450° C), it decomposes. This makes it impossible to increase the working temperature, which would, in turn, increase performance when converting heat into electricity. Moreover, these oils are difficult to handle and contaminating. In that sense, there are already advanced programs to research the replacement of oil with another fluid, such as water or a gas that would allow the working temperature to be increased and simplify plant design, lowering its cost. These programs involve German and Spanish research groups working at the PSA, as well as the most important firms in that sector (see, for example, Zarza 2008). In sum, the challenges posed by the use of these technologies involve the optimization of tubes, of the heat-bearing fluid, of storage systems and collectors, and the expansion of global markets on the basis of public incentives.

Another technology being developed in the area of thermoelectric solar energy is based on a central receiver at the top of a tower. A field of rectangular heliostats focuses solar radiation on the receiver from which the resultant heat is extracted by a liquid or gaseous fluid. The first such plants operating commercially were built in Sanlúcar la Mayor (Seville) by Abengoa: PS-10 and PS-20, with capacities of 11 MW and 20 MW respectively. For the time being, their costs are higher than those of plants based on cylindrical-parabolic collectors, and their degree of development is somewhat slower. But they offer certain advantages, such as being able to operate at higher temperatures, and adapting to more irregular terrain. The process of improvement and optimization—still in its initial stages—is similar to what was described above, including the thermal storage devices, which are conceptually similar.

Nuclear fission

Along with fossil fuels and renewable energy sources, nuclear fission is presently an essential energy source in the most developed countries. In Europe, 30% of electricity is nuclear, while in Spain it is 20%. Nuclear energy has some advantages that make it attractive as part of the future energy menu. The main ones are its total independence of any kind of climatic or environmental conditions, which allows a plant to operate for a very high percentage of the hours in a year, as can be seen in figure 6. That explains how the nuclear sector in Spain, with an installed capacity of 7,700 MW, generated almost twice as much electricity as wind power, when the latter has a total installed capacity of 15,100 MW. Another positive factor to be taken into account is its relative independence from oscillations in the price of uranium because, over the useful life of the plant, fuel counts for barely 6% of the total building and operation costs. In figure 11, the cost of the raw material for nuclear plants is compared to that of other conventional energy sources.

Moreover, this is an industrial sector with considerable experience in safety, despite widespread opinion to the contrary. In fact, the most advanced and demanding safety protocols come specifically from the nuclear industry.

Its drawbacks are well known: from an economic standpoint, the enormous investments necessary to build the plants, with a very long period of depreciation, are the counterpart to the low cost of its fuel; from an environmental and safety standpoint, the potential seriousness of accidents when the plant is functioning—although there are very few—and, most of all, the generation of radioactive residues that are difficult to manage and store. The problem of residues is certainly the most serious drawback and, in public opinion, it has undoubtedly predominated over the more positive aspects of this energy technology. It therefore merits special consideration.

Generally speaking, there are two types of residues—short duration and long duration. Typically, the former have a half-life of 30 years (the half-life is the time that has to pass in order for a material's radioactivity to be reduced by half). The majority of residues fall into this category, and the universally accepted solution is to store them in a depository until their activity has dropped to the level of natural background radioactivity. El Cabril, in Cordoba, is a typical example of this sort of storage and, when properly managed, its effects on the environment are imperceptible.

The serious problem is residues with very long half-lives, measurable in tens or hundreds of thousands of years. That is the case of depleted fuel rods. Some countries have chosen to build Deep Geological Depositories (DGD) sufficiently hermetic to guarantee the stability of residues deposited there for geological time periods. Clearly, the difficulty lies not only in finding places that meet the necessary physical conditions, but also in getting a part of public opinion to accept this. Other countries, such as Spain, choose to build a Temporary Centralized Depository (TCD) at surface level, allowing safe custody of those residues for much shorter periods of time—about a century—while techniques are perfected for eliminating them or transforming them into inert matter. Indeed, the management or elimination of residues is one of the problems whose resolution is most pressing if we want nuclear energy to have a future. The principles of such a transformation are known—techniques of separation and transmutation—but their development is barely beginning. This is due to the complexity of the technology, and also the difficulty of experimenting with nuclear technology in the face of such strong public opposition.

In fact, the development of technology to neutralize the most dangerous residues is linked to what are known as fourth-generation reactors. Right now, there are 439 functioning commercial reactors in the world—104 in the United States and 59 in France—with a power capacity of 373,000 MW. Thirty-eight more are under construction in Finland, France, Eastern Europe and Asia (World Nuclear Association 2008). All of them are of second or third generation, operating with (slow) thermal neutrons and using the isotope ^{235}U for fuel. That isotope is very rare in nature, constituting only 0.7% of natural uranium. The most promising lines of the fourth generation operate with rapid neutrons and can use most existing residues for fuel, such as ^{238}U, which is the most abundant uranium isotope (it is the other 99.3%). They can even use thorium, which is even more abundant, and that alternative has been seriously studied in India. Fourth-generation reactors and devices using rapid-neutron technology—for example, accelerator driven systems (ADP)—could potentially solve many of the problems associated with residues and would be immune to an eventual long-term scarcity of conventional fuel (if we could use both types of uranium and not only the scarce fissionable isotope, reserves would automatically multiply by more than one hundred).

The unarguable challenges in the nuclear sector are thus the treatment of residues and fourth generation reactors, which are related to each other from a technological standpoint. But advances in this field take time and, at a level that can be exploited commercially, they will not be available for another twenty to thirty years. So most Western countries, with the noted exception of France and Finland, are faced with the difficulty of an improbable resurgence over that entire period, which could lead to a loss of knowledge and technical capacity. In contrast, many other parts of the world, especially Asia, will continue to build and operate second and third-generation nuclear reactors.

Conclusions

Given the situation described in the previous paragraphs, it seems neither realistic nor sensible to suggest abandoning any of the available energy sources, with the due precautions and in the time frames permitted by each technology. In the short term, there is a pressing need to prepare substitutes for petroleum by-products in the transportation sector, where we cannot avoid considering second-generation biofuels. Coal will continue to be an abundant, though potentially highly contaminating source, and it is necessary to make advances in its use with the capture and storage of CO_2.

But at this time, the most important challenge may well be to encourage renewable energies in order to make them a significant percentage of the total supply. This is still far from the case, but Spain is playing a leading role. Wind has proven its potential as a massive source of energy and must continue to broaden its presence on the global market. Solar energy is more abundant, but has the problem of dispersion discussed above. At some point in the near future, it will have to become the dominant and truly massive, sustainable

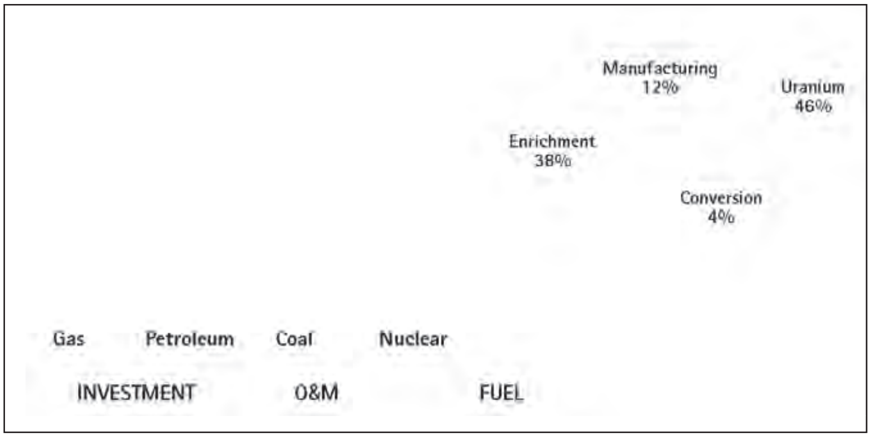


Figure 11. The distribution of costs in different types of electric power plants.

and unlimited renewable energy source. That will call for the solution of technological problems that limit its spread and affect its current high price, and will require decisive public support. In order to manage renewable energies and to meet the future needs of the transportation sector, energy storage technologies already occupy an outstanding place in energy research programs. So much so, that no sustainable scheme is conceivable without sufficient mastery of this sort of technology.

Unfortunately, nuclear fusion will arrive later on, and it is not likely to help alleviate the situation in upcoming decades. But fission reactors exist. They have been tested and have evolved toward ever-safer designs that use fuel more efficiently. I do not believe that it would be reasonable, in a period of energy crisis, to abandon this energy source, even though its survival largely depends on its public image. In the short term, the main problem is how to prolong the useful life of existing reactors and their replacement with third-generation technology. But the fundamental challenge in this area is to advance towards rapid fourth-generation reactors that make it possible to recycle residues and use fuel in an optimum manner.

No miracle has arrived to instantly solve the problem of supplying energy to humanity. It must be approached from all possible directions, and not only from a technological standpoint, as political and financial considerations are also important for each and every one of the available energy sources. Nor should we forget the educational and informational aspects, which are so important in a situation in which most of the population considers the energy problem to be solved and takes its continued supply for granted, yet refuses to accept the sacrifices inevitably associated with energy production from the standpoint of both economics and land use.

Bibliography

Ballesteros, M. "Estado del desarrollo tecnológico de los biocarburantes." *Energía*, vol. 202, 2007, 24–28.

British Petroleum. *BP Statistical Review of World Energy*, June 2008.

CIEMAT. *Análisis del Ciclo de Vida de Combustibles Alternativos para el Transporte*, 2005.

Club Español de la Energía. *Balance Energético de 2007 y Perspectivas para 2008*, Madrid, 2008a.

–, *Energía: Las tecnologías del Futuro*, 2008b.

European Commission. *The Sustainable Nuclear Energy Technology Platform*, 2007.

European Photovoltaics Industry Association. *Solar Generation V*, September 2008.

European Wind Energy Association. *www.ewea.org*, 2008.

Fundación para estudios sobre la energía. *El futuro del carbón en la política energética española*, Madrid, 2008.

Global Wind Energy Council. *Global Wind 2007 Report*, 2008.

International Energy Agency. *World Energy Outlook*, 2006.

–, *Key World Energy Statistics*, 2008.

Signes, V. et al. *Procedimiento para la revalorización energética de la fracción orgánica de residuos orgánicos e instalación.* International Patent Application ES2008/000077, 2008.

Socolow, R. H. "Can we bury global warming?" *Scientific American*, July 2005, 39–45.

Solar Paces. *The CSP Global Market Initiative*, 2004.

United Nations. *Human Development Report*, 2006.

World Nuclear Association. *www.world-nuclear.org*, September 2008.

Worldwatch Institute. "Biofuels for transport." *Earthscan*, 2007.

Zah, R. et al. *Ökobilanz von Energieprodukten: Ökologische Bewertung von Biotreibstoffen.* St. Gallen, Switzerland: Empa, 2007.

Zarza, E. et al. *Almería GDV: The first solar power plant with direct steam generation*, 2008.

climate change on the planet earth

SERGIO ALONSO

To my wife and companion, Mercedes, in memoriam.
To our daughters Aurora, Patricia and our son Carlos.

Introduction

The year 2008 marks the twentieth anniversary of the establishment of the Intergovernmental Panel on Climate Change (IPCC). Its creation grew out of an agreement between the World Meteorological Organization (a part of the United Nations) and the United Nations Programme. Its goal was to supply independent scientific information—in principle, to politicians—about questions concerning climate change. Almost ten years earlier, in the first World Climate Conference, attention was drawn to the increase in human activities, indicating that they might produce climatic alterations on a regional and even a planetary scale. Some years later, the role of CO_2 in climate variations was evaluated, along with other gasses capable of contributing to the so-called greenhouse effect. There was also a call for objective, balanced, and internationally coordinated scientific judgment that would shed light on the consequences of an increased concentration of greenhouse gasses in the Earth's atmosphere, and the socio-economic effects they might produce. This environmental concern, which was officially made public about thirty years ago, although it was actually older, led to the establishment of the IPCC in 1988. In 2007, the Norwegian Nobel Committee decided that the Nobel Peace Prize should be "shared, in two equal parts, between the Intergovernmental Panel on Climate Change (IPCC) and Albert Arnold (Al) Gore Jr. for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change."[1]

Some of the terms appearing in this introduction will be dealt with in more detail further on, but some should be clearly defined from the start. First, we should point out that the planet undergoes climate change on a continuous basis. We can be certain that, in the past, the climate was different than it is now, and that it will continue to change in the future. At first, terminology was somewhat confusing, with a coexistence of terms such as climate variation, climate variability, climate change, and climatic change. Finally (and unfortunately) two meanings remain in use. In scientific terms, climate change, means any change undergone by the planet's climate, regardless of its cause. This option is used, for example, by the IPCC. However, the United Nations Framework Convention on Climate Change, which arose from the so-called Rio Summit of 1992, and the Kyoto Protocol (established following the Convention), use the same terminology to refer to climate change attributed directly or indirectly to human activity, which is superimposed on natural variability. Therefore, when climate change is mentioned, care must be taken to make it clear which of the two meanings is intended. Notice, for example, that the Nobel Foundation's declaration specifies that it is referring to

[1] This paragraph is taken from the official announcement of the awarding of the prize by the Nobel Foundation.

climate change induced by humanity. Later on, we will see that this Climate Change of anthropic origin can be explained in terms of an intensification of the natural greenhouse effect. That intensification derives from a change in the composition of our atmosphere brought about by human activity.

The contents of this contribution include, in the next section, the reasons why the planet's climate changes, whether natural or anthropic. In section 3, we will review recent observations of changes experienced by the climate. The following section will contain arguments based on numerical simulations of climate, which attribute those changes to human activity. Section 5 offers a few indications about the use of computer models to simulate the Earth's climate. On the basis of the trustworthiness of such computer models, section 6 deals with the generation of climate scenarios for the future. Our conclusions are presented in section 7, followed by the bibliography employed.

Why does the climate change?

The climate is dynamic, changing and even unrepeatable. It is the consequence of the energy the Earth receives from the Sun, and of the exchanges of energy among different parts of what is called the Climate System, which we can understand as a synonym for the Planet Earth. Those parts or subsystems are:

a) The atmosphere, the planet's gaseous envelope, where we perceive the climate.

b) The hydrosphere, consisting of oceans, seas, lakes, and so on.

c) The lithosphere, the solid emerging crust of the continents, where we live.

d) The biosphere, made up of all living beings, including mankind, and

e) the cryosphere, which consists of all the ice that covers parts of the oceans and continents.

From a broad viewpoint, the climate can be defined as the state of the Climatic System, including its statistical properties. That is precisely what relates this definition of climate with the most classic and restricted one, which consists of a statistical description of environmental variables (for example, temperature, wind, surface humidity, and precipitation), using mean values and measurements of dispersion over long time periods, far superior to the typical periods of atmospheric weather.

The subsystems of the Climatic System mentioned above have very different dynamics. While some experience appreciable and continuous change (the atmosphere, for example, with its succession of quite different weather conditions—sunny, cloudy, windy, rainy, and so on), others change quite slowly, some so slowly that their variability merits little consideration over the course of a single human lifetime, or even several generations (that would be the case of the lithosphere, for example, except for the most superficial layer). When the energy we receive from the Sun reaches the Earth, it is distributed among all the subsystems and is exchanged among them, establishing relations according to the dynamics of each. The differences among these exchanges give rise to the great variety of climates in different regions of our planet, which we know so well, and which are a manifestation of the climate's spatial variability.

But climate is also characterized by variability over time. The Sun's energy does not arrive in equal amounts at all times, nor do the subsystems of the Climatic System always behave exactly the same. Therefore, we should not expect the energy flows that occur to invariably coincide over time. In certain intervals of time, their statistics can coincide more or less, but there is no reason to think that this must always be that case.

Next, we will analyze in some detail the origin of variability, that is, what causes changes in the Earth's climate. Some of these causes are natural, others are not—meaning that they have to do with human activity. The extant level of knowledge about the mechanisms we will see below is generally high, but we must not forget that, whenever there is a lack of knowledge (and there always is, of course) there will be a certain degree of ignorance, which leads to uncertainty in the interpretation of the observed phenomena.

First of all, we must begin by speaking of the Sun and its relation with the Earth. Its energy travels through space as radiation (called solar or short-wave radiation). It reaches the Earth, which intercepts it no matter what part of its orbit it is in or what time of the year. Not all the energy intercepted is used by the Climatic System. A fraction of it (called albedo) is returned to space through different processes of reflection mainly by clouds and the Earth's surface. Planetary albedo is around 30%. Finally, the radiation that is not absorbed by the atmosphere reaches the surface, which heats up and, in turn, emits its own radiation (called terrestrial or long-wave radiation). A large part of that radiation is absorbed by the atmosphere, which then re-emits it, either towards the surface or upwards, thus returning energy to space. For the entire planet, in average terms over time, there is an overall balance of energy, but not in the planet's different parts, nor at all times. It is these specific differences that affect the climate (see Kiehl and Trenberth 1997).

But how can the balance of energy be altered? According to what has been said, there could be three reasons:

a) Changes in the energy intercepted by the Earth. These may be due to changes in the Sun's emissions of radiation as a result of solar activity itself, or to changes in the position of the Earth in its orbit around the Sun.

b) Changes in the Earth's albedo. These, then, would be due to cloudiness (both degrees of cloud cover and types of clouds), changes in the reflective properties of the ground (types of ground and vegetation), and changes in the particulate matter suspended in the atmosphere. These particles are known as "aerosols."

c) Changes in the flow of long-wave energy from Earth to space. In this case, the changes would be due to a modification of the absorbent properties of the atmosphere as a result of changes in its composition.

Changes in solar activity have been recorded. The most popular may well be what is called Maunder's Minimum, which is though to have occurred between 1350 and 1850, coinciding with the so-called Little Ice Age (Hoyt, Schatten, and Nesme-Ribes 1994; Eddy 1976). Since that time it is estimated that radiation may have increased between 0.04% and 0.08%, with an increase of 0.05% between 1750 and the present (Wang, Lean, and Sheeley 2005).

But the Earth does not occupy a fixed position in relation to the Sun; it has a very approximate elliptical orbit—with the Sun at its focus—whose eccentricity changes cyclically over a period of about 100,000 years. That means that the Earth is not the same distance from the Sun, year by year, at the same point in its orbit—which is also changing. Moreover, the inclination of the Earth's axis with respect to the plane of its orbit (obliquity) is not constant. It is as if the Earth were a huge top, so the prolongation of its axis of rotation points to different places in the celestial dome in cycles lasting around 41,000 years. Also, the orbital ellipse changes its orientation in space, leading to what are called the precession of equinoxes. That means that the astronomical seasons take place in different parts of the orbit with cycles lasting approximately 19,000 and 23,000 years. The final result is that, even if the energy emitted by the Sun were constant, what actually affects the system varies, and is also distributed differently over the planet's surface. All of this constitutes what is called Milankovitch's Theory of Cycles, which, along with certain internal mechanisms, makes it possible to explain the succession of geological eras (Berger 1988).

The processes we have described are external to the Climatic System and in no way depend on human activity. Another possible cause of planetary Climate Change, which is also both external and natural but has no relation to the solar radiation received by the Earth, is the impact of meteorites or comets. This is something difficult to predict, but its consequences are important when the objects are big enough. Their impact against the surface of the planet can cause a cloud of dust or water of such magnitude that incident solar radiation cannot reach the Earth's surface with the intensity it had before impact. In those conditions, the temperature can drop appreciably, leading to climate change. The extinction of some species, including dinosaurs, in what is called the K/T Boundary, seems to have this origin (Álvarez et al. 1981).

This cause, which we can qualify as exceptional, allows us to bring in those related with albedo. Following impact, there must have been a considerable increase in albedo because of the increased amount of aerosols (particulate matter) in the atmosphere. This would have reflected a very high fraction of solar radiation back into space. In consequence, the Climatic System would suddenly have had much less energy to heat the ground and, thus, the previous balance of radiation would have been altered. The result must have been a lowering of the temperature at ground level. Without reaching those extremes, something similar happens each time there is a volcanic eruption. Their effect on temperature has been observed following large eruptions and depends on the intensity of the eruption, and on how high up in the atmosphere the generated particles reach. The effect, which can last several years, has been widely studied (see, for example, Yang and Schlesinger 2002).

The aerosols we have considered up to now are of natural origin but, besides these, the Earth's atmosphere also contains many others stemming from human activity. Generally, they reduce air quality and many of them also lead to health problems. From a climatic standpoint they have two effects. One directly affects albedo, leading to lower temperatures. The other has an indirect effect, modifying the conditions in which clouds are formed and how long they last. The final result of this indirect effect is not well known. Nowadays, it is the subject of uncertainty.

Clouds' role in albedo depends on cloud cover, the type of cloud, and how long it lasts. Thus, high clouds (cirrostratus clouds, for example) allow solar radiation through, but absorb terrestrial radiation, while medium clouds (altocumulus clouds, for example) almost completely impeded the passage of solar radiation. The first case will result in a rise in temperatures, while in the second they will fall.

Albedo also depends, as mentioned above, on the reflective properties of the planet's surface. A frozen surface (high albedo, of 70% to 90%) is not the same as bare earth, prairie, or the ocean's surface (low albedo, <10%). Different types of terrain and ground-use mean that the climatic treatment of the Earth's surface is a complex problem and a source of uncertainty.

At this point, we cannot avoid commenting on one type of behavior that is characteristic of the Climatic System. Often, the effects of a process act on its own causes, generating a sort of cyclical, unending behavior called feedback. Feedback is typical in what are called non-lineal or dynamic systems, and the Climatic System is one of them. The following example is relatively straightforward: let us suppose that, for whatever reason, the planet's surface temperature rises. One of the consequences will be the partial melting of its ice. Surface albedo will diminish, leading to decreased reflection of solar radiation. There will thus be more energy available to the system, and the temperature will rise further. The additional heating will lead to greater ice melting, reducing albedo even more, and so on and so forth. This, then, is a positive feedback cycle known as ice-albedo feedback. It was already identified in the nineteenth century (Croll 1890). In the Climatic System, there are many other positive feedback cycles like this one, but there are also negative ones. When those feedback processes act at the same time, it becomes very difficult to obtained detailed knowledge of the results, even though it is clear that they exist. The only possible way of dealing with the problem is through numerical simulation of those processes.

The last way of modifying the balance of radiation to be mentioned here might well have been the first: it is the main way of explaining the climate change the planet is experiencing today.

First, we will consider the role the atmosphere plays in exchanges of solar and terrestrial radiation, which is known as the Greenhouse Effect (GE). We have already mentioned that part of the radiation coming from the Sun—about 30%—is reflected back into space. If the Earth did not have an atmosphere, the planet's surface would have an average temperature of -18ºC, barely enough to maintain the energy equilibrium between penetrating solar radiation and terrestrial radiation (infrared) that the Earth would emit at that temperature. The Moon, which has no atmosphere, has an average temperature like that. But since the Earth *does* have an atmosphere, things are radically different. The atmosphere's constituents absorb relatively little solar radiation (especially where there are no clouds) but some of them are very good at absorbing the infrared radiation emitted by the Earth and by the atmosphere itself. This leads to a warming of the lower layers of the atmosphere, which modifies the balance of radiation, reaching an average temperature of 15ºC at ground level. This behavior by the atmosphere, which reacts differently to solar radiation than to terrestrial radiation, is the GE, whose name comes from its relative similarity to the behavior of such structures. The main cause of GE is water vapor (approximately 80% of the total effect) and the second cause, at a considerable distance, is carbon dioxide (CO_2). The GE (to which the adjective "natural" is often added) is decisive in the planet's climate, which has allowed the existence of life, at least as we know it. The gasses that contribute to the GE are known as greenhouse gasses (GHG). That said, it should be obvious that the GE is also affected by aerosols and that the role of clouds can also be discussed in those terms.

Any change in the composition of the atmosphere, or in the concentration of its components, alters its properties of absorption, consequently altering the GE as well. The atmosphere's composition has been changing for as long as the Earth has existed. Nitrogen (N_2) and oxygen (O_2) predominate, although the major contributors to the GE are water vapor (whose concentration does not surpass 4% of the atmosphere's volume) and CO_2 (with a much smaller concentration, currently around 385ppm[2]). If the atmosphere's composition changes, the GE will be modified and thus, the planet's mean surface temperature will change. Before the industrial revolution, the mean global concentration of carbon dioxide was around 280ppm, while it is now about 385ppm, as mentioned above. In these conditions, the planet's natural GE has been undergoing modification ever since the Industrial Revolution began. As the concentration of CO_2 has increased (those of other GHGs are also rising, including methane, nitrous oxide, CFCs, and so on) the GE has enhanced, more energy has become available in the lower layers of the atmosphere and, thus, conditions have arisen for warming on a planetary scale. This is not modern speculation; in the late nineteenth century, the Nobel scientist Svante Arrhenius estimated the effect of a 40% increase or decrease in atmospheric CO_2 on temperature, indicating that the glaciers could shrink or expand (Arrhenius 1896). Actually, by the end of the seventeenth century, there was already knowledge of the different behavior of certain substances with regard to solar and terrestrial radiation, which is the basis for GE.

By analyzing air from bubbles trapped in core samples extracted from polar ice, it is possible to obtain information about the evolution of the concentration of GHGs in past periods. These can also be compared to current levels. Figure 1 shows the value of carbon dioxide, nitrous oxide, and methane concentrations over the last 650,000 years. We can see that current values far surpass earlier ones, even in the warmer glacial periods. These are shown in Figure 1 as shaded bands. The lower part also shows variations in the concentration of deuterium, δD, in arctic ice. This serves as an indirect indicator of temperature variations. Note the values of δD in earlier warm

[2] Parts per million, a measure of concentration for scarce parts. It is equal to a molar fraction of µmol/mol. Similarly, a molar fraction of n mol/mol is represented in ppb (parts per billion) and p mol/mol is expressed as ppt (parts per trillion). Considering the non-ideal behavior of gasses, concentrations of volume different than those mentioned above are sometimes employed (ppm, ppb, ppt).

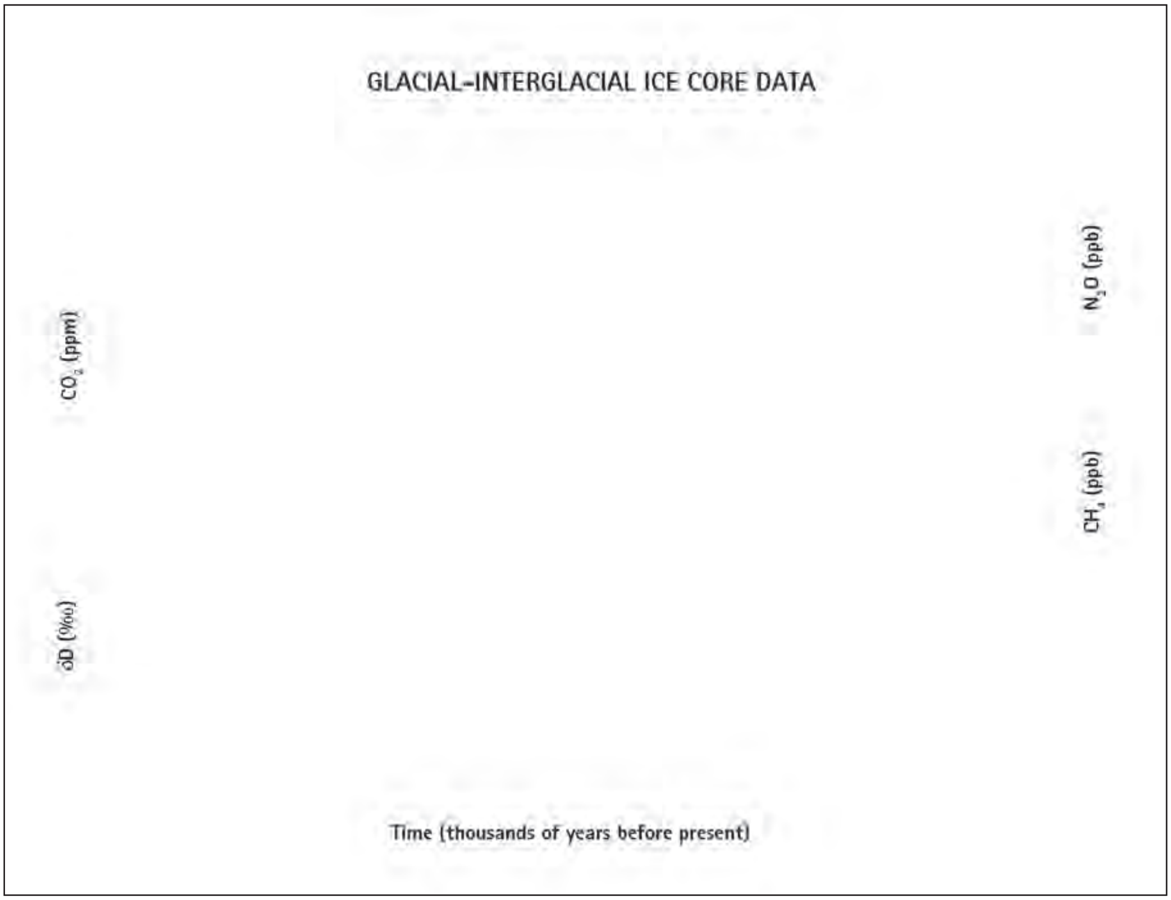


Figure 1. Variations of deuterium (∂D) in antarctic ice, which is a proxy for local temperature, and the atmospheric concentrations of the greenhouse gases carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) in air trapped within the ice cores and from recent atmospheric measurements. Data cover 650,000 years and the shaded bands indicate current and previous interglacial warm periods.

periods and in the present one, and the large difference in concentrations of GHGs. Unlike the present period, in which the relation between GHG and temperature is clearly established and the anthropic origin of the change in GHG is proven, there is still much to be discovered about many aspects of this relation in the past. It is thought that, in the Quaternary, changes in the concentration of CO_2 may have resulted from the simultaneous effects of biological and chemical processes in the ocean. They may also have been affected by changes in temperature (Köhler et al. 2005). Concentrations of CO_2 did sometimes surpass current levels in earlier periods, millions of years ago, but these are thought to have been the result of tectonic processes, such as volcanic activity, which determined changes of concentration (Ruddiman 1997).

Recently, as a result of the European EPICA research project, the time range has been increased to 800,000 years. The same conclusions hold with regard to concentrations of GHG indicated in the description of Figure 1 for the last 650,000 years (Lüthi et al. 2008; Loulergue et al. 2008).

Figure 2 shows variations in the concentration of CO_2, CH_4, and N_2O, but for shorter time periods (panels a, b, and c). The scale on the left of those panels shows the concentration of the corresponding GHGs, while the scale on the right represents what is called radiative forcing, which is equivalent to the intensification of the GE that implies increased concentrations of GHGs, as expressed in radiation units (Wm^{-2}). These three panels indicate that the change experienced by GHGs following the Industrial Revolution has no recent precedent: while the atmospheric concentration of CO_2 increased only 20ppm over the 8,000 years preceding industrialization, since 1750 it has increased more than 100ppm. Approximately two thirds of this increase is due to the burning of fossil fuels and the remaining third is due to land use change. Panel d represents the rate of change of the combined forcing of the same three GHGs, which gives an integrated value of $1.66Wm^{-2}$ since 1750. This amount is by far the greatest of all possible forcings associated with the different mechanisms responsible for climate change analyzed in this section.

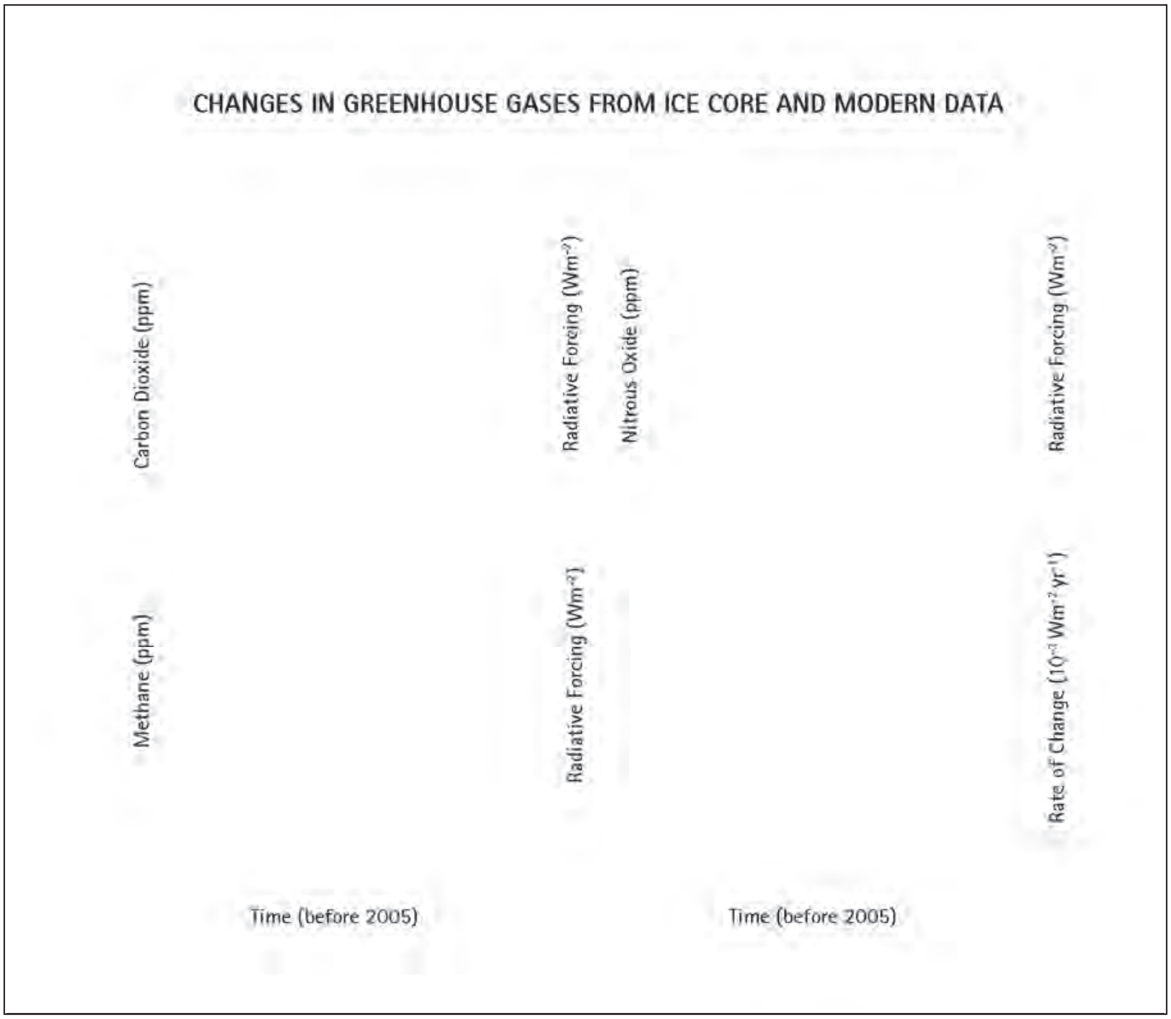


Figure 2. The concentrations and radiative forcing by (a) carbon dioxide (CO_2), (b) methane (CH_4), (c) nitrous oxide (N_2O) and (d) the rate of change in their combined radiative forcing over the last 20,000 years reconstructed from antarctic and Greenland ice and firn data (symbols) and direct atmospheric measurements (panels a,b,c, red lines). The grey bars show the reconstructed ranges of natural variability for the past 650,000 years. The rate of change in radiative forcing (panel d, black line) has been computed from spline fits to the concentration data. The width of the age spread in the ice data varies from about 20 years for sites with a high accumulation of snow such as Law Dome, Antarctica, to about 200 years for low-accumulation sites such as Dome C, Antarctica. The arrow shows the peak in the rate of change in radiative forcing that would result if the anthropogenic signals of CO_2, CH_4, and N_2O had been smoothed corresponding to conditions at the low-accumulation Dome C site. The negative rate of change in forcing around 1600 shown in the higher-resolution inset in panel d results from a CO_2 decrease of about 10 ppm in the Law Dome record.

In essence, what we have presented so far are the climate drivers related to the balance of radiation on a global scale. As was indicated above, the climate is a consequence of energy flows in different parts of the Climatic System. Now is when a large number of processes with their own internal dynamics come into play, with a great wealth of time scales, making the system truly complex. As a result, the Climatic System is very difficult to deal with. To study it in its entirety calls for numerical simulation. What must be clear is that whenever the functioning of one part of the machinery is modified, the end result will be a change of climate (see IPCC 2007).

Nowadays, when we speak of (human-induced) climate change, we are referring to climate change observed today, which is a consequence of the intensification of the GE. In the final analysis, this is a change in how Planet Earth functions as a consequence of human activity. That is what has come to be known as Global Change, leading some researchers, including Nobel Scientist Paul Crutzen, to say that the planet has entered a new era characterized by anthropic impact. That is why they propose this era be called the "Anthropocene" (Crutzen and Stoermer 2000).

Observing change

But is the climate really changing? Many people wonder, and ask the specialists. From a general viewpoint, the answer is yes. This planet's climate has always been changing. And now? In the Anthropocene?

Yes, now too. There are two aspects of current climate change that should be mentioned. The first is that, unlike earlier change, it has such a short time scale that change is appreciable over a period comparable to a human lifetime. The second is that humanity has never before had the capacity to interfere with climate on a global scale. It so happens that this planet's climate made life, including human life, possible. Now, the human species is capable of modifying that climate. These two characteristics make it possible to state that, strictly speaking, there is no past precedent for current climate change.

In this section, we will offer some of the evidence for current climate change. The following one will deal with procedures that have led to the conclusion that human activity is responsible for the observed changes.

In its fourth, and most recent report (IPCC 2007), the IPCC indicates that, compared to its third report (the third is designated by the acronym, TAR, and the fourth, AR4), there are now better data bases, more evidence, greater geographic coverage and a better understanding of uncertainties. As a result, AR4 indicates that the warming of the Climatic System is unequivocal, as can be deduced from observations of increased mean atmospheric and oceanic temperatures on a planetary scale, extensive melting of snow and ice and the global rise in the average sea level.

In TAR, calculations of warming of the mean global air temperature at ground level between 1901 and 2000 gave a linear trend of 0.6 ± 0.2°C per century. This was surpassed by AR4's calculations of that rate for the period between 1906 and 2005, which was 0.74 ± 0.18°C per century. The acceleration of warming becomes even clearer when we use only the last fifty of those one hundred years (1956–2005), and even more so in the last 25. In those cases, the resultant linear trend is 1.28 ± 0.26°C per century and 1.77 ± 0.52°C per century, respectively.[3] The temperature increases noted here are very likely unprecedented on Earth, at least in the last 16,000 years.

Changes in temperature extremes have also been observed, and these are consistent with warming of the lower layers of the atmosphere. Thus, the number of cold and frosty nights has diminished, while the number of warm days and nights, and heat waves, has increased.

If we analyze the spatial distribution of these trends (which are greater on land than over the oceans) and the seasonal values, we will find important differences. The same occurs with separate calculations of maximum and minimum temperature trends. For example, the results of an analysis of temperature trends on the Balearic Islands over a thirty-year period ending in 2006 (OCLIB 2007) showed a linear trend of 4.83 ± 1.85°C per century for the maximum temperature, with 5.14 ± 1.89°C per century for the minimum. The maximum value for the minimum temperature appeared in the summer (8.01 ± 3.17°C per century), while the maximum value for the maximum temperature (7.99 ± 3.01°C per century) appeared in the spring. It is important to note the large differences encountered here with respect to global values, even with the highest one quoted before, which corresponds to a period of 25 years.

The average temperature of the ocean has also risen, at least to depths of about 3,000 meters. It is estimated that, since 1955, the ocean has absorbed around 80% of the excess heat resulting from the GE. This results in the expansion of seawater and significantly contributes to sea level rise.[4]

Moreover, we must point out important changes in the cryosphere. For example, the surface area of arctic sea ice has diminished an average of 2.7% per decade, and that reduction process intensifies in northern hemisphere summers, where it reaches 7.4%. In the summer of 2007, the reduction of surfaces with at least 15% ice coverage was especially notable, after AR4 were developed. Such covered surfaces reached a summer minimum of 7.5 million square kilometers (averaged between 1979 and 2000) while, in the summer of 2007, only 4 million square kilometers were covered. That is the smallest surface area since Earth-observation satellites have existed. Values for the summer of 2008 show a slight recovery compared to 2007, but still far below the previously indicated average.[5]

Figure 3 indicates observed changes in the last century-and-a-half in the mean global surface temperature (panel a), the average sea level (panel b) and the surface of the Northern hemisphere covered with snow (panel c). The relative scale at the left of figure 3 shows the variation of those changes with respect to the average value between 1961 and 1990.

Global rainfall measurements are also being affected by current climatie change. To start with it must be said that there has been a continuous increase in the total content of water vapor in the atmosphere, which is coherent with the temperature increase in the troposphere. Precipitation has been modified to an unequal extent in different geographic areas. While it has significantly increased in eastern parts of North and South America, northern Europe, and northern and central Asia, the climate is now drier in the Sahel, the Mediterranean, Southern Africa, and part of Southern Asia. If we look at the extremes, on one hand the occurrence of strong rains over land has become more frequent, but on the other more intense and lasting droughts have been observed since the nineteen seventies, particularly in the tropics and subtropics,

[3] Warming has been observed in the average global temperature at surface level and in the troposphere. At higher levels—the stratosphere, for example—cooling of beween 0.3 °C and 0.9 °C per decade has been observed since 1979, although this has diminished in recent years.

[4] Variation in sea level is a complex problem lying outside the scope of this text. From a climatic standpoint, the main contributions, in almost equal measure, are the expansion of sea water (including the salinity effect) and the melting of continental ice. On geological time scales, there have been very important changes in sea levels. For example, it is estimated that, during the ice ages, level was over 100 meters lower than today.

[5] Information drawn from http://nsidc.org/arcticseaicenews/index html, consulted on 17 August, 2008.

sometimes in combination with flooding in those same geographical areas.

It is difficult to obtain figures on the trends of global precipitation, due mostly to the characteristic discontinuity of the variable and to measurement methods. As an example on a much smaller scale, the results offered below are from an analysis of precipitation trends in the Balearic Islands over a series of 55 years through 2006 (OCLIB 2007). Smoothing the annual precipitation series with a 5-year filter generates a tendency of -170 ± 123mm per century, which becomes -192 ± 38mm per century when the annual series is filtered with a 30-year average. Consideration must be given to the fact that the normal precipitation in the Balearic Islands is close to 600mm per year, which represents a decrease in rainfall tending towards 30% in one hundred years. This reduction has not been equally spread among the seasons, nor for all types of precipitation. Decreases have been greater in fall and winter, and much less so in spring and summer, linked to a decrease in the number of days with moderate rainfall, although the number of days with weak rainfall has increased, as has the number of days with strong rains, though to a lesser degree.

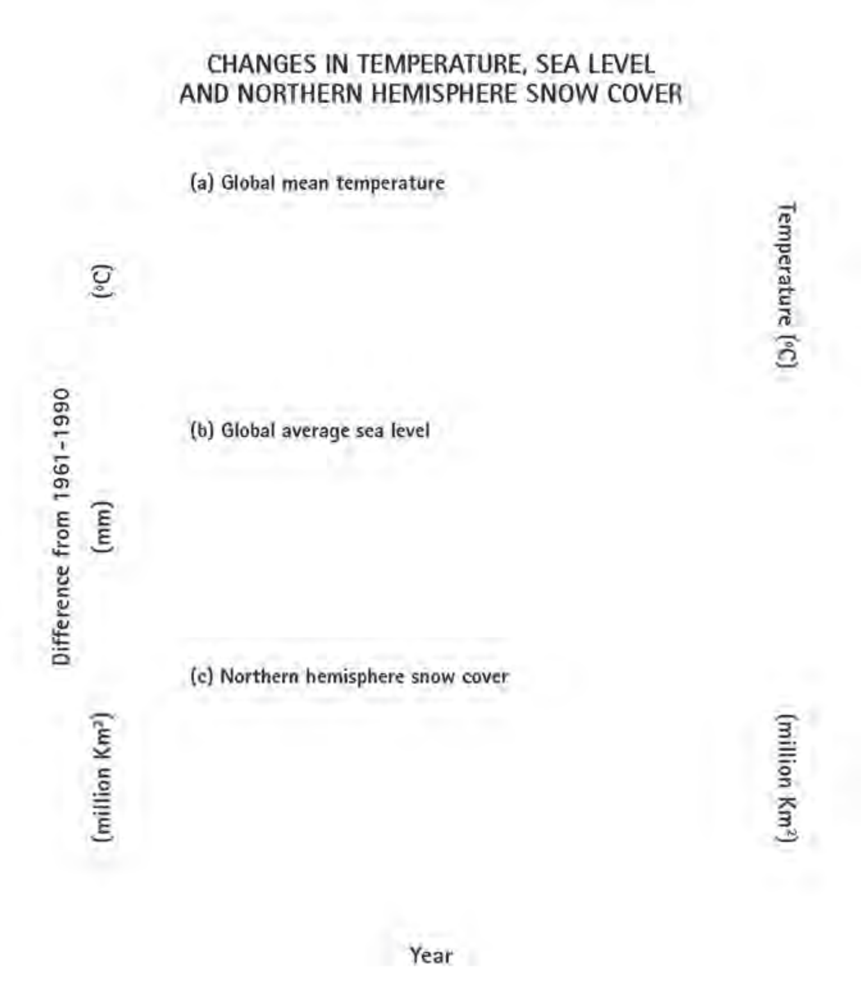


Figure 3. Observed changes in (a) global average surface temperature; (b) global average sea level from tide gauge (blue) and satellite (red) data and (c) Northern Hemisphere snow cover for March-April. All differences are relative to corresponding averages for the period 1961-1990. Smoothed curves represent decadal averaged values while circles show yearly values. The shaded areas are the uncertainty intervals estimated from a comprehensive analysis of known uncertainties (a and b) and from the time series (c).

Observed changes in rainfall data are explained, in part, by the previously mentioned increase in atmospheric water vapor content, but also by the change in patterns of atmospheric circulation characteristic of the climate's natural variability. These include North Atlantic Oscillation (NAO) and the phenomenon El Niño/Southern Oscillation (ENSO).

Scintists are also confident about changes observed in some other extreme phenomena not mentioned here (for example, increases in the number and intensity of tropical Atlantic cyclones). But for others (tornados, lightening, hail, Antarctic sea ice, and dust storms) there are not yet enough reliable results to allow us to be certain that they have experienced variation in the present climate.

For more information on the changes observed it is necessary to consult AR4 (IPCC 2007).

Attribution of observed climate change

The term "attribution" is used here to indicate the process by which we evaluate whether the observed changes are consistent with quantitative answers to the different causes of planetary climate change simulated with well-tested models, and not consistent with other physically possible alternative explanations. In this section, we will take it for granted that the climate can be simulated in a sufficiently adequate manner; in the following one, we will try to offer arguments that make it clear that this is indeed the case.

Ever since the IPCC drew up its first report in 1990, the subject of attribution has been addressed. In that first report (FAR) there was not sufficient observational data on anthropic effects on climate. The second report (SAR) concluded that overall evidence suggested a discernible human influence on the twentieth century's climate. TAR indicated that the greater part of warming observed in the last 50 years was probably due to increased concentration of GHGs. Since that report, confidence in the evaluation of humanity's effect on climate change has increased considerably. We have more evidence, and the methodology of attribution has improved. All of this appears in AR4 and will be summarized below.

Attribution of current climate change will be carried out here using the results for temperature, the variable most clearly determined, and whose simulation is most resolved. The observed evolution of temperature will be compared with what models are able to simulate. Figure 4 offers the results of a comparison of mean global temperatures with what climate models simulate for the twentieth century in different circumstances. In both panel a and panel b, the black curve represents the evolution of the global mean surface temperature. The values deduced from the scale on the left are temperature differences with respect to the average in the period 1901–1950. The red curve on panel a represents the mean evolution of the simulated temperature. It is obtained by averaging out the results of each of the individual models, whose different processes are represented in ocher. For this simulation, the models include known causes of climate change —specifically, natural ones, including volcanic eruptions—and those that are a consequence of human activity, using the known evolution of atmospheric concentrations of GHGs and aerosols. The result of this attribution experiment can be summed up by saying that there is a strong correlation between the evolution of observed and simulated temperatures, that the envelope of individual simulations almost completely includes the curve of observations, and that the average of the models closely approximates that of the observations when conveniently filtered by a time average (not shown in the figure).

Panel b presents the results of simulating the evolution of temperature using only natural causes of climate change. As before, it shows both the individual processes of models, in light blue, and the average of all the simulations, in darker blue. But here, the same conclusions cannot be drawn. The natural forcings can only explain the evolution of temperature through approximately the middle of the past century. In fact, a comparison of the two panels does not reveal large differences in the two simulations for that time period. The differences arise in the second half of the twentieth century. It is necessary to include anthropic causes in the simulations in order to explain the temperature trend in the second half.

This type of experiment had already been carried out in TAR (IPCC 2001), but the conclusions were not as trustworthy as in AR4. Moreover, equivalent studies have now been carried out for the different continents, separately for land and sea, and for other different temperature variables. The results are coherent with what has been stated above.

Climate research should always tend to reduce uncertainty while also achieving increasing realism in its simulations. Figure 4 shows important discrepancies between the simulations and mean surface temperature calculated by direct measurements around 1940. Analysis of the origin of the temperature observations concludes that there is a bias in the observed values as a result of the method employed to measure the sea's surface temperature, which obviously forms a part of the planet's surface temperature (Thompson et al. 2008). If the observed values were corrected, the discrepancy would be reduced, bringing the observed temperature evolution closer to the simulation. At the time that AR4 was published, the above was not yet known, but the results were still considered sufficiently realistic to indicate that "most of the observed increase in global average temperatures since the mid-20th century is very likely[6] due to the observed increase in anthropogenic GHG concentrations."

Simulation of the Earth's climate with models

Knowledge of the mechanisms that determine climate, set out in section 2, is partial but sufficient to allow us to simulate it (not in a laboratory, of course, but using complex models run by powerful computers). It has become possible to reproduce the current and past climates with sufficient accuracy, as well as the fundamental known traits of the climate in far earlier geological eras. Thanks to that, attribution exercises have been carried out, as indicated in section 4, and we can also think about inferring the possibilities of the future climate, including man's role in it. This last matter will be addressed in the following section.

Let us now look at climate simulation models in some detail. In the first place, we should state that such models are not an invention of climate researchers; in physics and other sciences, models are generally employed and they have turned out to be extraordinarily useful for advancing knowledge. In general terms, a model is a simplification of reality used as a tool to describe and explain phenomena of scientific interest. Models are sometimes constructed through mathematical equations that describe empirical relations among variables characteristic of the system being studied. For example, such relations can be obtained on the basis of an adequate statistical treatment of those variables. At other times, previous and independently established physical laws are used to establish the relations among variables. Moreover, in this case, they allow an interpretation of why this relation exists because, in fact, that is what these laws express. Finally, there are also mathematical equations that relate variables but are in this case based on physical laws.

In all cases, a set of equations is obtained that makes it possible to offer an approximate (remember, these models are simplifications) description of reality. It is precisely this fact that makes it possible to at least partially explain the discrepancies that appear between

[6] IPCC uses this term to indicate that the probability surpasses 90%.

a simulated description of reality generated by a model, and the reality of observations of a real phenomenon.

Once the set of equations that constitute a model is obtained, those equations must be written in such a way as to furnish quantitative information about the system being studied. In the case we are discussing here, at the very least, they would have to furnish values for temperature and precipitation in order to reveal the fundamental traits of climate. Moreover, they would have to do so for the entire planet and, actually, at different levels of the atmosphere, from the lowest ones, at ground or sea level, to the highest. And that is only the part that deals with the atmosphere, because in other subsystems, it will be necessary to know many other variables (for example, the salinity and temperature of the oceans, ice mass, and the properties of soil and vegetation), at different levels or depths as well. The conclusion we must draw from all this is that the model's equations must be applied to a large number of points in space. Many mathematical operations have to be carried out in order to determine all the variables that describe the state of the Climatic System at a single instant in time. But in order to characterize climate, we must know what happens, not only at a specific moment, but over the course of sufficiently long time spans. That is, an enormous succession of individual instants.

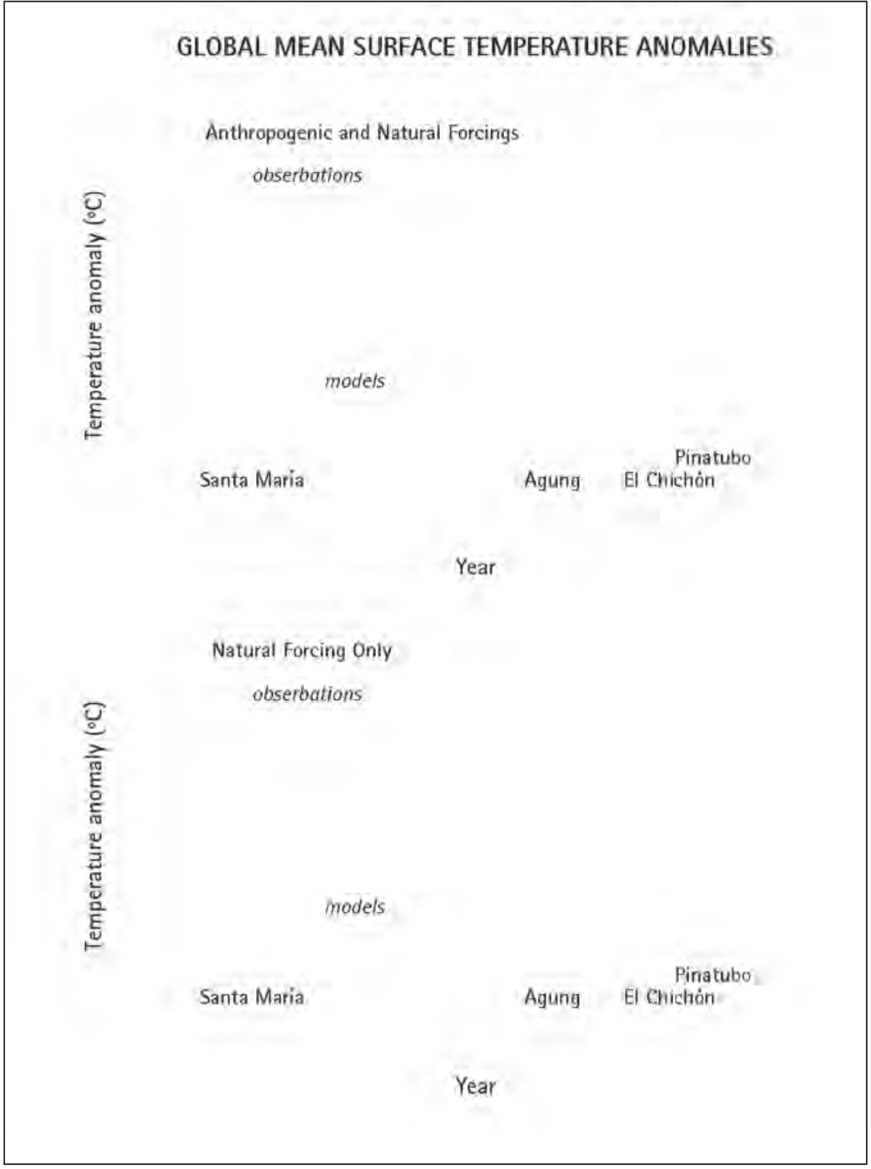


Figure 4. (a) Global mean surface temperature anomalies relative to the period 1901 to 1950, as observed (black line) and as obtained from simulations with both anthropogenic and natural forcings. The thick red curve shows the multi-model ensemble mean and the thin lighter red curves show the individual simulations. Vertical grey lines indicate the timing of major volcanic events. (b) As in (a), except that the simulated global mean temperature anomalies are for natural forcings only. The thick blue curve shows the multi-model ensemble mean and the thin lighter blue curves show individual simulations. Each simulation was sampled so that coverage corresponds to that of the observations.

How can we approach such a huge task? The answer is not immediate. First of all, if we want to obtain useful climate information in a reasonable time, we must use very powerful computers—the most powerful in the world. In order to do so, we must, again, simply the model, writing it in a form that is adequate for computer work. Once this is done, computers will be used to carry out the millions and millions of mathematical operations needed to obtain climate simulations for various decades, centuries, and so on, in a reasonable amount of time. Numerical simulations of climate are often mentioned in order to designate the means by which the desired climatic information is obtained.

The most advanced models of climatic simulation include formulas that address processes in the atmosphere, the oceans, the Earth's surface, and the cryosphere, atmospheric chemistry and the modeling of aerosols. They also deal in a linked way with atmosphere-oceanic interactions. Some models include mechanisms for maintaining energy flows at reasonable values, but nowadays, due to advances in research, most of them do not need this adjustment because the flows obtained directly by the simulations are already realistic. Those climate simulation models that include equations for the treatment of the processes mentioned here are generically called Atmosphere/Ocean General Circulation Models (AOGCMs). Many models exist, generally linked to leading research centers around the world, and their climate simulations offer different results, all of which are plausible. There are intercomparison projects and programmes in which results are contrasted in order to verify performance, which also makes it possible to establish confidence levels for the results. The IPCC itself bases a large part of its evaluation reports (see chapters 8 and 9 of AR4, IPCC 2007) on simulations. Confidence in climate simulation has been obtained by verifying that the results are sufficiently realistic in comparison with observations. Those results affect the different subsystems of the Climatic System and known of variability modes of the current climate, including the phenomena of El Niño/Southern Oscillation (ENSO) and North Atlantic Oscillation (NAO), patterns of anticyclonic blockage and the variability of monsoons. But verification by contrast with the present climate is not the only source of confidence. From a conceptual viewpoint, the primary source is that these models utilize physical laws that were independently established before the problem of climate simulation was even addressed. Moreover, it has become possible to simulate important traits of the climate of the last 2,000 years, as well as earlier climate change, such as the warm period in the Holocene, some 6,000 years ago, and the variability of the ice ages. It goes without saying that the results are reliable enough to foster confidence in the use of such models, despite the fact that there are still areas of uncertainty.

One of the main advantages to using models to simulate climate is that processes included in those models can be activated or deactivated at will. It is enough to eliminate the set of equations that affect a specific process in a given model. That model is then capable of simulating the planet's climate with or without the activity of the process (or processes) under study. Thus, for example, following a volcanic eruption, the additional effect of expulsed aerosols can be included, or the intensification of the GE can be eliminated while pre-industrial concentrations of GHGs are being considered. That, precisely, is the basis for the attribution of climate change dealt with in the previous section.

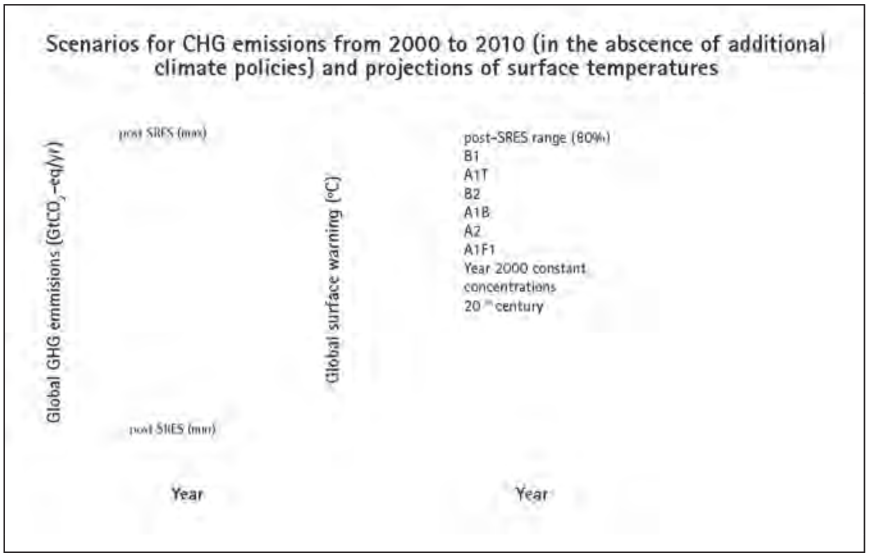


Figure 5. Left Panel: Global GHG emissions (in GtCO$_2$-eq) in the absence of climate policies: six illustrative SRES marker scenarios (coloured lines) and the 80th percentile range of recent scenarios published since SRES (post-SRES) (gray shaded area). Dashed lines show the full range of post-SRES scenarios. The emissions include CO$_2$, CH$_4$, N$_2$O and F-gases. Right Panel: Solid lines are multi-model global averages of surface warming for scenarios A2, A1B and B1, shown as continuations of the 20th-century simulations. These projections also take into account emissions of short-lived GHGs and aerosols. The pink line is not a scenario, but is for Atmosphere-Ocean General Circulation Model (AOGCM) simulations where atmospheric concentrations are held constant at year 2000 values. The bars at the right of the figure indicate the best estimate (solid line within each bar) and the likely range assessed for the six SRES marker scenarios at 2090-2099. All temperatures are relative to the period 1980-1999.

If we do not want to use large computers, or do not have access to them, there are also more modest solutions, which are not necessarily less useful. It is possible to gain access to a second level of climate simulation using a new simplification of the Climatic System. In other words, it is possible to simplify the complexity of the model—which is already, itself, a simplification of reality—in order to be able to work with personal computers or the like. In such cases, it is a matter of making sure that the simple models offer simulations that are compatible with those being carried out with AOGCMs.

To give us an idea: at the maximum extreme of simplification, we could consider the Earth a sphere that receives energy from the Sun and maintains the equilibrium of that energy with the energy it reflects, and that which the Earth itself radiates into space. In such conditions, a temperature—called the equilibrium temperature—is determined. It turns out to be around -18ºC and is very different than the mean temperature on Earth, which is about 15ºC. These same figures were mentioned above when discussing the natural GE. In other words, the equilibrium temperature is obtained by a maximum simplification of the system (specifically, by eliminating the atmosphere), which makes conditions more similar to those on the Moon than on the Earth. Including the atmosphere allows us to assign a temperature increase of some 33ºC to the GE. If we really think about it, we realize that is a spectacular amount, especially when compared to what is thought to be the temperature oscillation associated with geological eras or abrupt climate changes. None of these is even half the amount indicated for warming due to natural GE (Masson-Delmotte et al. 2005).

With other simple models—though less simple than the one described above—it is possible to calculate the distribution of the equilibrium temperature for different latitudes on Earth, to establish elemental considerations about the clouds' role, and to determine other potential climates when all the ice has melted, or when the Earth is totally covered with ice, as well as the transitions between such states, and so on.

One advantage of simple models, compared with more complex ones, is that they allow us to carry out a large number of different experiments—by changing some of the conditions of the simulation—because they need much less time to resolve the equations than more complex models.

Climate projection for the future

It is important to emphasize that climate models are the most important, if not the only, tools for carrying out simulations of the planet's climate. In order to be able to use them with any sort of guarantee, experiments have been carried out to reproduce the present climate and the past climate, and to explain the climate change being experienced by the Earth. Since the basic equations come from physical laws and the simulation is realistic, there is great confidence in the use of such models. Clearly, there are still aspects to be discovered with regard to how the Climatic System functions, and this lack of knowledge produces uncertainty. Nevertheless, by accepting the results of the simulation when they are verified by observation, we are indicating that the knowledge we already possess about how that System works is sufficient, and what is still unknown would not be able to substantially modify the simulations. If that were not the case, that is, if our ignorance implied important consequences for such simulations, research would already have detected it.

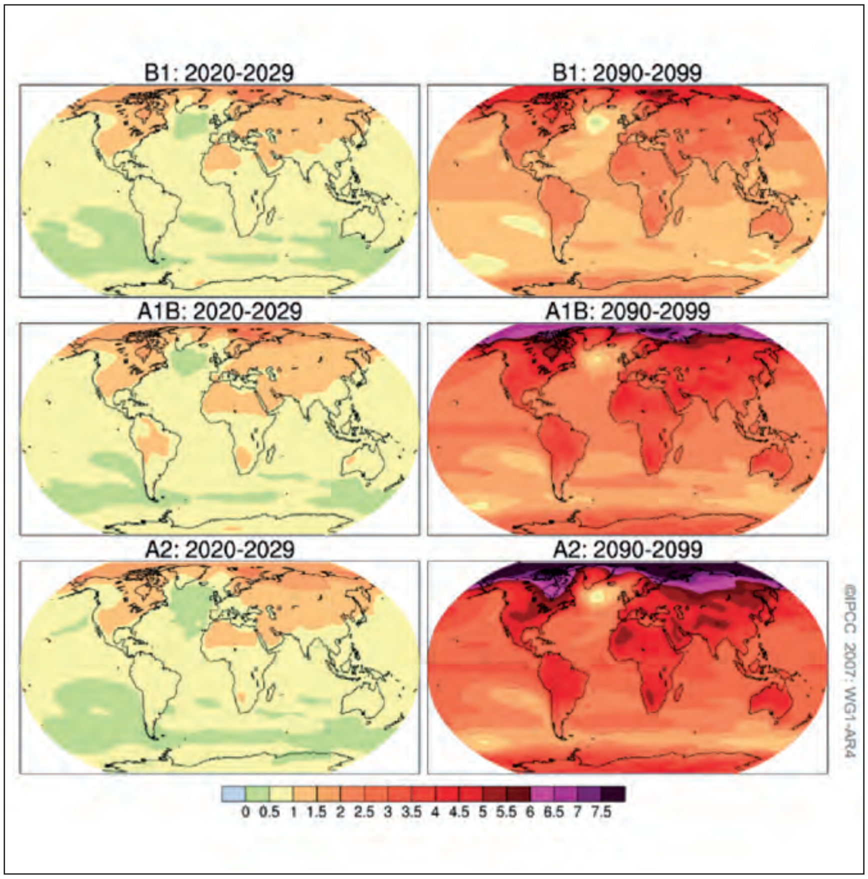


Figure 6. Projected surface temperature changes for the early and late 21st century relative to the period 1980 to 1999. The left and right panels show the AOGCM multi-model average projections (°C) for the B1 (top), A1B (middle) and A2 (bottom) SRES scenarios averaged over the decades 2020 to 2029 (left) and 2090 to 2099 (right).

That being said, it should be clear that simulation of the present climate is not the same problem as simulation of the future climate. In the first case, we know what changes took place in the past, leading up to the present. We know how radiation intercepted by the Earth has changed, and we know how the atmospheric composition has changed—not only with regard to the concentration of GHGs but also, for example, to volcanic eruptions. The forcing of models with real and known conditions has made it possible to reconstruct the present climate. But from now on, we do not know what the conditions of the Earth's atmosphere will be, yet that knowledge is imperative if we are to simulate the future climate.

We know from the past, for example, that annual emissions of CO_2 of fossil origin have increased from an average of 6.4GtC[7] per year in the nineteen nineties, to 7.2GtC per year between 2000 and 2005. These emissions, along with those of the past, have partially determined the concentration of CO_2 in the atmosphere, just as other processes have done with other GHGs. The problem of determining the concentration of GHGs on the basis of emissions is not a simple one: it is necessary, once again, to resort to simulation using models. In this case, they are models of the cycles of carbon and other elements. It is necessary, for example, to take into account how carbon is fixed in the soil and in the seas ("carbon sinks"), which in turn depends on many factors.

Supposing that this problem is resolved, it will still be necessary to know how future GHG emissions evolve. What will definitely be clear by now is that this depends on many conditions, most of which are fundamentally socioeconomic in character and difficult to determine. In response, work is being done with different plausible hypotheses generally called scenarios. Ever since the earliest IPCC reports (FAR and SAR), attention has been paid to defining emissions scenarios, which were initially included in the reports themselves. Following the second report, however, specific work on scenarios was commissioned (IPCC 2000), which generated those scenarios currently being used to project the climate into the future. They are called SRES, an acronym that reflects the character and title of the work: Special Report on Emissions Scenarios.

In a nutshell, work is being done with four storylines (A1, A2, B1, and B2) conditioned by "forces" such as population, economy, technology, energy, agriculture, and soil use. In A1 and A2 more weight is given to economic growth, while in B1 and B2 environmental aspects take the fore. Also, whereas A1 and B1 project on the basis of a globalized world, A2 and B2 emphasize regional and local solutions. Each of these lines generates different scenarios, making a total of 40.

[7] GtC are gigatons of carbon, that is, a thousand million tons of carbon.

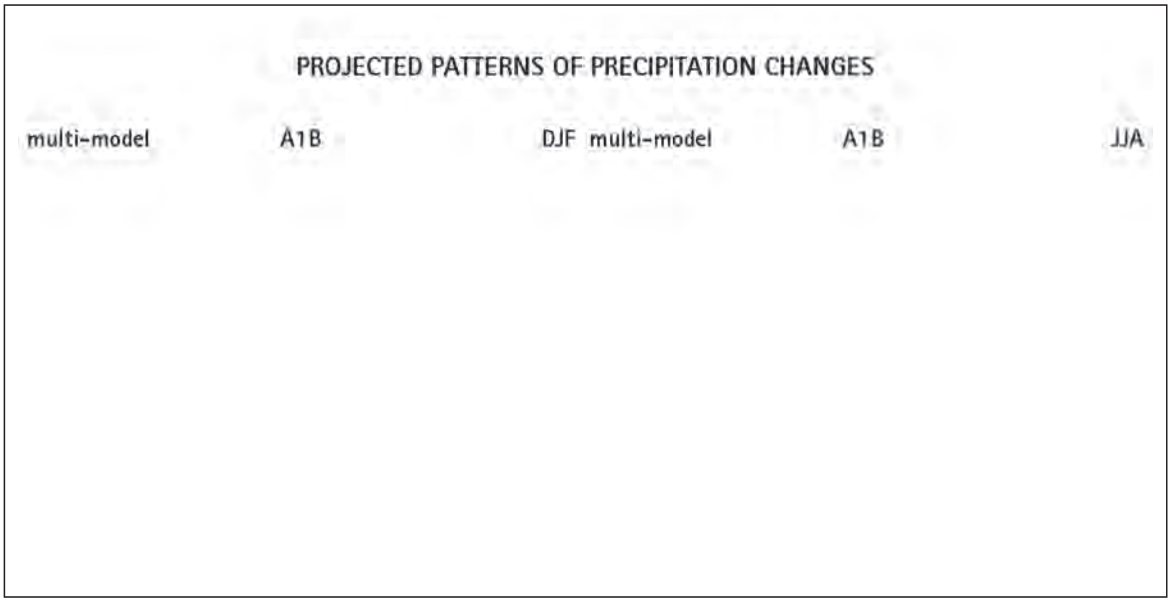


Figure 7. Relative changes in precipitation (in percent) for the period 2090-2099, relative to 1980-1999. Values are multi-model averages based on the SRES A1B scenario for December to February (left) and June to August (right). White areas are where less than 66% of the models agree in the sign of the change and stippled areas are where more than 90% of the models agree in the sign of the change.

Normally, these are organized as families, coinciding with the name of their lines, except for A1, which has the following breakdown:

- A1FI, with intensive use of fossil fuels,
- A1T, with use of non-fossil energy sources,
- A1B, with a balanced use of different sources.

Clearly, we do not know what path humanity will take from here on, so all of those scenarios are considered equally probable.

Each of these SRES emissions scenarios is associated with concrete GHG emissions values over the course of the twenty-first century. Then, using adequate models, future concentrations of GHG are deduced, and the future evolution of those concentrations allows us to project the climate into the future, thanks to climate simulation models. The result is a group of climate projections for each of the SRES being considered. Because they differ from the climatic conditions set as reference, they lead to different future scenarios of climate change. Those scenarios or projections can be global, or limited to specific regions of the world's geography.

The left panel of Figure 5 shows the evolution of GHG emissions during the twenty-first century. The figure includes emissions of all GHG in what is called equivalent CO_2. In calculating this, account is taken of the same GE intensification effect as all the GHGs being considered. As well as all the SRES scenarios described above, results are given here for other scenarios that appeared after the Special Report (IPCC 2000) was published. These modify the contribution of certain "forces" that affect the storylines being considered. The right panel shows projected mean surface temperatures for various families of scenarios and the projection that would correspond to no increase of GHG over the amounts registered in the year 2000. It should be pointed out that, even if that were the case, the temperature would continue to rise, though at a much slower rate.

An analysis of projections for the first two decades of this century offers results that depend very little on the scenario being considered and the model employed (0.2ºC per decade). That is not the case, however, for the final decades of the century; they strongly depend on the scenario being considered, and also on the model employed. For example, the mean multi-model estimation for scenario B1 at the end of the century is 1.8ºC (probably with a rage of 1.1ºC to 2.9ºC), while for scenario A1FI, it is 4.0ºC (probably with a range of 2.4ºC to 6.4ºC), which is always higher than the mean for the period from 1980 to 1999. Note that those values are far above those observed for the increase in mean surface temperature during the twentieth century.

These temperature projections have been used to evaluate the effect on the global average sea level (including the contribution of ice melting in Greenland and Antarctica as well). The rise at the end of the twenty-first century—depending on which scenario is chosen, of course—would lie between a minimum of 0.18 to 0.38 meters for scenario B1 and a maximum of 0.26 to 0.59 meters for scenario A1FI. Those values are relative to the global average sea level between 1980 and 1999.

The AOGCM models make it possible to carry out global climate projections in which spatial and temporal variability is apparent. AR4 includes many such projections (see IPCC 2007, chapter 10), only a few of which are presented here. Figure 6 shows maps of multi-model mean surface temperatures with a clear predominance of values in the arctic region, where the temperature could increase by more than 7ºC by the end of the century. In general, the projected warming for the twenty-first century is expected to be greater over land and at higher latitudes of the northern hemisphere, and lesser over the South Seas and part of the North Atlantic.

Figure 7 shows projections for seasonal rainfall. While global amounts are expected to rise, in the majority of terrestrial sub-tropical regions they will probably decrease. In the upper latitudes, precipitation will probably be greater.

Projections for other important aspects of the climate have also been obtained. Generally, it could be said that all of them continue the tendencies observed in the twentieth century but most show increases in those tendencies.

Special mention should be made of the melting of ice in Greenland, although the time scale is more than a century. Some 125,000 years ago, the temperature in the North Atlantic zone remained higher than at present for a prolonged period of time. The reduction of the ice mass led the sea levels to rise between 4 and 6 meters. Now, if the temperature remained between 1.9 and 4.6ºC higher than pre-industrial levels for at least a thousand years, melting Greenland ice would cause a rise in planetary sea levels of 7 m.

One of the most important applications of climate projections is the analysis of the consequences of climate change or, as it is generally called, the impact of climate change. This has considerable social importance because its effects are local. In order to determine them, it is necessary to have climatic projections with much greater resolution than those offered by global models. This is done with different methodologies and is generally called "downscaling." One of the most common employs regional-scale models nested in global models and run in a linked, simultaneous way. That is dynamic downscaling. Another possibility involves using empirical statistical relations that have been determined for the present climate—they are supposed to remain valid in the future—in order to gain resolution on the basis of future climate projections obtained with AOGCM. There are also methodologies that combine the two mentioned above. More information on downscaling is available in chapter 11 of AR4 (IPCC 2007).

Conclusions

During the Anthropocene, planet Earth is experiencing a change of climate that, strictly speaking, has no precedent in the past. The burning of fossil fuels and general human activity have modified the atmosphere's composition, increasing the concentration of GHGs to levels never before attained, at least in the last 800,000 years. The GE, which has allowed life to exist on Earth, is being intensified anthropogenically, leading to an increase in the global mean surface temperature in the twentieth century that has no antecedents, at least in the last 16,000 years. Along with this change of temperature, a rise in sea levels has also been observed, as well as the reduction of snow coverage on the continents, and sea ice in the Arctic Ocean. Moreover, climate patterns are changing, including rainfall, NAO, and the phenomenon ENSO, among others. The frequency with which certain extreme phenomena occur is also changing.

If GHG emissions continue at the current rate, observed climate change will accelerate in the present century. In fact, even if the concentrations of those gases remained at their current levels, temperature increases and the resulting effects would continue to occur for decades, though with lesser intensity.

The social and economic consequences of the changes observed have already become significant in some zones (changes of habitat, exhaustion of certain species' capacity to adapt, modification of crop seasons, problems with water resources, changes in the distribution and occurrence of certain diseases, and so on), but it is believed that they will become even more significant as warming intensifies. From a human standpoint, the most disadvantaged societies, with the lowest levels of development, will be the most vulnerable.

Global warming can no longer be stopped; we are already suffering the consequences of what we began with the Industrial Revolution. It is clear that we have to reduce emissions, and that is intrinsically good for the environment in general. But we must also strive to adapt to the coming climate and understand that, beyond living with a certain level of risk, it will be necessary to face the cost of adapting. At any rate, that will be much less than the cost of doing nothing. Policymakers have to play their role and the society also the own. Obviously, as members of society, scientists, too, must participate. Research must be intensified, eliminating doubts, improving climate projections, offering clues as to how to reduce climatic vulnerability and risk, and seeking out more efficient means of energy use, less contaminating systems, and so on.

We will undoubtedly have to make some slight changes of lifestyle so that developing countries can attain an adequate level of wellbeing. The future humanity expects nothing less of us.

Bibliography

Álvarez, L. W., W. Álvarez, F. Asaro, and H. V. Michel. "Asteroid Extinction Hypothesis." *Science* 211, 1981, 654–656.

Arrhenius, S. "On the influence of carbonic acid in the air upon the temperature on the ground." *Philos. Mag.* 41, 1896, 237–276.

Berger, A. "Milankovitch theory and climate". *Reviews of Geophysics* 26 (1988): 624-657.

Croll, J. *Climate and Time in Their Geological Relations: A Theory of Secular Changes of the Earth's Climate*. 2nd ed. New York: Appleton, 1890.

Crutzen, P., and E. F. Stoermer. "The 'Anthropocene.'" *Global Change Newsletter* 41, 2000, 12–13.

Eddy, J. A. "The Maunder Minimum." *Science* 192, 1976, 1189–1202.

Hoyt, D. V., K. H. Schatten, and E. NESMES-RIBES. "The hundredth year of Rudolf Wolf's death: Do we have the correct reconstruction of solar activity?" *Geophys. Res. Lett.* 21, 1994, 2067–2070.

IPCC. *Special Report on Emissions Scenarios.* Nakicenovic, N., and R. Swart (eds.). Cambridge and New York: Cambridge University Press, 2000.

IPCC. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Da, K. Maskell, and C. A. Johnson (eds.). Cambridge and New York: Cambridge University Press, 2001.

IPCC. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller (eds.). Cambridge and New York: Cambridge University Press, 2007.

Kiehl, J., and K. Trenberth. "Earth's annual global mean energy budget." *Bull. Am. Meteorol. Soc.* 78, 1997, 197–206.

Köhler, P., F. Joos, S. Gerber, and R. Knutti. "Simulating changes in vegetation distribution, land carbon storage, and atmospheric CO_2 in response to a collapse of the North Atlantic thermohaline circulation." *Clim. Dyn.* 25, 2005, 689–708.

Loulergue, L., A. Schilt, R. Spahni, V. Masson-Delmotte, T. Blunier, B. Lemieux, J.-M. Barnola, D. Raynaud, T. F. Stocker and J. Chappellaz. "Orbital and millennial-scale features of atmospheric CH_4 over the last 800,000 years." *Nature* 453, May 15, 2008, 383–386.

Lüthi, D., M. Le Floch, B. Bereiter, T. Blunier, J.-M. Barnola, U. Siegenthaler, D. Raynaud, et al. "High-resolution carbon dioxide concentration record 650,000–800,000 years before present." *Nature* 453, May 15, 2008, 379–382.

Masson-Delmotte, V., A. Landais, N. Combourieu-Nebout, U. Von Grafenstein, J. Jouzel, N. Caillon, J. Chappellaz, D. Dahl-Jensen, S. J. Johnsen, and B. Stenni. "Rapid climate variability during warm and cold periods in polar regions and Europe." *Comptes Rendus Geoscience* 337, 2005, 935–946.

OCLIB. "Informe de seguimiento del convenio Universitat de les Illes Balears-Conselleria de Medi Ambient del Govern de les Illes Balears, sobre el Observatori del Clima de les Illes Balears." Unpublished technical report. Grup de Meteorologia, UIB, 2007.

Ruddiman, W. F., ed. *Tectonic Uplift and Climate Change*. New York: Plenum Press, 1997.

Thompson, D. W., J. J. Kennedy, J. M. Wallace, and P. D. Jones. "A large discontinuity in the mid-twentieth century in observed global-mean surface temperature." *Nature* 453, May 29, 2008, 646–649.

Wang, Y. M., J. L. Lean, and N. R. Sheeley. "Modeling the sun's magnetic field and irradiance since 1713." *Astrophys. J.* 625, 2005, 522–538.

Yang, F., and M. Schlesinger. "On the surface and atmospheric temperature changes following the 1991 Pinatubo volcanic eruption: a GCM study." *J. Geophys. Res.-Atmos.* 107, April 2002: doi10.1029/2001JD000373.

Acknowledgements:
My thanks to Professor C. Ramis, whose suggestions have considerably improved the quality of this text. The figures are exact reproductions of those appearing in IPCC documents and are included here with their tacit authorization.

the economy of the firm

VICENTE SALAS FUMÁS

Introduction

The firm is a central institution in the functioning of any economic system in which people meet their needs through the division of labor, cooperative production, and the exchange of goods and services. As part of the system, firms serve to produce goods and services for sale on the marketplace, a necessary function allowing each person to combine specialization in work with the satisfaction of his or her multiple needs. Firms take the form of a legal entity with its own trade name. The heterogeneity of firms with regard to size, the variety of goods and services they offer on the market or the activities and resources they control internally, awakens intellectual interest on the part of the social sciences in general and economics in particular. Why they exist, what their nature is, how they are structured and function internally, and what factors influence their changing nature over time—all these are questions addressed by economic research into firms.[1]

Firms arise from the decisions of people, firm executives, who also direct the assignment of resources within the scope of their responsibilities. In complex firms, oversight of resources, generically known as "management," has to be shared by numerous specialists, leading to a professional setting of considerable quantitative and qualitative importance in developed societies. Alongside the study and *positive knowledge* of firm reality carried out by economics and other social sciences, there is also a *normative knowledge* of decision making and management practices, which is included in the training of firm executives and managers. The existence of centers specialized in training professional management personnel throughout the world bears witness to the importance of this knowledge.

Thus, there are two large knowledge bases about firms, both of which are characterized by a dynamic generation and renewal of contents that makes them separate but also interdependent: one of these bases is linked to the *why* of the phenomena being studied (positive analysis), and that is the area on which the social sciences have concentrated. Its ultimate goal is to learn about the consequences that one firm's reality or another will have for social wellbeing. The other knowledge base concerns *how* to act in the event of specific problems (normative analysis) and this is the area occupied by the disciplines of professional management, whose ultimate goal is to contribute to the specific wellbeing of those making decisions

[1] DiMaggio (2001) and Roberts (2004) offer an integrated view of the recent past and future of firms from different disciplines—sociology and economics—that are, nevertheless, complementary. See also: Malone et al. (2003) and Salas Fumás (2007).

within a firm, especially increasing private profits.

A summery of the main aspects of positive and normative knowledge of firms and their management is beyond the scope of the present text, not only because of the multitude of questions involved, but also because of the diversity of academic disciplines interested in them. Therefore, we will limit ourselves here to that part of positive knowledge mostly attributable to economic research into firm theory. Thus, the present text will be organized in the following manner: the first section outlines antecedents in the field of general economic interest and the firm's place in this setting. The second deals with research into a firm's limits or delimitation in the market to which it belongs. The third section reviews advances in the economics of internal organization of firms, while the forth addresses questions such as the legal identity of a firm and the social relations that link firm economics with other social sciences. In our conclusions, we will evaluate this theory's contributions to firm management and society's expectations with regard to good *performance* by firms.

Antecedents and the general framework of firm economics

In what are called *market economics,* the relations between firms, or between firms and their consumers, workers, investors, and so on, are regulated by prices that indicate the relative value of resources available in alternative uses when the needs that must be met outstrip available means. Market economics customarily include the institution of private property so that price is the monetary reward to whoever produces or sells what others demand. "Market" is also synonymous with free firms, which means equality among citizens when deciding to create a new firm and participate with it in the offer of goods and services, shouldering the consequences of that decision (financial sufficiency). The production of goods and services for sale on the marketplace is therefore made, in most cases, in competitive conditions, that is: allowing the possibility of choice to all those agents related to the firm, especially those who purchase and pay a price for its products. Competition creates pressure for continuous improvement and innovation as responses intended to insure survival and the obtaining of rewards commensurate with the resources employed in firm activity. It thus seems realistic for economists to analyze the *raison d'être* and existence of firms on the basis of their efficiency. In other words, the existence of firms, their nature and internal organization—which we observe and seek to explain with positive analysis—correspond to the goal of obtaining the best possible adaptation to the laws of competition that favor the creation of wealth (the difference between value or utility and cost of opportunity).

The key role of price in coordinating (identifying imbalances between supply and demand) and motivating competition (rewarding those who respond to those imbalances by producing more of what has the highest price) among members of a given social collective converts economic theory into a theory of prices and markets. For a long time, there was hardly a place in this theory for the economic study of firms beyond their consideration as one of several elements in the mechanism of markets, where they serve to make pricing possible. In fact, prices arise from the meeting of supply and demand, and in order to explain the creation of prices, it in necessary to identify the suppliers and demanders who compete in that market. That is where firms find their place. This concept of the role of firms in the market economy was so mechanistic and instrumental that they were described as "black boxes," in keeping with the absolute indifference with which economics considered their *raison d'être* and nature.

While the academic discipline of economics was contemplating them with indifference, firms were nevertheless gaining presence and visibility in society. Especially, they were growing larger and more diversified in the forms they adopted for their internal functioning. The division of labor became so much a part of their inner workings that, beyond functions and tasks directly related to production, posts were also created to oversee the assignment of resources—a function that market logic had assumed to be carried out by the system of prices. Administrative functions within firms are complex enough that the persons carrying them out seek professional training beforehand. *Business schools* have been created to respond to the training needs of business management (one of the most prestigious, the Harvard Business School, is now celebrating its one-hundredth birthday).

A sort of specialization has thus arisen that separates economics, as an academic discipline dedicated to the study of how markets function and prices are created, and professional business administration schools, which are dedicated to meeting the demand for trained specialists in management positions. Teaching and research into management has thus become the area for studying specialized administrative functions in firms, from personnel directors to general management, including finances, marketing, and operations. At first, this training revolved almost entirely around case studies and teachers' personal experiences. The situation changed in the 1960s when a report on the teaching of business administration in the US, commissioned by the Carnegie Corporation and the Ford Foundation, recommended that universities base their teaching of this subject on rigorous academic research, especially economics and behavioral sciences.[2]

In response to this recommendation, business schools broadened their teaching staff to include academic economists, along with professors and researchers from other scientific, technological, and social disciplines. At the same time, firms and management processes became the subject of growing intellectual interest. Research into firms took shape and gained substance, receiving contributions from a broad variety of academic disciplines. Economics is one of those disciplines, and economic research has a growing interest in firms themselves, without the need to subordinate that interest to the study of how markets function. This work has posed intellectual challenges to academic economists researching firms and has begun to receive attention. An article on the nature of firms, published by Ronald Coase as far back as 1937, was ignored until much later in the twentieth century. Coase considers the existence of firms, their internal nature, and the firm director's authority, as an anomaly in economic thought that reveals the great advantage of the marketplace and the system of prices in organizing economic activity. Coase asks: If the market and prices are so effective in their functions, why are there firms in which resource management is not carried out on the basis of prices but rather according to the orders and authority of managers?

Orthodox economics have always recognized the limitations or failures of the market to harmonize individual rationality (private profit) and collective rationality (social wellbeing) in specific contexts. But in those "situations of disharmony, economic policy's normative prescription calls for intervention by the state in order to reconcile conflicting interests. Coase warns that the market's limitations or failure to direct (coordinate and motivate) the processes of resource assignment cannot always be resolved by state intervention. When possible (that is, when legislation and transaction costs allow it) institutions will arise in the private sector (ways of directing resource assignment that are not based on market prices) that help overcome the market's limitations without direct intervention by the state. To Coase, firms exemplify an institution that arises in the private sector when the coordination of resource assignment is most efficient if carried out by the visible hand of the firm director rather than by the invisible hand of the market. Firm and market switch roles to organize exchange, exploiting comparative advantages and suggesting an institutional specialization in terms of relative comparative advantage.

Over time, economics' contribution to the study of firms has defined two fields of interest that remained separate until just a few years ago. One is the interest in explaining the limits *of firms,* while the other seeks to explain their *internal organization.* The limits of a firm have been defined horizontally and vertically, while their inner workings are viewed in terms of problems of coordination and problems of motivation. The study of the *horizontal limits* of firms has concentrated mainly on explaining the size of a firm in terms of its volume of production (or use of resources needed for that volume, including, for example, the number of workers employed). This explanation relies, fundamentally, on two predetermined variables: the efficient scale of production, and the size of the market. If the market is sufficiently large, competitive pressure will force firms to converge toward a size close to the scale that insures a minimum of production costs (efficient scale). Differences in production that minimize unit costs (differences in production technology and degrees of presence of growing returns to scale) explain the heterogeneity of firm sizes. When market size is small with relation the efficient scale, one can expect the market to be dominated by a single firm, in what has come to be known as a natural monopoly. From a dynamic perspective, a change in the horizontal limits of a firm can be explained by changes in technology or market size.

The study of the horizontal limits of firms is part of the broader neoclassical theory of production, in which production technology is summed up as a function that represents the most advanced technological knowledge available at the time being studied, in order to transform resources into goods or services of greater value or utility. This representation of technology and the price of resources is used to derive the functions of unit cost and supply mentioned above. When carefully studied, the theory of production explains the size of production units (production plants) but doesn't explain the influence of business administrators, which is what defines the perimeter of a firm, according to Coase. That theory fails to explain why some firm executives direct a single production plant while others direct several. The study of the limits of a firm and its internal organization—firm theory—includes contractual considerations such as availability of, and access to, information, and the capacity to process it, as well as

[2] Handler (1962, 1977) offers historical documentation of the growth of large firms in the US in the early twentieth century, and of the development of the professional manager. A view of the role of business schools in the training of professional directors can be found in Kochan and Schmalensee (2003), a book written on the occasion of the fiftieth anniversary of the Sloan School at MIT.

the merely technological considerations postulated by production theory. In a nutshell, figure 1 orders and sums up, on the basis of time and thematic areas, the main contributions of firm theory from a contractual perspective, in the broadest sense. This covers the rest of the materials that have drawn the interest of the economic theory of firms.

Vertical limits

The study of the *vertical limits* of firms is directly tied to Coase's observations regarding the co-participation of markets and firm executives in the coordination of economic activity and the assignment of resources. The limits of a firm coincide with the authority with which a firm director is able to direct the assignment of resources, while the market determines coordination among firms. How many resources a firm director can control, and how much activity his firm can carry out and place on the market, depend on the relative efficiency of a mechanism that coordinates one or the other. That efficiency is defined by comparing respective transaction costs. An important part of knowledge about firms' vertical limits revolves around determinants of transaction costs from a comparative perspective: first, firm verses market, and then, firm, market, and intermediate forms of organization that include non-standard contracts in relations among firms (long-term contracts, sub-contracting, franchising, alliances, joint ventures, and so on). Early research (Arrow 1969, Williamson 1975–1985; Klein, Crawford and Alchian, 1978) concentrated mostly on those attributes that facilitate an *ex ante* prediction, in terms of transaction cost, of comparative advantages when resources are directed by either a firm director or the market. Uncertainty and asymmetrical information among those involved in such an exchange, as well as the specificity of assets (of various types) invested in transactions are the attributes that must, according to this theory, be most clearly discerned when explaining the vertical limit of firms. Empirical evidence supports those conclusions.

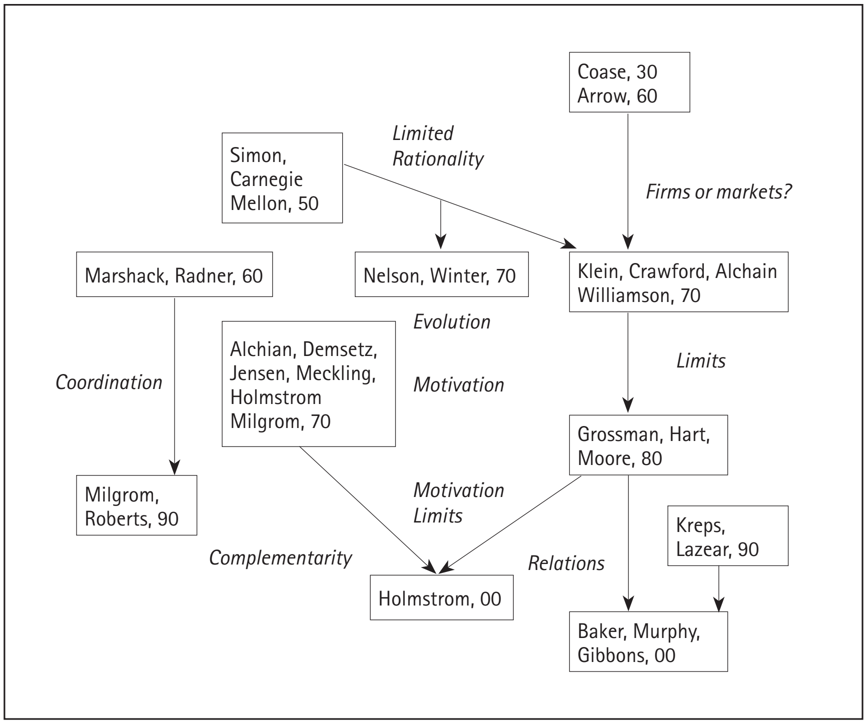


Figure 1. Contributions to the Economic Theory of Firms. Source: Salas Fumás (2007).

Given that specificity of assets and asymmetry of information are conditions that relatively favor the use of a firm rather than the market, and given that they occur in a very high portion of economic transactions, the TTC (theory of transaction costs) tends to predict a leading role for firms in the direction of resources in a higher percentage than is actually observed. On that premise, beginning with the work of Grossman and Hart (1986) and especially Hart and Moore (1990), the theory of property rights (TPR) offers considerations about brakes to vertical expansion by firms on the basis of an identification of the source of transaction costs to the firm, which are ignored by TTC. Particularly, the TPR emphasizes how the key to defining the limits of a firm is the distribution of ownership of non-human assets. In that sense, the TPR sees the definition of firm limits in term of the assets it owns. One important implication of this view of firms is that, since people cannot be owned—excluding slavery—workers are outside the limits of a firm.

When a firm expands its presence as a coordinating mechanism (taking on more activities under the direction of management), it is generally increasing the amount of non-human assets it owns, to the detriment of ownership by others outside the firm. Supposing that non-human assets are complementary with those specific assets that result from the investment in human capital by people; when a firm increases its control of non-human assets it is decreasing incentive to invest in human capital on the part of those who lose those assets that they previously owned. This opportunity cost of expanding the limits of a firm with more assets justifies the TPR's prediction that the distribution of ownership of non-human assets will be spread among a greater number of firms than is predicted by the TTC.

The TTC emphasizes the transaction (transfer between technologically separable units) as the basic unit of analysis whose inclusion in, or exclusion from, the perimeter of a firm is decided in the margin. The TPR, on the other hand, emphasizes decisions about assigning the property of non-human assets as a determinant of firm limits. In both cases, one detects a certain distancing with respect to the view of united management adopted by Coase in referring to the nature of firms. One way of reconciling these differing approaches is to include a contractual viewpoint in the analysis. Ownership of assets and the hierarchy of authority attributed to the firm and its director are economically significant to the degree that transaction costs require the regulation of transactions through incomplete contracts, that is, contracts that define the general framework of relations among agents even though there are many contingencies for which particular responses have not been predetermined.

Ownership of assets implies the capacity or power to decide about their use in any sense not previously dictated by contract, while the authority that Coase attributes to firm directors is fullest when job contracts are incomplete and the director has the contractual right to direct (order) his workers. Simon (1957) was a pioneer in identifying the economic importance of incomplete contracts in relations between firm and workers although he did not relate his discoveries to the authority of contractual origin proposed by Coase. In a world where all relations among persons were regulated by complete contracts (where anything that could happen in each relation was predetermined), the ownership of non-human assets and the authority of firm executives would be irrelevant since there would be no residual rights to decision, which do indeed exist when contracts are incomplete.[3]

The fact that contracts have to be incomplete (to avoid excessive transaction costs) along with the repetition of relations among agents, opens the path to another type of contract whose viability and efficacy affect decisions about the limits of a firm: implicit contracts. In effect, implicit contracts are those in which agreement between parties is based on the expectation of good faith in the other party's conduct and mutual trust. Trust more easily emerges in interpersonal relations in which the power to make decisions is shared by all agents involved, rather than in situations where decision-making is more one-sided. In order to exploit the advantages of implicit contracts (in terms of low transaction costs), a decision about the limits of a firm (in terms of the assignment of ownership of non-human assets employed in production) will largely depend on a certain distribution of ownership. This distribution will serve to balance power and strengthen trust, forming a basis for relations among agents in the successive stages of the chain of production (Baker, Murphy, and Gibbons 2001).

Internal organization

Economic research has separated the study of firm limits from the study of what occurs within a firm with given limits. To a high degree, the study of the inner workings of a firm has more-or-less explicitly concentrated on an analysis of the manner in which firm directors assign resources, as well as the directors own *raison d'être.* The internal organization of a firm (the order that stems from the internal division of work, the exchange of information and the distribution of power in decision making) is achieved through the coordination (determining what each person is supposed to do) and motivation (being interested in doing it) of the persons involved. The difference with respect to a market is that, in a firm, the manner of approaching coordination and motivation is influenced by the firm director's capacity to intervene.

Economic analysis begins by separating the study of coordination problems from the study of motivation problems, although both play a significant part in a very realistic concept of firms and their management: the *team* concept. In order to study coordination, the concept of *teamwork* is postulated. This is defined as collective action in which all involved persons share the same goal, which also represents the group's interests. For example: maximizing shared wealth. While each person acts for the general good, what he or she must decide or do to achieve maximum efficiency and results depends on the information and actions of the other team members (interdependence). In this context, coordinating individual action means influencing individual decisions (mainly by exchanging information) in order to harmonize them in the context of the existing interdependence, or else to change that interdependence in order to thus influence the needs of coordination itself (Marshack and Radner 1972; Milgrom and Roberts 1995). In team organizations, coordination becomes a relevant problem because there is a considerable cost involved in producing and transmitting information. Thus, the solutions to coordination problems discussed in articles, as well as those applied by firms, have had much to do with advances in information technology.

On the other hand, production or *team technology,* refers to complementarity between resources belonging to different people who cooperate in the production process so that the joint exploitation of technology creates more potential wealth than individual exploitation would. Alchian and Demsetz (1972) place this characteristic of technology at the origin of firms as we know them, analyzing the collective functioning of team production with the

[3] The empirical referent of the theory on the limits of a firm is made up by the firm's decisions to make, buy, or establish alliances with third parties. In recent years, terms such as *outsourcing* and *offshoring* have been added to characterize the relocation of activities of a firm's value chain on global markets. The theory of firm limits has fully entered studies of the multinational firm (Helpman 2006).

opposite suppositions of team organization. That is, that people who contribute complementary resources to production do so with the expectation of obtaining a maximum net individual reward regardless of the collective interest. Technology and teamwork lead only to coordination problems; team technology plus overlapping individual interests (also called coalition) lead to coordination problems and also to other motivation problems, which have attracted the greatest academic interest.

Team technology impedes a consideration of joint production as the sum of individual production by those who participate in it. The individual benefits that lead to participation in collective action and motivate the contribution of resources by each agent involved can only be defined in terms of joint production and contribution of resources, if these are observable. In principle, compensating participation in collective action with participation in joint output has the advantage of needing to measure only one variable. Nevertheless, it has disadvantages known as stowaway behavior (Holmstrom 1982). The alternative to measuring resource contribution (quantity and quality) demands specialization in that task and a capacity to address the question of how to insure the efficiency of whoever is monitoring the process. Alchian and Demsetz offer an organizational solution in which the monitor bilaterally contracts each participant in a collective action, agrees on a compensation commensurate with the amount of resources he or she has contributed, acquires the right to supervise his activity and direct his work, and retains the difference between what is produced and what he has agreed to pay, as his own compensation. In sum, firm theory gives economic meaning to capitalist firms as we know them; in which firm directors centralize contracts and act as supervisors and coordinators in exchange for residual profits.

The paradigm of team production, the bilateral character of contracts, and residual income (profit) for the firm director are the basis for successive contributions to firm theory, which characterize it as a nexus for contracts. Developments since the 1970s have taken account of the supervisor's limitations when precisely measuring the quantity and quality of resources, as well as the unequal assignment of risks implied by retribution in the form of profits when the result of collective action depends in chance factors as well as on the resources employed and the available technology. This has led to new contributions in the organizational design of firms, such as: 1) determining the number of hierarchical levels of supervision and control (Calvo and Wellisz 1978; Rosen 1982); 2) efficiently assigning risks (for example, by creating separate collectives for the functions of director of resources, coordination, and motivation, and the functions of risk assumption, which led to the complex capitalist firm, or corporation (Jensen and Meckling 1976); 3) determining the optimal distribution of decision-making power—centralization versus decentralization (Agnion and Tirole 1997; Alonso, Dessein and Matouschek 2008); 4) designing complex incentive systems to stimulate effort that cannot be observed by a supervisor—agency theory (Holmstrom 1979, 1982; Holmstrom and Milgrom 1987, 1991).[4]

With all of this work on the theory of firm limits and internal organization, the concept of a firm draws away from the idea of production and becomes that of a structure that governs the process of assigning resources, processing information, assigning decision-making power, evaluating work and awarding recompense.

The firm as a mini-economy or community of persons

The economic theory of firms has unwittingly used the terms "firm" and "firm director" almost interchangeably. A careful reading of the works of Coase, and Alchian and Demsetz, reveals that they are really explaining the existence of a firm director who carries out concrete functions in the general framework of the specialization and division of work. Coase's firm director directs resources with coordination, giving orders and taking advantage of his central position in the network of contracts with other owners of resources wherever the market and prices face imbalances in the economies of scale. Alchian and Demsetz's supervising firm director is needed to measure the quantity and quality of resource contributions that feed production with team technology. Contractual approaches inspired by those authors, which cast the firm as a nexus for contracts, do not explain whether that nexus is the firm director in person, or an undefined entity called the "firm." It can thus be reasonably stated that many firm theories proposed by the field of economics are really theories about firm directors. The figure that emerges from this literature complements Schumpeter's view of this social agent as the protagonist of innovation and creative destruction.

The best way to separate the firm from its director is to consider a firm as a legal entity recognized by law as a legitimate party in contracts and the capacity to own property. The common nexus among contracts, which is identified with firms, is generally a legal entity that enters into bilateral contacts with

[4] In reviewing this literature, we find that, rather than configuring a theory of firms, it defines a field of study: the economics of information. Many of the problems of transactions in conditions of asymmetrical information that are analyzed with models adverse selection or moral risk are not limited to firms; they also occur in the domain of markets. Therefore, agency theory is not firm theory, but rather a theoretical framework for the study of problems of moral risk, some of which are part of the problems of internal organization faced by firms.

the different agents with whom it has relations. In order for a firm director to be able to coordinate and motivate people within the firm, its contract with them must include the possibility that a third party, also contracted by the firm, can carry out those functions. In the TPR, where the limits of a firm are related to the non-human assets it owns, it is necessary to explain why this property belongs to the legal entity that is a firm, rather than to the physical person that is its director. In short, economic theories about firms will not be complete until they explain why the legal entity of the firm emerges as something different than the physical person that is its director.

Texts about these theories offer various possible answers, all of which are somehow related to the desire to economize transaction costs:

I. The firm, as a legal entity, is not affected by the temporal limitations that affect living people. A longer time span, with no finite demarcations, is relevant to the viability of relations based on reciprocity that sustain mutual trust (implicit contracts) and bring economic value to a good reputation (Kreps 1990).

II. The legal entity, complemented by the variety of legal forms a firm can take when constituted under law, offers the possibility of managing risks, directing firm resources and financing its assets, which would not be possible if people were unable to differentiate between personal assets and firm assets. The intellectual and technical entity of questions posed by a firm's chosen legal form and the response to available options when assigning management and control responsibilities have generated a field of highly relevant studies of modern capitalist firms such as that of *corporate government,* where law and economics merge.[5]

III. By concentrating ownership of non-human assets in the legal entity of the firm, rather than spreading ownership among the different persons who are linked by it, the firm's directors find efficient ways of coordinating and motivating workers in contexts of asymmetrical information that would not be feasible if ownership of those assets were spread among all workers (Holmstrom 1999). With this explanation of the firm as a legal entity that owns assets, Holmstrom combines in a single problem of organizational design, decisions concerning firm limits, the assignment of ownership of non-human assets, and decisions about the coordination and motivation of work in the firm, which depend on internal organization. With this inclusive view of , Holmstrom manages to define a firm as a *mini-economy* whose directors wield solutions for inefficiency derived from problems of asymmetrical information and external effects in a way that resembles how the state wields authority in overall society. There is, however an important difference: a firm is surrounded by markets that offer ways out, limiting possible excesses of power derived from the high concentration of assets that can be accumulated.

As a legal entity that owns assets whose accessibility and conditions of use are decided by its directors, a firm becomes a powerful lever that affects the conduct of those persons who combine their work and knowledge with those assets. Therefore, although it is formally true that a firm does not own human capital—the persons working there are outside its perimeter—its overall functioning is better understood when the workers are considered a part of it. A theory of the firm that includes employees who are combining effort and knowledge with assets belonging to the firm will be much closer to approaches to the study of firms carried out in other social disciplines, such as psychology and sociology. From the very start, those disciplines have considered firms to be a community of persons, minimizing the relevance of its other assets, which is the opposite of what economics has done. Moreover, the link between economics and other social disciplines studying firms becomes more solid when theories employ more relaxed views of the concept of rationality, which have recently emerged in economics and other studies.

The firm as a community of persons

Economic studies of firms are carried out under the premise of human behavior characterized as *rationality:* people know their preferences, and their behavior is coherent with them. Rationality allows academic research to simulate the private and social consequences of specific behavior and restrictions, recommending corrections or adjustments according to the foreseeable results. Nevertheless, economic rationality has been criticized for its unrealistic suppositions and for the aspects of human behavior it leaves unexplained. Williamson (1975), in the book that reconsiders the institutional comparison between market and firm (hierarchy) put forth by Coase forty years earlier, brings criticism of the hypothesis of rationality into firm theory. This criticism was initiated by Herbert Simon and his colleagues at Carnegie Mellon when they realized that economists' suppositions about the capacity to store and process information, implicit in utility maximization, are unrealistic in light of the physiology of the human brain.

This criticism by Simon and his colleagues led to the proposal of an alternative to the supposition of absolute rationality under which economic studies of firms and markets had been carried out. Known as *limited*

[5] Corporate government includes a more general subject, which is that of decisions about firm property (Hansmann 1996). Tirole (2001) offers a view of the problem of governing a firm that fits the contractual perspective adopted in this text.

rationality, this alternative supposition proposes that, while people are intentionally rational, their behavior is affected by the limits of their capacity to store and process information. These limits are as relevant to an explanation of reality as are those restrictions coming from technology. On this basis, the explanation of human behavior includes a supposition of heuristic decision-making, rather than the optimization predicted by the supposition of unlimited rationality. Evolution and adaptation in processes of transit from one equilibrium to another are steady states in the system and cannot be ignored as they have been by neoclassical economics, which only compares situations of equilibrium (Nelson and Winter 1982).

Laboratory experiments and the observation of reality offer evidence about human behavior that does not fit the supposition of consistency and transitivity of preferences associated with the most conventional rationality. They have led to the development of the specialized field of behavioral economics, which emphasizes *prospect* or reference-point theory, Khaneman and Tversky (1979), which questions the theory of expected utility used to analyze a considerable portion of behavior in risk situations when studying firms and markets. Then there is the theory of *social preferences* (Guth et al. 1982), which questions the classic supposition that people only consider their own payment when choosing among alternatives (see Camerer, Loewenstein and Rabin (2004) for a review of this literature). Behavioral economics marks an important reduction of the distance between economics and psychology, and one of the most recent steps in that reduction is the importation from the field of psychology of the concept of "happiness," as an alternative to economics' traditional concept of "utility," when expressing personal preferences (Frey 2008). Classic economics eschews introspection as a way of explaining how people form their preferences, opting instead for the idea of preferences as a means of inferring the utility of proposed alternatives. It draws on the supposition of rationality (consistency and coherence between preferences and conduct). Research into happiness, however, adopts forms of measuring utility developed by psychology—especially neuroscience—with clearly introspective goals. The experimental results of this research indicate that people not only evaluate tangible goods and services available to them through economic income, they also base their concept of utility on less tangible aspects, such as social conditions, relations, the capacity to decide and the possibility of developing their own competence. This leads to the idea of *utility associated with processes,* rather than utility based solely on results, which is predominant in the most orthodox economics.

Behavioral economics are modifying the way in which we analyze how markets function (Frey refers to happiness as a concept that will revolutionize the science of economics). They also help to explain certain particularities of how firms function, which were previously considered anomalies in neoclassical models. For example, one factor regularly observed in firms is the stability in their relations with workers, with long-term contracts and internal job markets (internal promotion is the dominant mechanism to cover job openings, rather than resorting to outside hiring). The stability of such relations is partially due to the impossibility of acquiring on the market, the specific knowledge that can only be acquired through learning routines that arise from the mutual adjustment and evolving adaptation to conditions in a specific context. In that sense, the limits of a firm can be explained on the basis of the need to protect and exchange valuable specific knowledge that constitutes a competitive edge in the marketplace (Teece 1986; Kogut and Zander 1992).

This theoretical work expands the conventional model of preferences and rationality to make room for empirical regularities that appear in a historical moment when the dominant firm model in Japan (organization by processes and in highly autonomous teams oriented towards clients) brought into question firm models generated according to the empirical referent of the dominant firm model in the US (hierarchical firms), especially in light of the proven commercial success of Japanese firms in markets where they compete with American ones. Research into happiness poses a new challenge to firm theory and the management practices therein, in that people's preferences about to how achieve results, as well as the results themselves, make it necessary to evaluate the internal organization (the design of jobs, the autonomy of the people doing those jobs, participation in decision making, mechanisms of socialization) as an end unto itself, rather than just a means to obtain better results.

Conclusion

Academic knowledge about firms is imperative to an understanding of the functioning of the overall economy because what happens inside firms is as important, quantitatively and qualitatively, as what happens among those firms (Simon 1991). It is difficult to develop a theory of firms, even if only in one academic discipline, such as economics, because the ideas and concepts generated by such a process are not at all precise in delineating the concept. In that sense, firms appear to be more-or-less explicitly associated with a technical production unit (plant), with the function of a firm director or a person in charge of directing resources, with a legal entity created under law, or with a community of people. Sometimes, firm theory addresses questions about the determinants of its limits or borders, and other times, questions associated with the solution of inner motivation or coordination problems.

In any case, economic analysis casts firms as entities that function to produce goods and services for the market in conditions of competition and financial sufficiency. To do so, they adopt one or another of the multiple judicial forms dictated by law. One aspect that differentiates a firm from the market to which it belongs is its condition as a nexus for contracts, which allows it to take the place of its director as that nexus. This central situation with regard to contracts helps to avoid multilateral contracts among investors, workers, and clients, which would be necessary in a market solution. The result is a savings of transaction costs. The fact that the nexus is a legal entity facilitates the accumulation of assets and the management of risks, as well as internal management in the face of asymmetrical information and external effects. None of this could be achieved if that nexus resided in a physical person, rather than a legal entity. While each judicial form imposes certain restrictions on how problems of coordination and motivation are resolved (in order to reduce aggregate transaction costs), they all leave enough freedom so that each can adopt solutions most compatible with the characteristics of the transactions in which agents are involved. Moreover, the dominant form taken by firms changes over time and in different countries with similar levels of economic development during the same time period, but this should not lead us to forget that the firm is a human invention and is thus subject to modification and transformation in keeping with technological conditions, including developments in information technology and institutional changes (legal systems).[6] Firm theory seeks to identify basic problems of coordination and motivation that are structurally permanent but allow different solutions in different surroundings and conditions.

Knowledge of firms as an institutional response to problems of exchange and collaboration arising from the division of labor should not be confused with knowledge of business administration (management) that is transmitted to those who hold, or seek to hold, management positions in the world of business. It would seem to be a good idea for the positive knowledge of firms offered by theory to become a part of the normative knowledge about how to run a firm that is taught in business schools, but that is not currently the case. One explanation for this distance between normative and positive knowledge is that firm theory presupposes an absolute rationality of behavior by agents, and its interest lies exclusively in discovering the implications of that individual rationality for collective wellbeing. This explanation fails when decisions and conduct by firm directors are difficult to reconcile with the rationality on which the theory is based. Advances in behavioral economics and the introduction of the concept of happiness in the introspection of preferences and utility serve to bring positive and normative knowledge about firms closer together. That way, firms may no longer be considered instruments or means to obtain the best possible economic results (greater income and consumption), as they are by conventional economic analysis; instead they may be evaluated in terms of the concrete solutions they adopt in the presence of internal problems of coordination and motivation. Means and results should be evaluated together, although one or the other might have greater social importance. References to ethics and social responsibility on the part of firms in recent years may well indicate that society is showing preferences as to how the *performance* of firms should be evaluated, above and beyond their results.

[6] With the development of information and communications technologies (TIC), new models of firms and firm business emerge—Microsoft, Intel, Cisco—as the empirical referents, replacing Ford and GM (referents until the 1980s) and Toyota (in the late twentieth century). The new firm model refers to the virtuality and network structure adopted by firm organization and breaks with the tradition of lasting relations (life employees, long-term contracts with suppliers, client fidelity) that were predominant in earlier firms. If the view of a firm as a human community has any reason to exist in this new setting, it must be reinvented (Castells 1996).

This text is largely based on the author's book, *El Siglo de la Empresa*, published by the BBVA Foundation.

Bibliography

Aighon, P. and J. TIROLE. “Formal and real authority in organizations.” *Journal of Political Economy*, 105, 1997, 1–29.

Alchian, A. and H. Demsetz. “Production, information and economic organization.” *American Economic Review*, 62, 1972, 777–795.

Alonso, R., W. Dessein, and N. Matouschek. “When does coordination require centralization?” *American Economic Review*, 98, 2008, 145–179.

Arrow, K. “The organization of economic activity. Issues pertinent to the choice of market versus non market resource allocation.” In US Joint Economic Committee, *The analysis and evaluation of public expenditures. The PBS system*, 1969.

Baker, G., R. Gibbons, and K. Murphy. “Relational contracts and the theory of the firm.” *Quarterly Journal of Economics,* 117, 2001, 39–83.

Calvo, G. and S. Wellisz. “Supervision, loss of control and the optimal size of the firm.” *Journal of Political Economy*, 87, 1978, 943–952.

Camerer, C., G. Loewenstein, and M. Rabin. *Advances in Behavioral Economics.* Princeton: Princeton University Press, 2004.

Castells, M. *The Rise of the Network Society. Vol. 1: The Information Age: Economy, Society and Culture*. Berkeley: University of California Press, 1996.

Chandler, A. *Strategy and Structure*. Cambridge: MIT Press, 1962.

—. *The Visible Hand*. Cambridge: Belknap Press, 1977.

Coase, R. “The nature of the firm.” *Economica*, 4, 1937, 386–405.

Dimaggio, P., *The Twenty-First-Century Firm*. Princeton: Princeton University Press, 2001.

Frey, B. *Happiness: A revolution in economics*. Cambridge: MIT Press, 2008.

Gintis, H., S. Bowles, R. Boyd and E. Fehr. *Moral Sentiments and Material Interests: The Foundations of Cooperation inEconomic Life*. Cambridge and London: MIT Press, 2005.

Grossman, S., O. Hart. “The costs and benefits of ownership: A theory of lateral and vertical integration.” *Journal of Political Economy*, 94, 1986, 691–719.

Guth, W., R. Schmittberger, and B. Schwarze. “An experimental analysis of ultimatum bargaining.” *Journal of Economic Behavior and Organization,* 3, 1982, 367–388.

Hansmann, H. *The Ownership of Enterprise*. Cambridge: Belknap Press, 1996.

Hart, O., J. Moore. “Property rights and the theory of the firm.” *Journal of Political Economy*, 98, 1990, 1.119–1.158.

Hayek, F. “The use of knowledge in society.” *American Economic Review*, 35, 1945, 519–530.

Helpman, E. “Trade, FDI and the organization of firms”, *Journal of Economic Literature,* 3, 2006, 589–630.

Holmstrom, B. “Moral hazard and observability.” *Bell Journal of Economics*, 10, 1979, 74–91.

—. “Moral hazard in teams.” *Bell Journal of Economics*, 13, 1982, 324–340.

—. "The firm as a subeconomy.*" Journal of Law Economics and Organization*, 1999, 74–102.

Holmstrom, B. and Milgrom. “Multitask principal-agent analysis: Incentive contracts, asset ownership and job design.” *Journal of Law, Economics and Organization*, 7, 1991, 24–52.

—. “Aggregation and linearity in the provision of intertemporal incentives.” *Econometrica,* 55, 1987, 308–328.

Jensen, M. and W. Meckling. “Theory of the firm: managerial behaviour, agency costs and ownership structure.” *Journal of Financial Economics* 3, 1976, 305–360.

Kahneman, D., A. Tversky. “Prospect theory: An analysis of decision making under risk.” *Econometrica*, 47, 1979, 263–291.

Kandel, E. and E. Lazear. “Peer pressure and partnerships.” *Journal of Political Economy*, 100, 1992, 801–817.

Klein, B., R. Crawford, and A. Alchian. “Vertical integration, appropriable rents and the competitive contracting process.” *Journal of Law and Economics,* 21, 1978, 297–326.

Kochan, T. and R. Schmalensee. *Management: Inventing and delivering its future*. Cambridge: MIT Press, 2003.

Kogut, B. and U. Zander. “Knowledge of the firm, combinative capabilities and the replication of technology.” *Organization Science,* 3, 1992, 383–379.

Kreps, D. “Corporate Culture”, in J. Alt, K. Shepsle edrs. *Perspectives on Positive Political Economy.* Cambridge: Cambridge Uni. Press, 1990.

Malone, T. H., R. Laubacher and M. Scott Morton (eds.). *Inventing the Organizations of the 21st Century.* Cambridge: MIT Press, 2003.

Marschak, J. and R. Radner. *Economic Theory of Teams*. New Haven: Yale University Press, 1972.

Milgrom, P. and J. Roberts. *Economics, Organization and Management*, New Jersey: Prentice Hall, 1992.

—. “Complementarities and fit: Strategy, structure and organizational change in manufacturing.” *Journal of Accounting and Economics*, 19, 1995, 179–208.

Nelson, R. and S. Winter. *An Evolutionary Theory of Economic Change.* Cambridge: Belkman Press, 1982.

Roberts, J. *The Modern Firm*. Oxford: Oxford University Press, 2004.

Rosen, S. “Authority, control and the distribution of earnings.” *Bell Journal of Economics*, 13, 1982, 311–323.

Simon, H. “A formal theory of employment relationship.” *Econometrica*, 19, 1951, 293–305.

—. “Organizations and markets.” *Journal of Economic Perspectives*, 5, 1991, 25–44.

Teece, D. “Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy.” *Research Policy*, 15, 1986, 285–305.

Tirole, J. “Corporate governance.” *Econometrica*, 69, 2001, 1–35.

Williamson, O. *Markets and Hierarchies*. Nueva York: Free Press, 1975.

—. *The Economic Institutions of Capitalism.* New York: Free Press, 1985.

frontier research in economics

JOAN ESTEBAN

Introduction

This essay deals with trends in economic theory over the past few decades. It is unabashedly subjective and partial. It does not attempt to provide an exhaustive panoramic view of current research in economics. Rather, I have chosen to focus on some of the recent developments that have tried to relax the highly restrictive assumptions under which General Equilibrium Theory (GE) had been built over the second half of the twentieth century. This enrichment in the description of the working of individuals, firms, government, and society at large, has also had the side effect of significantly increasing the inter-disciplinary nature of current research in economics. We are witnessing a remarkable overlap with political science and sociology, of course, but also with psychology, biology, and neuroscience.

Even with such a severe restriction in scope, I shall have to be more superficial than I would like. Also, my choice has the drawback of leaving completely uncovered important and dynamic areas of economics such as macroeconomics, finance, trade, and development, to mention a few.

The essay proceeds as follows. In the next section I start by giving a summary view of the GE model, undoubtedly the core paradigm in economics. Section 3 describes the major departures from the standard GE model. Then I move into a more in depth analysis of the recent contribution of the *behavioral* approach to individual decision making. Inspired by research in psychology and largely based on controlled experiments, behavioral research tries to carefully document patterns of individual behavior that deviate from the choices predicted by the classical rational behavior model. Section 4 provides a description of all the ingredients involved in a decision so that we can give a more structured account of the different results and what are the precise ingredients that are questioned. Section 5 gives a synthetic account of the main contributions in *behavioral economics*. Finally, Section 6 takes stock of the research reported, makes an evaluation of the net contribution and derives implications for future research.

General Equilibrium and Welfare Economics

Modern GE theory started in the 1950s. The 1954 paper by the Nobel laureates K. Arrow and G. Debreu on the existence of a competitive equilibrium and the 1959 book by Debreu *Theory of Value* can be taken as the beginning of four decades of an extremely fruitful effort to understand the working of competitive markets. The books by Arrow and Hahn (1971) *General Competitive Analysis*, W. Hildenbrand (1974) *Core and Equilibria of a Large Economy*, and A. Mas-Colell (1990) *The Theory of General Economic Equilibrium: A Differentiable Approach* can be considered the most significant landmarks of this endeavor.

GE consists of a very complex but schematic model that tries to capture the coordinating role played by markets in an otherwise atomistic and individualized society. It provides a rigorous proof of the Smithian claim that the invisible hand is sufficient to make mutually compatible the decisions taken by completely uncoordinated individuals and firms.[1] In this sense, the elimination of any reason for explicit or tacit coordination as well as of individual concern for others was crucial to the model. With all participants furthering their own interest (narrowly understood) and limiting their interaction just to supply and demand through the markets, the outcome would not turn out to be chaotic but orderly and efficient—in a sense to be made precise below.

Let us recall some of the assumptions needed to derive this mathematical result. The participating agents are an arbitrary number of individuals and firms. All participants can buy and sell commodities (labor, for instance) through the markets at the ruling prices. Most important, all participants are assumed to behave competitively, that is, to take prices as given.[2] In addition, individuals own all the shares over existing firms. The distribution of these shares across the population is arbitrary.

Initially, each individual owns a collection of commodities (which may well be only labor) and shares. By selling and buying they can obtain a new set of commodities (for instance, eight hours labor sold and bread and butter purchased). The amounts traded have to be within each individual's budget, where the monetary value of the budget is determined by the ruling prices. All individuals have preferences over the traded commodities, that is, they can rank any pair of bundles of commodities in terms of their desirability. Besides being complete, this ordering is assumed to be transitive, reflexive and satisfy a weak convexity condition.[3] Most important, these preferences depend on personal consumption only, hence eliminating any form of altruism. Then, individuals are assumed to act rationally. That is, among the trades affordable to them (the value of purchases cannot exceed the value of sales) they choose the one they rank highest in terms of desirability. Thus, for any given vector of market prices each individual consumer thus has a well-defined vector of demands and supplies of the traded commodities.

Firms purchase through the market commodities and labor supplied by other firms and by individual consumers (inputs) to turn them into a set of produced outputs to be sold in the market. How much output firms can produce from a given vector of inputs is conditioned by the production technology available to them. The set of feasible combinations of inputs and outputs is assumed to be convex.[4] For every vector of prices, each firm chooses the vector of inputs (purchases) and the vector of feasible outputs (sales) with a view to maximizing their own profits.

An equilibrium is a vector of prices such that when all purchases and sales are aggregated for the entire economy, supply is equal to demand in each of the markets. A good part of the research in GE until the mid-nineties was devoted to demonstrating the existence of such an equilibrium vector of prices under the weakest assumptions possible on individual preferences and production technology. Indeed the collection of individual decisions taken by egoistic individuals and firms without any coordination can turn out to be feasible rather than generate disorder. Besides proving that the concept of equilibrium is not vacuous—there always exist such equilibrium situations—GE theorists also obtained conditions under which this concept was not too lax: equilibria are determinate, thus excluding continua of equilibria.

The most remarkable results of GE theory—the "Two Fundamental Theorems of Welfare Economics"—prove that these market equilibria have interesting efficiency properties. W. Pareto defined a basic efficiency requirement that has become fundamental in economics: a situation is (Pareto) efficient if by reallocating commodities in the economy it is not possible to improve the well-being of someone without harming somebody else. Notice that distributional justice is completely absent from this notion. An allocation in which one person owns everything while the rest are starving to death is efficient as long as individual preferences never reach satiation.

The First Fundamental Theorem establishes that all competitive equilibria are Pareto efficient. Therefore, market exchange among self-regarding participants also leads to an efficient use of the existing resources. The Second Fundamental Theorem says that every efficient allocation of commodities can be implemented as a competitive equilibrium, with an adequate redistribution of the initial resources. It follows that a socialist planned economy cannot do better than competitive markets—with an appropriate one-time redistribution of resources.

[1] This issue—markets versus socialist planning—became salient in the Cold War political debate.

[2] This assumption is more plausible when all participants are so small that they cannot influence prices by their action. It is obvious that this is not the case. Governments have had to set up competition agencies more or less effectively trying to guarantee that firms will not collude to manipulate prices. Workers too are to a large extent unionized in order to keep wages (and working conditions) up.

[3] If bundle A is strictly preferred to B, then any convex linear combination $\lambda B + (1 - \lambda) A$, ($\lambda > 0$) is strictly preferred to B.

[4] If the combinations A and B are feasible, so is $\lambda B + (1 - \lambda) A$, $0 \leq \lambda \leq 1$.

How much to redistribute and how to do it without distorting the working of the markets clearly is a question complementary to GE theory. These kind of questions pertain to *Welfare Economics*. If the government has to choose it has to be that there are some sort of "social preferences" ranking alternative policies by the social desirability of their outcomes. As early as 1951, K. Arrow (Nobel laureate, 1972) demonstrated that it was not possible to aggregate individual preferences into a social preference ranking, if this had to satisfy a set of reasonable properties. Welfare Economics—as well as Public Economics in general—ended up by assuming that somehow social priorities could be encapsulated into a *social welfare function*. The role of the government was then modeled in the same spirit as individual choice: to maximize social welfare under feasibility constraints. The contributions of P. Diamond and J. Mirrlees (Nobel laureate, 1996) in the mid-seventies set the basis of modern public economics by rigorously rooting the theory of government intervention on the foundations of GE theory.

This summary, of course, only records the most essential results of GE theory and the associated welfare economics.[5] At the January 1994 meeting of the *Econometric Society* one of the most distinguished contributors to the development of GE theory, Andreu Mas-Colell, gave an invited lecture on "The Future of General Equilibrium."[6] His presentation transpires the perception that GE theory had already reached its peak and that the attention of young researchers had already turned towards other issues that had been left aside by GE theory. We are going to review some of these new lines of research. But, before moving to the marrow of my essay it is imperative to stress one fundamental contribution of GE theory: mathematical rigor. This precisely is Mas-Colell's (1999) last line: "I would hope that the role of theorem proving is kept alive, perhaps not as strong as before, we may have overdone it, but with substantial presence" (p 214).

Major recent departures from the standard model

The extremely stringent assumptions of the GE model were obvious to all theorists, but were considered the price to pay to have a clean model of the working of the markets.

One obvious reservation is that in many markets there aren't sufficient enough firms so as to justify the assumption of competitive behavior. There are situations in which there exists a monoplist, and there are even situations in which a monopolist is considered to be "natural," as in the case of the supply of electricity, cable TV, and so forth. A monopolist is not a "price taker," and it may take into account that the quantity it decides to produce will affect the price at which it is sold. In this case, the market equilibrium will typically not be Pareto efficient. The same result applies if there are several firms in the market, but their number is not large enough for each of them to act as if it had no effect on prices. This has given rise to the field of *industrial organization* mostly developed in the late eighties and nineties.[7]

A second major departure from the classical GE model has been the study of the role of information in the eighties and nineties. In the standard model, all participants are assumed to have the same information—which might mean the relevant probabilities in case of uncertainty. However, it is plain that this is not always the case. The classic example (Akerlof, Nobel laureate 2001) is the market for second hand cars: the seller, who has owned the car, has more information on its quality than does the buyer. In such a situation, one can show that equilibria are typically not Pareto efficient. Akerlof's used car example ("the market for lemons") is a parable that applies to a multitude of situations in which information about trade is not symmetric. Other examples include the insurance market (where the insured may know more about the risk than the insurer), employment contracts (where the employed "agent" may know more than the employing "principal"), and so on (Akerlof, Mirrlees, Stiglitz, Vickrey, all Nobel laureates).[8]

Yet another major deviation from the classical model has to do with externalities, namely, situations in which the consumption of one agent might directly affect the wellbeing of another. In particular, cases in which there is a public good—a good that can be used by many individuals simultaneously, such as hospitals, transportation, education, defense—fall in this category. Again, it was shown that competitive markets cannot be relied upon to result in a Pareto efficient allocation in these situations.

It so happened that all these deviations from the classical GE model were analyzed using game theory. Game theory started out as the analysis of parlor games at the beginning of the twentieth century. In the 1940, O. Morgenstern and J. von Neumann wrote the first book on the topic, which also suggested that the theory is the correct way to analyze all social and economic situations. The approach was soon refined by J. Nash, who held that all such situations should be analyzed from the level of the individual decision maker up. Nash suggested the notion of "equilibrium", currently named after him (Nash Equilibrium), which requires that each decision maker chooses his or her best course of action in accordance with what the others are doing.

[5] See Mas-Colell et al. (1995) for a state-of-the-art presentation of the General Equilibrium model and microeconomic theory.

[6] See Mas-Colell (1999).

[7] The books by Tirole (1988) and Vives (2001) give a comprehensive overview of the topic.

[8] See Stiglitz (2002) panoramic presentation in his Nobel lecture.

Game theory was perfectly suited to dealing with non-competitive markets, but only after the contributions of J. Harsanyi did it become apparent that situations of asymmetric information were also amenable to game theoretic analysis. This also made game theory the natural method to analyze problems of externalities and public goods. Hence game theory became the standard tool of analysis in microeconomic theory. Since the mid-seventies, economic theory has become dominated by game theory. Over recent decades, game theory has also proven a fundamental tool for macroeconomics and even political science. It seemed that any problem in the social sciences can be thought of as an application of game theory.[9]

Current research, however, is grappling with several fundamental problems, which suggest either that not all major problems can be relegated to game theoretic analysis, or that such analysis might not be complete. One can classify the different recent departures from the paradigm in three categories: (i) how individuals and firms decide; (ii) how governments decide; and (iii) how agents interact.

Even in such narrow an area of economics there are too many developments to permit a coherent and comprehensive presentation. I shall focus on the frontier research in individual rational choice only.[10] However, before moving on, I shall give a sketchy picture of the main lines of progress on group decision and on the departure from the competitive assumption.

It is plain that government economic policies are not decided on the basis of maximizing a social welfare ordering. Rather, in democracies political parties propose policies and citizens vote over the proposed manifestos.[11] This is the orderly way that modern democracies are designed to resolve the opposing interests that characterize all societies. Therefore, if we want to understand the policies actually enacted we have to explain how political parties choose their political platforms and how they behave once in office.[12] This approach is what has become to be known as *positive political economy.*[13] Of course, the literature has also explored the known fact that large firms and special interest groups are also effective in influencing the government in its decisions by lobbying through various channels.[14]

The political system is itself endogenous and responds to the nature of different, often opposing interests within the society. Acemoglu and Robinson (2006) have studied the role of social and economic change in forcing the political system to change. Specifically, they argue that democracies have evolved as a commitment device: tension over the distribution of wealth and the threat of a revolution have historically forced monarchs to share their wealth. However, a promise to do so may not be credible. Democracy, according to this view, involving giving voting rights to larger fragments of society, allowed commitments to be credible and thus averted conflicts. But can the social contract always be modified in response to social changes? Clearly not. In the second half of the twentieth century there have been over 16 million civil deaths in civil wars. This exceeds by a factor of five the number of battle deaths in the same period. The threat of civil wars currently is so endemic that the World Bank considers political instability the most serious impediment for the effectiveness of foreign aid in promoting growth in developing countries. In economics, Hirshleifer (1991), Grossman (1991), and Skaperdas (1992) have pioneered the study of conflict, but we are still far from a satisfactory understanding of the causes of civil conflict.[15] How economic agents interact is simply not modeled by GE theory. How the products produced by firms reach consumers or other firms is left out of the model.[16] This lack of interest in the actual operation of economic transactions is partly due to the exclusive focus on competitive markets in equilibrium. Indeed, if markets are competitive and in equilibrium there is no point in looking for a better deal. All the requests get (magically?) satisfied and there is no chance of finding a seller offering a lower price. If, however, there are different prices for the same commodity or if certain requests (such as an individual's request to sell his or her labor) might not get satisfied, it certainly does matter how the transactions actually operate.

Most of the research on the actual working of economic interactions has been investigated by *development economics*. The most superficial observation of underdeveloped countries makes it plain that their economies cannot be conceived as a set of competitive markets in equilibrium. We owe to this branch of the literature most of the insights on the role of specific institutions, networks, and social rules in channeling economic interactions.[17] At a more formal level, the existence of network connections and their impact on economic interactions has caught the attention of economic theorists. The essence of the model bears similarities with the local interaction models in physics with a crucial twist: the nodes are optimizing individuals or firms that can create or sever ties.[18] This approach has given new insights in different areas of economics such as education[19] and the labor market.[20]

[9] Other major contributors to game theory include L. Shapley, J. Aumann, and R. Selten (the latter two are also Nobel Laureates, alongside J. Nash and J. Harsanyi).

[10] Note that I am skipping the new views on the behavior of firms beyond profit maximization. As an excuse, let me cite Mas-Colell (1999) who says "I am not sure that (...) we will end up seeing the theory of the firm at the heart of economics."

[11] The first classical model of a democracy is due to Downs (1957).

[12] Oemer (2001) is a rigorous and comprehensive book on political competition.

[13] See Persson and Tabellini (2000) for a very influential book in this area.

[14] See Grossman and Helpman (2001).

[15] See Esteban and Ray (1999) for a general model of conflict. Fearon (1996), Powell (1999) and Ray (2008b) have developed arguments to explain why society may fail in agreeing on a new social contract that Pareto dominates the costly outcome of civil conflict.

[16] As a consequence, the GE model is unable to analyze how the economy can reach the precise vector of prices that will clear all the markets.

[17] Ray (1998) is the basic advanced textbook in development economics. See also, Ray (2008a) for an overview of the recent developments in this area.

[18] See Jackson (2008b) for an survey of recent contributions and the books by Goyal (2007) and Jackson (2008a) for an extensive presentation.

[19] See Benabou (1993; 1996).

[20] See the survey by Calvo-Armengol and Yoannides (2008).

"Classical" rational choice

Many of the new directions of research in economics are driven by the need to relate theory to facts more strongly. This is especially true of modern "behavioral economics," as some colleagues term it. Since observed individual behavior is often at odds with the decisions that derive from the standard rational choice assumptions, economists have turned their attention towards the patterns of behavior that psychologists have been identifying by means of controlled laboratory experiments. The pioneering work of psychologists Kahneman (Nobel laureate in Economics, 2002) and Tversky[21] has recently had a profound influence in economics. In addition to opening the minds of economists to the findings in psychology, it has also triggered a remarkable boom in experiments on individual and group behavior. C. Camerer, E. Fehr, D. Laibson, and M. Rabin, to mention a few, are among the economists that have worked more intensively in this field.[22]

To proceed in an orderly fashion, I find it useful to separate the essential ingredients of the individual decision problem.

The first ingredient is an informational input. Individuals observe some information about the state of world. In standard consumer theory this information refers to prices and incomes and, in an uncertainty environment, to the probability of each of the possible realizations of the states of the world. This information typically constrains the set of actions available to each individual. The second ingredient is the computation of the consequences (deterministic or probabilistic) that derive from each possible action. For instance, we take the action of working for eight hours and purchase meat and fish. Or, we can give up on consumption by ten euros and spend them in buying lottery tickets with given probabilities on a set of prizes.

The third ingredient is individual preferences. These preferences are assumed to generate a ranking over all possible consequences (independently of the actions taken as such) according to their desirability, as explained before. When the consequences are probabilistic, individuals rank actions by their expected utility, that is, the weighted average of the utility of the various realizations, using probabilities as weights. The standard model assumes that these preferences are egotistic, independent of what others obtain.

The last ingredient is rational choice. By this we mean that each individual is able to solve the constrained maximization problem consisting of identifying the action in the feasible set that has the most desirable consequence.

Psychology and individual decisions

Experimental work on individual behavior consists of confronting a set of individuals with a situation (as controlled as possible) in which classical decision theory has an unequivocal prediction so that the researcher can contrast actual with predicted individual choices. There is a rich variety of such experiments exploring different types of violations of the predictions of the standard model.[23] For instance, a repeatedly studied experiment consists of subjecting a sample of individuals to the *ultimatum game*. Individuals are randomly matched in pairs and one is assigned the role of proposer and the other that of the receiver. The proposer announces a division of a given amount of money among the two players. Then the receiver either accepts—and the money is divided according to the proposed allocation—or refuses—and both players receive zero. If players care about their own material interest only, the second player should accept any strictly positive amount of money. Knowing this, a selfish proposer should give an arbitrarily small amount to the receiver and cash the rest. As it turns out, in all specifications of this experiment there is a substantial proportion of proposers that propose divisions that are quite close to the egalitarian one and of receivers that are ready to give up even a non-negligible positive prize in order to "punish" unfair proposals.

This experiment is but one example of the plethora of patterns of choice that are currently being tested by behavioral economists. Some of these experiments challenge certain specific ingredient of our previous description of the decision process. But some are less targeted and try to identify violations of the predictions of standard rational choice theory.

Let us go through some relevant complexities that standard rational choice theory dismisses by assumption at each of the ingredients of a decision. Some have been studied by behavioral economics, but many are still to be carefully explored. As it will become clear, my position is somewhat ambivalent. On the one hand, I think that economics has to enrich its basic model of rational choice. But, on the other hand, I am quite skeptical—if not critical—with many of the claims of behavioral economics. In this respect I feel more in line with the critical positions of Gul and Pesendorfer (2008) and Rubinstein (2006).

The first ingredient of choice is the processing of information. There are many channels through which the acquisition of information may affect decisions. In the first place, individuals categorize information. This has been an object of study by social psychologists for the past five decades, but only recently have economists started paying attention to

[21] Tversky died in 1996. See the essay of Laibson and Zeckhauser (1998) on Tversky's contributions.

[22] Camerer and Rabin invited lectures at the Ninth World Congress of the Econometric Society provide an overall view of the potentials of marrying psychology and economics [see also the discussion by Ariel Rubinstein]. They are all published in Blundell et al (2006).

[23] See Della Vigna (2008) for a comprehensive survey of the different types of experiments.

it. It is immediately clear that such a process can bias our decisions. Fryer and Jackson (2008) study how efficient processing of information leads to a coarser categorization of the types of experiences that are less frequently observed, lumping them together. As a result, decision makers make less accurate predictions when confronted with such objects and this can result in discrimination. Secondly, individuals have to have an idea of which information is relevant to the decision at hand and which not. In order words, they need to entertain a "model" linking the possible actions to their consequences, as we shall soon discuss. However, there is psychological evidence that individuals tend to censor evidence that refutes their view of the world.

Moreover, Benabou and Tirole (2006) argue that this censoring mechanism may serve a function by supporting a belief that the degree of social mobility is higher than it actually is, and thereby inducing individuals to make a higher effort than rational choice would warrant. Such unrealistic beliefs are especially necessary in countries with a limited social net, such as the US. Finally, the rational choice assumption that individuals will use information to perform Bayesian updatings of the relevant probabilities may be unwarranted. As argued by Gilboa et al. (2007) there are instances in which individuals cannot have prior beliefs that are represented by probabilities to start with because rational, justified beliefs fail to pin down a numerical probability.

The second ingredient consists of mapping actions onto consequences, either deterministically or probabilistically. This step presumes that in view of the evidence individuals can identify a model that fits the data and that this model is unique. It is plain that this is generally not the case. More formally, Aragones et al. (2005) show that given a knowledge base, finding a small set of variables that obtain a certain value of R2 is *computationally hard*, in the sense that this term is used in computer science. Because of this fact, rhetorical statements contributing no new evidence can induce a change of behavior, as they may make one aware of certain relationships among known variables, which are obvious post hoc, but have not been conceived of before. Multiple theories and as many decisions are compatible with a given stock of evidence. Picketty (1995) developed a model in which the belief in a particular theory induces decisions that generate evidence confirming this theory. Different individuals can sustain different theories and parents have an incentive to transmit their own theory to their children. Finally, another way of making decisions in the absence of a determinate interpretation of evidence is via social imitation. Banerjee (1992) developed a model of herding in which individuals make inferences from the observation of the behavior of others regarding the information they might have had and this leads them to act in a similar fashion.[24] An alternative line is the effect of social identity in behavior. Akerlof and Kranton (2000) were the first to call the attention of economists to the role of individual identification with categories or prototypes. I think that this is an important line of research, unfortunately still largely unexplored.[25]

The third ingredient is individual preferences. This possibly is the front where classical economic theory is more restrictive and unrealistic. In the first place, it assumes that preferences are defined over own consumption only. This excludes altruistic motivations of which we have ample evidence. There is a vast literature on altruism, especially reciprocity based. Rabin (1993), Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Falk and Fischbacher (2006) have focused on the interaction between pairs of players where each player attempts to infer the motive of its partner and then modifies its social preferences accordingly, giving greater weight to partners who are believed to be benevolent and less weight to those who are selfish or malevolent. Additionally, individuals may value actions per se, independently of the valuation of their consequences. This refers to the self (self-esteem/pride) and to others (ethical judgments). See Lindbeck et al. (2006) for an economic analysis in which parents seek to instill work norms in their children which are sustained by guilt and Tabellini (2007) for the adoption and transmission of values of generalized morality.[26] Finally, individuals seem to face changes in their preferences both through time[27] and through restrictions in their choice set.[28]

The fourth ingredient is decision making proper. That is the process of combining all the information available and turning it into the choice of a particular action. The assumption of rationality means that individuals are supposed to choose the action they value highest within the actions available to them. Therefore, one can say that individuals do not behave rationally only when the analyst can offer an alternative action different from the chosen one and that the individual accepts as preferable.

The joint consideration of the available information, the link between actions and consequences and the valuation of these consequences involves considerable reasoning and the reasoning capacity depends on education, training, and past experience. It follows that we can conclude that individuals behave non-rationally only if they insist on their choices after being shown by the analyst that better

[24] See the recent overview on herding by Rook (2006).

[25] Esteban and Ray (1994) launched the idea that social conflict is driven by the sense of identity with one's own group, combined with the complementary sense of alienation with respect to the others. However, they just axiomatize an index of polarization rather than go into developing a model of identity formation.

[26] For the particular case of norms pertaining to work, seminal contributions by Moffit (1983) and Besley and Coate (1992) consider the case in which there is stigma associated with living on welfare. Lindbeck et al. (1999) have extended this analysis to include voting over welfare benefits. Cervellati et al. (2008) let moral judgments on effort determine self-esteem and esteem for the others.

[27] The psychological cost of a given time postponement of a prize is higher if this happens immediately than if it does later in the future. Laibson (1997) has termed these time preferences as displaying "hyperbolic discounting."

[28] Here is a prototypical example: I definitely prefer not to smoke. I also would like to have lunch with a friend who happens to be a smoker. I may commit not to smoke by not having lunch with my friend. I know that if I have lunch with her I may give in to the temptation of smoking. This idea has been thoroughly explored by Gul and Pesendorfer (2001).

choices existed for them. We shall come back to this notion of rationality based on Gilboa and Schmeidler (2001).[29] In the case of certainty, rational choice amounts to solving a maximization problem under feasibility constraints. However, already in the 1950s Simon (Nobel laureate, 1978) argued that individual rationality was "bounded" in the sense that human computing capacity is limited. But this line of enquiry has not had much following. Notice that "bounded rationality" does not imply non-rationality in the sense above. The reasoning required for a decision in the case of uncertainty is much more complex as it calls for the additional consideration of the probability of the occurrence of every consequence possible. Classical rational decision theory assumes that individuals value each action by the weighted sum of the valuation of the consequences, using the relevant probabilities as weights.[30] Experimental evidence seems to confirm regular violations of some of the axioms—the so-called Allais and Ellsberg paradoxes. In order to reconcile theory and behavior, Kahneman and Tversky (1979) proposed *prospect theory* for modeling decision under risk. Based on behavior—and not on axioms—they claim that probabilities do not enter linearly in the valuation of an action but through a weighing function that exaggerates low probabilities and moderates large probabilities. Also the valuation of each consequence is measured as a deviation from a reference outcome.[31] Notice that this approach continues to assume that there exist well-defined probabilities for each of the possible consequences of an action. Yet this is rarely the case. Gilboa and Schmeidler (2001) postulate that decisions are based on the results observed in previous cases that are considered "similar" to the current problem. From a set of axioms they derive that the value of an action is the sum of the utility levels that resulted from using this action in past cases, each weighted by their similarity to the current problem.

As far as rational choice is concerned, we can conclude that, while there is little controversy on rational decision making in certainty, the case of uncertainty is still unsettled. Summing up, we have seen that actual behavior appears to display deviations in each of the ingredients of a decision problem. There certainly is room for enriching our modeling of how information is acquired and processed, of how individuals link consequences to actions, or even what are the different dimensions that agents value (in addition to the material tradable commodities). However, none of these changes appears to have much to do with the basic notion of making of an optimal decision under certain constraints. In the next section we shall discuss what can we learn from the findings of behavioral economics and the extent to which they challenge the assumption of rational individual decision making.

Behavioral economics: taking stock

Where does the exploration of the links between psychology and economics lead us? The vast and solid empirical evidence that there is a number of psychological factors that matter when individuals make decisions has been seen by some behavioral economists as a challenge to the core paradigm of economics on rational choice.[32]

I shall argue that it is unclear how experimental evidence can be extrapolated outside the laboratory—what can we learn from it?—and that the enrichment of the behavioral description of decision makers is likely to have more influence on models of applied economics than on the paradigm and core assumption that individuals act rationally (and essentially) out of self-interest.

What can we learn from experimental evidence?

In the first place, we obtain a controlled confirmation that individuals behave differently from what prescribes classical economic theory. Indeed, individuals care for things other than their own material consumption, such as the actions they may take per se, moral judgments of the self and of others, etc. Decision makers also do not perfectly process information and often violate some axioms of rational behavior. However, it is not always obvious how such evidence should be interpreted and, even if it were unequivocal, whether the violations found in experiments should form part of the standard modeling of individual behavior.

In a sense, the experiment by Kahneman and Tversky (1984) showing that framing does have an effect on individual behavior makes one skeptical about what we can learn from experiments. First, there is a suspicion that the observed behavior has been induced by the particular way the choice problem had been presented to the participants.[33] Should we then conclude that people generally violate the most basic assumptions of the theory, or should our conclusion be that *sometimes*, given certain very clever formulations, people may act in highly irrational ways?

Implicit in the work of experimental behavioral economists is the belief that there is a natural pattern of behavior that was not properly captured by classical decision theory and that can be identified by means of critical experiments. Camerer and Loewenstein (2003) tell us that "behavioral

[29] See a more detailed analysis in Gilboa et al (2008).

[30] Properly speaking the assumption is that individual choices respect the axioms proposed by von Neumann and Morgenstern (1944), which imply that the valuation of an action is its expected utility.

[31] The case of Ellsberg paradox is examined in C. R. Fox and A. Tversky (1995).

[32] Rabin and Thaler (2001) refer to classical expected utility theory as a "dead parrot," from the comedy sketch by Monty Python.

[33] This and other arguments on the difficulty of extracting conclusions from behavioral experiments are carefully examined in Levitt and List (2007).

economics increases the explanatory power of economics by providing it with more realistic psychological foundations" and that "the conviction that increasing the realism of the psychological underpinnings of economic analysis will improve economics on its own terms." Also, Rabin (1998) asserts that "because psychology systematically explores human judgment, behavior, and wellbeing it can teach us important facts about how humans differ from the way traditionally described by economics." Therefore, the purpose is to capture the *true* nature of individual decision making from factual observations in experiments or by other means.[34]

Can we capture this "nature" of decision-making by experiments? Further, does this "nature" exist in a meaningful sense?

Leaving apart the reservations on the effective ability to control experiments, it still remains unclear what is the exact "nature" we are measuring. In order to illustrate my point let me take the most popular experiment that we have described before: the ultimatum game. The costly refusal of "unfair" proposals is interpreted as showing that individuals also care about things other than their personal monetary payoff.[35] However, it can also be that this costly rejection of an unfair proposal is an emotional reaction that momentarily obscures what reason would have dictated. The extent to which reason overrides emotions varies across individuals—possibly depending on education and training—and, in any case, only the dictates of reason should be taken to conform to rational behavior. If we are interested in the choices that a specific group of individuals will make in a given circumstance, it might be critical to know whether they will react bluntly or whether they will make cold calculations.[36] However, it seems natural that a general theory of individual behavior should abstract from the fact that we may momentarily deviate from rationality.

This raises a fundamental question to which we shall return: whether rationality is something positive or normative. Should society train citizens to be rational?[37] In fact, we do through the compulsory educational system...

Even for a given degree of sophistication in reasoning specific experiences or training can have a profound effect on behavior. In trying to empirically identify the notions of equity actually used by individuals, the work of Amiel and Cowell (1992) is very pertinent to substantiating the point I am making. Students were shown a series of two lists of (ten) incomes and asked to rank them in terms of their relative inequality. The purpose was to test which of the different criteria used in economics could find wide acceptance. Among these criteria they tested the *principle of progressive transfers* popularized by Atkinson (1970). This principle says that if we transfer one euro from any single person to someone poorer the resulting distribution is less unequal. The result of relevance here is that this principle found wide support among economics students—who were directly or indirectly familiar with the concept—and quite modest among the other students. Indeed, we can more easily interpret information when we have been told how to organize it.

The previous question of whether there is a "nature" of decision making that can be captured by experiments was somewhat rhetoric. The point is that with the present state of knowledge it is not possible for the experimentalist to conduct critical experiments. As discussed in detail by Levitt and List (2007), even the most carefully designed experiment cannot guarantee that all other influences have been effectively controlled by the analyst. Therefore, while experiments have an extremely useful role in highlighting deviations from prescribed behavior, they cannot in general unequivocally identify the causes of such deviant behavior. I find it very important that these experiments be continued, but I am persuaded that this will be a long term project that will require time, effort, and patience.

Behavioralism and rational choice

Behavioralism will have more influence in models of applied economics than in redefining the core paradigm of individual rational choice. Let me present two arguments in support of my point.

My *first argument* is that there are still too many aspects of behavior of which we have but a very imperfect understanding. We see that individuals may be motivated by altruistic feelings, for instance. However, we are still not able to understand the causes of the variation of these feelings across the population. Some researchers have seen altruism as driven by the search of the benefits of reciprocity. But even reciprocity can be in material benefits or it can be a reciprocity of attitudes. Other researchers see altruism as deriving from moral convictions. We also observe that the degree of altruism depends on the proportion of the group that behaves altruistically. We have only conjectures as to how all these aspects interact. So far we don't know whether moral values, response to observed behavior by others, tendency to reciprocate, and the like are exogenous parameters or at least partly result from the variables we are trying to analyze. It is obvious that without exactly knowing (or

[34] I shall not discuss the current attempts at exploring the link between decision making and brain activity. See Gul and Pesendorfer (2008) for a critical view.

[35] This interpretation is reinforced by the interesting result that when the proposer is replaced by a machine that it is known to select proposals randomly, then the second player accepts unfair proposals much more easily.

[36] Even if rejection truly were the result of moral disappointment, the experimenter should test whether leaving the decision of rejection for the next day would alter the results.

[37] "Human beings, Romans argued, consist of two elements: an intelligent, rational spirit, and a physical body. [...] In fully rational people—such as elite Romans, of course—the rational spirit controlled the physical body. But in lesser human beings—barbarians—body ruled mind. [...] Where Romans would calculate probabilities, formulate sensible plans and stick to them through thick and thin, hapless barbarians were always being blown all over the place by chance events." P. Heather, *The Fall of the Roman Empire*. Oxford University Press, 2005, 69.

hypothesizing) what determines what, these behavioral features cannot be incorporated into a general model.

My *second argument* is that, even if we knew much more about individual behavior, how many specificities do we want the paradigmatic model to take on board? When the objective is to predict the demand for a given product (a new car, for instance) textbook consumer theory is of modest help only. The sales department of large companies know all too well that there are many motivations other than price to buying a product, that a certain share of the market reacts to the pride of driving a new car, while another share carefully reads consumer reports, and so on. By correctly mimicking the reaction of each type of consumer they are able to estimate the potential demand with remarkable precision. However, most researchers would consider that this kind of exercise does not belong to economics as a science.

How has economics dealt with features that do not fit with the assumptions of the core model? For a long time, modern economics has identified "anomalies" such as public goods—the enjoyment of which does not reduce the supply available, such as public TV broadcasting or law-and-order (Samuelson 1954)—, Giffen goods—whose demand increases with its price (Marshall 1895)—, inconsistencies in inter-temporal choices (Strotz 1956), or the social status effects on consumption (Duessenberry 1953). However, the recording of such anomalies in actual behavior did not erode classical theory of rational choice. Rather, economics reacted by developing "auxiliary" models to examine how each such departure from the classical assumptions could modify the intuition derived from the GE equilibrium model.

Classical decision theory is not meant to be descriptive in the literal sense of the word. The contribution of the GE model has not been to produce theories that actually predict anything with any precision, but a new way to think about the world that is truly illuminating. Giving up accuracy for insight is a familiar trade-off in economics, and perhaps the social sciences at large. How far one should go in skipping specificities in behavior is debatable. Should research in economics proceed as in the aforementioned cases and also develop "auxiliary" models while preserving the essence of the GE model as the core paradigm? My position as of today is in the affirmative, at least as long as we cannot neatly identify the exogenous determinants of the observed behavioral patterns.[38]

How realistic a theory has to be?

There is little doubt that, for specific applications, one would like to have as accurate a theory as possible. However, for theoretical applications, such as the derivation of the welfare theorems, it is not obvious that more accurate assumptions result in more accurate, let alone more useful, conclusions. The reason is that theoretical applications use models that are known to be false as a way to sort out and test arguments. Certain assumptions, which are certainly incorrect when tested in a laboratory, can be more useful for certain purposes and less for others. There is a danger that an experimental finding such as framing effect might, when put together with other theoretical assumptions, lead to a result that is less realistic than the assumption that framing does not matter.

Thus, the question we should ask ourselves when we deal with general economic thought is not whether a particular assumption is accurate. Rather, as pointed out by Milton Friedman (Nobel laureate, 1976) long ago, we should ask whether it leads to more accurate conclusions when coupled with other assumptions and, importantly, whether it suggests a reasonable trade-off between accuracy and strength. If we end up rejecting all assumptions, and therefore saying nothing, the accuracy of our models will be of little consolation.

Closing comments

Some researchers have been tempted to interpret the observed deviations in behavior as a challenge to the assumption of rationality. As we have seen, many deviations in behavior are due either to mistakes in processing the information, to framing or to a misunderstanding of the relationship between actions and consequences, or due to temporary perturbations in preferences or in time discounting (provoked by emotions and the like). As pointed out by Gilboa and Schmeidler (2001) and Gilboa et al. (2008), all these deviations have the following in common: if exposed to the analysis of their behavior, decision makers would wish to change their choices. For instance, they would want to eliminate identifiable errors in reasoning or blunt reactions. Thus, what is irrational for a decision maker are those types of behavior that will not be robust to analysis; they are likely to change when talking to an expert or brainstorming the decision with other decision makers. It appears more useful to focus on those deviations from classical theory that pass this test of robustness, that are "rational" in this sense. Other violations are sometimes thought provoking and often amusing, but they need not qualify as a basis for responsible economic analysis.

I wish to conclude this essay with a few words for the immediate future of research in behavioral economics. There is nowadays a burst of departures

[38] I personally find more embarrassing the assumption of competitive behavior when, for instance, the critical oil market is controlled by the OPEC countries.

from the standard rational choice model, all motivated on the grounds of psychological evidence. In Rubinstein's (2008) words, a model "that has at its core fairness, envy, present-bias and the like is by now not only permitted but even preferred." All this variety of departures certainly produce intellectual excitement, but it also produce perplexity and a sense of lack of direction. Every newly identified pathology is cheerfully welcome. In my view, an effort should be made to introduce some order in this chaotic landscape. Research should concentrate on a few types of deviations only. The ones that may be more critical from the perspective of economics —as Gul and Pensendorfer (2008) recommend. Once the implications have been well understood we may move to a further enrichment of our modeling of individual decisions.

Acknowledgements
I am thankful to Xavier Calsamiglia, Itzahk Gilboa, Clara Ponsat, and Debraj Ray for their comments. I retain full responsibility for the opinions expressed here.

Bibliography

Acemoglu, D., and J. A. Robinson. *Economic Origins of Dictatorship and Democracy.* Cambridge University Press, 2007.

Akerlof, G. A., and R. Kranton. "Economics and Identity." *Quarterly Journal of Economics* 115, 2000, 715–753.

Amiel, Y., and F. A. Cowell. "Measurement of income inequality: experimental test by questionnaire." *Journal of Public Economics* 47, 1992, 3–26.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler. "Fact-Free Learning." *American Economic Review* 95, 2005, 1355–1368.

Arrow, K. J. *Social Choice and Individual Values.* John Wiley and sons, 1951.

Arrow, K. J., and F. H. Hahn. *General Competitive Analysis.* North Holland, 1983.

Atkinson, A. B. "On the measurement of inequality." *Journal of Economic Theory* 2, 1970, 244–263.

Banerjee, A. V. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics* 107, 1992, 797–817.

Benabou, R. "Workings of a City: Location, Education, and Production." *Quarterly Journal of Economics* 108, 1993, 619–652.

—. "Equity and Efficiency in Human Capital Investment: The Local Connection." *Review of Economic Studies* 62, 1996, 237–264.

Benabou, R., and J. Tirole. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121, 2006, 699–746.

Besley, T., and S. Coate. "Understanding Welfare Stigma: Taxpayer Resentment and Statistical Discrimination." *Journal of Public Economics* 48, 1992, 165–183.

Blundell, R., W. K. Newey, and T. Persson (eds.). *Advances in Economics and Econometrics Theory and Applications.* Vol. 2. Cambridge University Press, 2006.

Bolton, G. E., and A. Ockenfels. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90, 2000, 166–193.

Calvo-Armengol, A., and Y. Ioannides. "Social Networks in Labor Markets." In L. Blume and S. Durlauf (eds.), *The New Palgrave Dictionary of Economics.* Palgrave MacMillan Press, 2005.

Camerer, C. F., and G. Loewenstein. "Behavioral Economics: Past, Present, Future." In C. Camerer, G. Loewenstein, and M. Rabin (eds.), *Advances in Behavioral Economics.* Princeton University Press, 2003.

Caplin, A., and A. Schotter (eds.). "Foundations of Positive and Normative Economics," *Methodologies of Modern Economics*, Vol. 1. Oxford University Press, 2008, forthcoming.

Cervellati, M., J. Esteban and L. Kranich. "Work Values, Endogenous Sentiments and Redistribution." 2008, unpublished.

Downs, A. *An Economic Theory of Democracy*. Harpers and Bros., 1957.

Debreu, G. *Theory of Value*. Yale University Press,1959.

Vigna, S. della. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature*. 2008, forthcoming.

Duesenberry, J. S. *Income, Saving and the Theory of Consumer Behavior*. Cambridge, Massachusetts: Harvard University Press, 1949.

Esteban, J., and D. Ray. "On the Measurement of Polarization." *Econometrica* 62, 1994, 819–852.

—. "Conflict and Distribution." *Journal of Economic Theory* 87, 1999, 379–415.

—. "On the Salience of Ethnic Conflict." *American Economic Review* 98, 2008, forthcoming.

Falk, A., and U. Fischbacher. "A Theory of Reciprocity." *Games and Economic Behavior* 54, 2006, 293–315.

Fearon, J. "Rationalist Explanations for War." *International Organization* 49, 1995, 379–414.

Fehr, E., and K. M. Schmidt. "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics* 114, 1999, 817–868.

Fryer, R., and M. O. Jackson. "A Categorical Model of Cognition and Biased Decision Making." *The B.E. Journal of Theoretical Economics*: Vol. 8 : Issue 1 (contributions), 2008, article 6.

Fox, C. R., and A. Tversky. "Ambiguity Aversion and Comparative Ignorance." *Quarterly Journal of Economics* 110, 1995, 585-603.

Gilboa, I. and D. Schmeidler. *A Theory of Case-Based Decisions*. Cambridge University Press, 2001.

Gilboa, I., A. Postlewaite, and D. Schmeidler. "Probabilities in Economic Modeling." 2007, unpublished.

Gilboa, I., F. Maccheroni, M. Marinacci, and D. Schmeidler "*Objective and Subjective Rationality in a Multiple Prior Model*." 2008, unpublished.

Goyal, S. *Connections: An Introduction to the Economics of Networks*. Princeton University Press, 2007.

Grossman, G. M. and E. Helpman. *Special Interest Politics*. The MIT Press, 2001.

Grossman, H. I. "A General Equilibrium Model of Insurrections." *American Economic Review* 81, 1991, 912–921.

Gul, F., and W. Pesendorfer. "Temptation and Self-Control." *Econometrica* 69, 2001, 1403–1435.

—. "The Case for Mindless Economics." In Caplin and Schotter (eds.), 2008, forthcoming.

Hirshleifer, J. "The Paradox of Power." *Economics and Politics* 3, 1991, 177–200. Jackson, M.O. (2008a) "Social and Economic Networks," Princeton University Press.

—. "Network Formation." In L. Blume and S. Durlauf (eds.) *The New Palgrave Dictionary of Economics*, Palgrave MacMillan, 2008b.

Kahneman, D., and A. Tversky. Prospect Theory: An Analysis of Decision under Risk," *Econometrica 47*, 1979, 263–291.

—. "Choices, values and frames." *American Psychologist 39*, 1984, 341–350.

Laibson, D. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112, 1997, 443–77

Laibson, D., and R. Zeckhauser. "Amos Tversky and the Ascent of Behavioral Economics." *Journal of Risk and Uncertainty* 16, 1998, 7–47.

Levine, D. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1, 1998, 593–622.

Levitt, S. A. D., and J. A. List. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21, 2007, 153–174.

Lindbeck, A., and S. Nyberg. "Raising Children to Work Hard: Altruism, Work Norms and Social Insurance." *Quarterly Journal of Economics*, 121, 2006, 1473–1503.

Lindbeck, A., S. Nyberg, and J. Weibull. "Social Norms and Economic Incentives in the Welfare State." *Quarterly Journal of Economics* 114, 1999, 1–35.

Mas-Colell, A. "The Future of General Equilibrium." *Spanish Economic Review* 1, 1999, 207–214.

Mas-Colell, A., M. D. Whinston, and J. R. Green *Microeconomic Theory*. Oxford University Press, 1995.

Moffit, R. "An Economic Model of Welfare Stigma." *American Economic Review* 73, 1983, 1023–1035.

Neumann, J. von, and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Persson, T., and G. Tabellini. *Political Economics. Explaining Economic Policy*. MIT Press, 2000.

—. *The Economic Effects of Constitutions*. MIT Press, 2003.

Piketty, T. "Social mobility and redistributive politics ." *Quarterly Journal of Economics* 110, 1995, 551–584.

Powell, R., *In the Shadow of Power: States and Strategies in International Politics*, Princeton University Press, 1999.

Rabin, M. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83, 1993, 1281–1302.

Rabin, M., and R. Thaler "Anomalies: Risk Aversion." *Journal of Economic Perspectives* 15, 2001, 219–232.

Ray, D. *Development Economics*. Princeton University Press, 1998.

—. "Development Economics." In L. Blume (ed.) *The New Palgrave Dictionary of Economics*. Palgrave MacMillan, 2008a.

—. "Inevitable Costly Conflict," unpublished, 2008b.

Rook, L. "An Economic Psychological Approach to Herd Behavior." *Journal of Economic Issues* 40, 2006, 75–95.

Rubinstein, A. "Discussion of Behavioral Economics," in Blundell et al. (2006).

Samuelson, P. A. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics*, 36, 1954, 387-389.

Skaperdas, S. "Cooperation, Conflict, and Power in the Absence of Property Rights." *American Economic Review* 82 (4), 1992, 720–739.

Stiglitz, J. "Information and the Change in the Paradigm in Economics." *American Economic Review* 92, 2002, 460–501.

Strotz, R. H. *"Myopia and Inconsistency in Dynamic Utility Maximization." Review of Economic Studies* 23, 1956, 165-180.

Tabellini, G. "The Scope of Cooperation: values and incentives." *The Quarterly Journal of Economics*. 2007, forthcoming.

Tirole, J. *The Theory of Industrial Organization*. MIT Press, 1988.

Vives, X. *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press, 2001.

radical innovations: an economist’s view

NATHAN ROSENBERG

Why is there no economics of science?

I would like to begin with this question: Why is there still no recognized discipline called "The Economics of Science"? After all, economics as a discipline has shown strong imperialistic tendencies in recent decades. It has successfully colonized many fields, but it has yet to colonize science. We now have an economics of education, an economics of health, an economics of voting behavior, an economics of marriage, an economics of divorce, and an economics of crime. With respect to the latter, the economics of crime, it turns out that crime pays, especially when, as is often the case, the likelihood of apprehension and punishment is low! As some indication of the elevated status of this kind of research, one of its most eminent practitioners (Gary Becker) was awarded a Nobel Prize in Economics.

Why then do we not now have an economics of science—or, rather, since it is now just beginning to happen, why did it take so long? This question is particularly pertinent in view of what we think we have long known about science. That is to say, it has long been an article of faith that scientific research yields huge economic benefits.

There is at least a partial answer that suggests itself to the question of why an economics of science has taken so long to emerge: economics is a discipline that studies the principles involved in achieving an efficient use of scarce resources. But to talk about efficiency of resource use requires an ability to make some explicit comparison of costs and benefits. Now, we do in fact know a great deal about the costs of linear accelerators, synchrotron radiation machines, Hubble telescopes, the mapping of the human genome, etc. Indeed, some years ago the US Congress decided to cancel the construction of a superconducting supercollider when the costs threatened to escalate up to $11 or $12 billion. (In fact, it cost well over $1 billion just to close down the project!)

But, while it is relatively straightforward to calculate the costs of conducting science, it is extraordinarily difficult to calculate the benefits. And if one insists on considering only the narrow economic benefits, it would be difficult to make any sort of case at all for some of the projects of so-called "Big Science." (What purely economic case can be made for the Hubble Telescope?)

Now, it is of course true that the history of scientific research in the twentieth century was full of unexpected benefits that have flowed from scientific research. But the general acknowledgment of the likelihood of unanticipated benefits hardly constitutes a guide to determining the size of the annual public subsidy to science, or the allocation of a budget of given size among the many competing possible uses in different fields of science. In a nutshell, the uncertainties concerning the possible benefits of basic scientific research are simply immense, and it is difficult to make a rigorous application of economic principles in a realm where the benefits of resource use are essentially unmeasurable.

What also follows from what has been said is that, in order to think in a useful way about science and technology in modern society—including the society of the twenty-first century—it is necessary to acknowledge that one must unavoidably learn to live with a high level of uncertainty. Nevertheless, I would like to insist that this need not diminish the usefulness of economic analysis, at least so long as we do not harbor unreasonable expectations about what can be achieved by abstract reasoning alone. For economic analysis alone can never provide a neatly packaged solution to policy-making with respect to the extremely complex issues with which we are concerned. Nor should we expect it to do that. But it can be an invaluable guide in looking for significant cause-effect relationships, in trying to understand how institutions and incentives shape human behavior, and in attempting to make sense of an immense body of historical and empirical data that are available to serious scholars, and from which important lessons for policy-making and institution-building might be derived.

Institutional changes in the twentieth century

If one looks back upon the last one hundred years and asks what were the distinctive features that dominated the realm of economic activity, my first reply would be that it was the application of scientific knowledge and scientific methodology to a progressively widening circle of productive activities. But this statement, by itself, is not very informative. In fact, it can serve only as a platform from which to raise other, deeper questions: In precisely what ways has science played this role? Which aspects of the scientific enterprise have played the role, and under what circumstances? And what were the changes in the manner in which science was institutionalized in the course of this century that made the twentieth century so different from the nineteenth?

A dominant factor, of course, was that, in the years after the World War II, national governments in industrial countries became, in varying degrees, the patrons of scientific research, especially of basic research. In considerable measure this reflected the critical role that science, and scientists, had played in shaping the conduct and the outcome of the war, culminating with the horrific weapon forged by the Manhattan Project that brought the war in the Pacific to an abrupt conclusion. The Cold War served as a further, powerful rationale for massive government contributions to the support of science, which dwarfed all pre-war precedents. But there were also powerful and quieter forces at work.

It may be useful here to recall Alfred North Whitehead's oft-quoted observation that: "The greatest invention of the nineteenth century was the invention of the method of invention." (Whitehead 1925, 98.) The twentieth century, of course, was not only to inherit, but also to institutionalize, that method of invention. Whitehead understood that this invention involved the linking of scientific knowledge to the world of artifacts. But he also understood that this linkage was not easily achieved, because a huge gap typically exists between some scientific breakthrough and a new product or process. Although the sentence just quoted from Whitehead's book, is well known, his subsequent observation is not, but deserves to be: "It is a great mistake to think that the bare scientific idea is the required invention, so that it has only to be picked up and used. An intense period of imaginative design lies between. One element in the new method is just the discovery of how to set about bridging the gap between the scientific ideas, and the ultimate product. It is a process of disciplined attack upon one difficulty after another." (Whitehead 1925.)

What appears to matter more than the quality of a country's basic science, as judged by the usual academic or Nobel Prize Committee criteria, is the extent to which the activities of the scientific community can be made to be responsive to the needs of the larger society. It is regrettable that this is a question that is not very much discussed, and is poorly understood. It is often obscured by much of the rhetoric of academic science, with its overwhelming emphasis on the importance of the independence and the autonomy of the individual scientist. The fact of the matter is that, in the course of the twentieth century, and with varying degrees of success, industrial societies have created increasingly dense networks of institutional connections between the conduct of scientific research and the needs of the larger social system.

Within the university world, this includes a number of engineering disciplines that emerged late in the nineteenth and twentieth centuries, such as electrical engineering, chemical engineering, aeronautical engineering, metallurgy, and computer science. Indeed, although it is not widely realized, in recent years government R&D expenditures at American universities devoted to the engineering disciplines have exceeded expenditures devoted to the physical sciences. Far and away the largest recipients in the most recent years were the life sciences, receiving more than 50% of federal financial support.

In addition to new academic disciplines, the other key institutional innovation of the twentieth century was, of course, the industrial research laboratory. These laboratories monitored frontier research within the university community and elsewhere, although for many years it was the application of relatively elementary scientific concepts and methodologies that dominated their contributions to industry. In the course of the century, however, and especially after the World War II, research at many of these laboratories became increasingly sophisticated. By 1992 the Directory of American Research and Technology counted about 12,000 non-government facilities that were active in some form of "commercially-applicable" scientific research. And, according to the National Science Foundation's figures, more than 30% of all basic research in the US was financed by private industry.

The industrial research laboratory is essentially an institutional innovation in which the scientific research agenda is largely shaped by the needs of industrial technologies. The role of industrial scientists is to improve the performance and reliability of those technologies, as well as, of course, inventing entirely new ones. Thus, the industrial research laboratory has rendered science more and more an institution whose directions are increasingly shaped by economic forces and concentrated on the achievement of economic goals. Science has become gradually incorporated, in the course of the twentieth century, into a crucial part of the growth system that has propelled industrial societies along their long-term growth trajectories.

That growth system, in which technological change played a central role for two centuries, is now reinforced by a powerful scientific research capability that has strong public and private components, varying among countries according to their histories, cultures, their political systems and their current social priorities. For further details, see United Nations Industrial Development Organization, Industrial Development Report 2005, *Capability building for catching-up, Historical, empirical, and policy dimensions* (Vienna, 2005).

In addition to the institutional requisites, the successful exploitation of scientific knowledge has flourished best in industrial countries that have offered potential innovators ready access to capital as well as strong financial incentives, and have nourished and educated effective managerial and engineering cadres. Thus, nineteenth-century Czarist Russia produced many brilliant scientists and inventors, but their presence exercised a negligible impact in a society that lacked an adequate managerial, engineering and financial infrastructure. On the other hand, America's emergence to a position of technological leadership in a number of industrial sectors, before the World War I, occurred in a period when its achievements in basic science were limited and, with few exceptions, of no great international consequence. In this respect America in the late-nineteenth and early-twentieth centuries bears some interesting resemblances to Japan in the second half of the twentieth century. Both countries managed to achieve rapid industrial growth with no more than a modest scientific base because of their great aptitude for borrowing and exploiting foreign technologies.

On the other hand, the relative stagnation of the British economy in the twentieth century has occurred in spite of continued brilliant performances at the scientific frontier. Until not very long ago the British scientific community continued to receive more Nobel Prizes per capita than the United States. But, at the same time, the British failed to maintain competitiveness even in many inventions that had originated in Britain—radar, the jet engine, penicillin, and the CT scanner. Moreover, the revolution in molecular biology that began with the discovery of the double helical structure of the DNA molecule in the 1950s was, to a remarkable degree, a British achievement—indeed, a Cambridge University achievement. Nevertheless, British firms played only a minor role in the emerging biotechnology industry, while there were several hundred biotechnology firms in the US, including the very small number of such firms that quickly enjoyed some degree of commercial success.

I wish to draw two conclusions. Looking over the entire course of the twentieth century, scientific achievement alone, however brilliant, was not readily translated into superior economic performance. Strong complementary institutions and incentives have been necessary, not the least of which has been venture capital firms. Moreover, when such institutions and incentives have been present, even a comparatively modest scientific capability has been sufficient to generate high levels of economic performance.

The endogeneity of science

I have argued that the institutional changes of the twentieth century have rendered science a more endogenous activity. I mean this in the specific sense that, where such institutional innovations have occurred, science has come to be more directly responsive to economic forces. But I must now expand upon a particular aspect of that observation. That is, I want to suggest that the research agenda of science has been more and more determined by the need to improve the performance of technologies that were already in existence. In the twentieth century world, science and technology have become intimately intertwined. Science has indeed come to play an expanding role in influencing the world of technology, but causality has worked in both directions: the scientific enterprise of the twentieth century also needs to be explained in terms of its responses to the needs, and the exigencies, of technology.

In fact, a major, neglected theme in twentieth century science is that prior progress in the technological realm has come to play a critical role in formulating the subsequent research agenda for science. The natural trajectory of certain technological improvements has served to identify and to define the limits to further improvement, which, in turn, has served as a focusing device for subsequent scientific research.

Consider the aircraft industry. In this industry, improved performance continually brought the technology—the aircraft—to performance ceilings that could be pierced only by understanding some aspects of the physical world better. As a result, the introduction of the turbojet had a profound impact upon science as well as upon the aircraft industry, by progressively pushing against the limits of scientific frontiers and by identifying the specific directions in which this new knowledge had to be further enlarged before additional technological improvements could occur.

Thus, the turbojet first led to the creation of a new specialty, supersonic aerodynamics, "…only to give way," according to one authority, "to aerothermodynamics as increasingly powerful turbojets pushed aircraft to speeds at which the generation of heat on the surface of the aircraft became a major factor in airflow behavior. Eventually, turbojet powered aircraft would reach speeds at which magnetothermodynamic considerations would become paramount: [that is to say] temperatures would become so great that air would dissociate into charged submolecular ions." (Constant1990, 240.) Thus, the greater speeds made possible by jet engines also required advancing the frontiers of scientific knowledge in order to be able to accommodate the design requirements of high speed jet aircraft.

I suggest that a central feature of high technology industries is that this kind of sequence has become prominent. That is, technological progress serves to identify, in reasonably unambiguous ways, the directions in which scientific research needs to be conducted, and at the same time it holds out the prospect of a large financial return should the research prove to be successful.

The mechanisms at work may take a variety of forms. In the case of the jet engine, functioning at increasingly high speeds, the technology pointed to specific natural phenomena in a specific environment. In the telephone industry, on the other hand, transmission over longer distances, or the introduction of new modes of transmission, have been particularly fruitful mechanisms in the generation of basic research. For example, in order to improve overseas transmission by radiotelephone it was essential to develop an expanded appreciation for the ways in which electromagnetic radiation interacts with various atmospheric conditions. Indeed, some of the most fundamental of all scientific research projects of the twentieth century have been direct outgrowths of the attempt to improve the quality of transmission of sound by telephone. Dealing with various kinds of interference, distortion or attenuation of electromagnetic signals that transmit sound has profoundly enlarged our understanding of the universe.

Two fundamental scientific breakthroughs, one by Karl Janskyin the late 1920s and another more recently by Penzias and Wilson, both occurred as a result of attempts to improve the quality of telephone transmission. This involved, specifically, dealing with sources of noise. Jansky had been put to work to deal with the problems of radio static after the opening up of the overseas radiotelephone service. He was provided with a rotatable radio antenna with which to wor. In 1932 he published a paper identifying three sources of noise: from local thunderstorms, from more distant thunderstorms, and a third source, which Jansky identified as "a steady hiss static, the origin of which is not known." It was this "star noise," as it was first called, that marked the birth of an entirely new science: radio astronomy, a discipline that was to prove one of the greatest sources of scientific advance of the twentieth century.

Jansky's experience underlines one of the reasons why the attempt to distinguish between basic research and applied research is extremely difficult to carry out consistently. Fundamental breakthroughs often occur while dealing with very mundane or practical concerns. Attempting to draw that line on the basis of the motives of the person performing the research—whether there is a concern with acquiring useful information (applied) as opposed to a purely disinterested search for new knowledge (basic)—is, in my opinion, a hopeless quest. Whatever the ex ante intentions in undertaking research, the kind of knowledge actually acquired is highly unpredictable. This is in the nature of serious scientific research. Historically, some of the most fundamental scientific breakthroughs have come from people, like Jansky, who certainly thought that they were doing very applied research.

Bell Labs' fundamental breakthrough in astrophysics was also directly connected to improving telephone transmission, and especially in the use of communication satellites for such purposes. At very high frequencies, rain and other atmospheric conditions became major sources of interference in transmission. This source of signal loss was a continuing concern in the development of satellite communication. It led to a good deal of research at both the technological and basic science levels—e.g., the study of polarization phenomena (Dinn l977, 236–242).

Arno Penzias and Robert Wilson first observed the cosmic background radiation, which is now taken as confirmation of the "big bang" theory of the formation of the universe, in 1964 while they were attempting to identify and measure the various sources of noise in their receiving system and in the atmosphere. They found that: "The radiation is distributed isotropically in space and its spectrum is that of a black body at a temperature of 3 degrees Kelvin." (Fagen 1972, 87.) Although Penzias and Wilson did not know it at the time, the character of this background radiation was precisely what had been predicted earlier by cosmologists who had formulated the "big bang" theory. They subsequently received a Nobel Prize for this momentous finding.

There is, I am suggesting, a compelling internal logic to certain industries, e.g., the telephone system, that forcefully points the research enterprise in specific directions. Consider further some of the material needs of that system. The invention of the transistor and the discovery of the transistor effect were the results of a deliberate attempt to find a substitute for the vacuum tube in the telephone industry. The vacuum tube was unreliable and generated a great deal of heat. After the transistor had been invented, its actual production required standards of material purity that were quite without precedent for industrial purposes. Since transistor action was dependent on introducing a few foreign atoms to the semiconducting crystal, remarkably high standards of semiconductor purity had to be attained. Something of the order of a single foreign atom for each 100,000,000 germanium atoms meant that the telephone system simply had to attain levels of purity that presupposed a good deal of fundamental research into the structure and behavior of materials, especially crystallography.

The invention of the transistor in 1947 had an enormous impact on the direction of scientific research. Solid state physics had attracted only a very small number of physicists before the arrival of the transistor. In fact, before the World War II, it was a subject that was not even taught in most American universities. However, there was a huge redirection of scientific resources within a few years after the announcement of the transistor effect. In fact, within a matter of years, rather than decades, solid-state physics had become the largest subdiscipline of physics. The huge mobilization of scientific resources in this field, in universities as well as private industry, was clearly a response to the potentially high payoffs to such research that were signaled by the arrival of the transistor.

The growth of the telephone system also meant that equipment and components had to perform under extreme environmental conditions, from geosynchronous satellites to transatlantic cables. These extreme environmental conditions have one particularly important consequence: there are likely to be severe economic penalties for failing to establish very high standards of reliability. There are compelling reasons for the attainment and maintenance of high standards that are absent in, say, consumer electronics, not to mention a brick factory. The failure of a submarine cable, once placed on the ocean floor, involves extremely high repair and replacement costs in addition to a protracted loss of revenue. Similarly, communication satellites had to be remarkably reliable and strong simply to survive the act of being launched and placed into orbit. The instrumentation had to survive extremes of shock, vibration, temperature range, radiation, etc.

Thus, high standards of reliability are not a marginal consideration but the very essence of successful economic performance in this industry. This consideration had a great deal to do with the high priority that Bell Labs attached to materials research over a period of several decades. Important advances in polymer chemistry, for example, were achieved at Bell Labs in order to understand the morphology of polyethylene, because of premature failure of cable sheathing employing this material on the floor of the Atlantic Ocean.

The importance of high standards of reliability has also been a basic underlying condition in the thrust of research in other specific directions.

The decision to undertake a basic research program in solid state physics, which culminated in the development of the transistor, was strongly influenced, as suggested earlier, by these (as well as other) sources of dissatisfaction. But the transistor suffered from reliability problems of its own in its early years. These problems emerged in the early 1950s as the transistor experienced a widening range of applications. The defects were eventually linked to certain surface phenomena. As a result, a major research thrust into the basic science of surface states was undertaken that eventually solved the reliability problems but, in doing so, also generated a great deal of new fundamental knowledge in surface physics.

The development of optical fibers is particularly apposite to our present concerns. Although its attractiveness as a new mode of transmission was increased by space and congestion constraints, its feasibility was rooted in another set of technological breakthroughs of the 1950s. It was the development of laser technology that made it possible to use optical fibers for transmission. This possibility, in turn, pointed to the field of optics, where advances in knowledge could now be expected to have high financial payoffs. As a result, optics as a field of scientific research has had a great resurgence in the last few decades. It was converted by changed expectations, based upon past and prospective technological innovations, from a relatively somnolent intellectual backwater to a burgeoning field of scientific research. The causes were not internal to the field of optics but were based upon a radically altered assessment of new technological possibilities—which, in turn, had their roots in the earlier technological breakthrough of the laser.

This discussion has implications, I believe, that are of fundamental importance to an understanding of the economic role of science in the twentieth century. Although the impact of new scientific knowledge upon industry is continually emphasized in public discussions, very little attention is devoted to causal forces flowing in the opposite direction. But modern high technology industries set in motion immensely powerful forces that stimulate and influence scientific research. It does this in several ways: by providing observations, or formulating problems that could only have occurred in specific industrial contexts, such as the telephone or the aircraft industry; by providing new techniques of instrumentation that vastly enlarge the observational, measurement and calculating capabilities of the scientist; and most important of all, by raising the economic payoff to the performance of scientific research and therefore powerfully increasing the willingness of private industry to finance such research.

It should be understood that the remarkable accomplishments at Bell Labs in the twentieth century were by no means typical of other sectors of American industry—indeed it was quite unique in many respect—but many other American firms developed strong scientific capabilities of great economic value, an assertion that is reinforced by an earlier assertion that there were somewhere around 12,000 industry laboratories in the US in 1992.

A fair generalization is that American firms learned how to exploit scientific knowledge and methodology, and to link these forces through organization and incentives, and they managed these more successfully than did other OECD countries.

The increasingly multidisciplinary nature of research (and innovation)

There is another feature of the scientific enterprise that demands attention because of its important implications for the future. The multidisciplinary nature of research in the realms of both science and technology, increasingly apparent in the second half of the twentieth century, will doubtless intensify in the next century.

History suggests that the crossing of disciplinary boundaries is not something that usefully emerges from some kind of deliberate plan, strategy, or committee meeting; rather, it is something that occurs, when it does occur, because of the peculiar logic of scientific progress. It has happened, historically, when certain problems emerged at the frontier of a particular discipline, such as cell biology, that required a better understanding of the role of certain processes that were the specialty of scientists in a different discipline, e.g., chemistry. The outcome, biochemistry, has thus been a natural outgrowth of the changing requirements of an expanding body of research knowledge. Similarly, geophysics emerged as an independent subdiscipline of geology when it became possible to apply the methodologies, that had first developed in physics, to the understanding of the structure and the dynamics of the earth, as well as, eventually, other planets. Here, as on other occasions, the introduction of new technologies of instrumentation has led to a beneficial crossing of certain disciplinary boundary lines.
The established lines between physics and chemistry have been crossed on a number of occasions in the past for similar reasons.

The increasing importance of the ability to exploit the knowledge and the methodologies of more than one discipline has become apparent not only at the level of basic science but in applied sciences and engineering as well. In recent years, medical science has benefited immensely, not only from such "nearby" disciplines as biology and chemistry, but from nuclear physics (magnetic resonance imaging, radioimmunoassays), electronics, and materials science and engineering. In pharmaceuticals there have been remarkable advances deriving from such fields as biochemistry, molecular and cell biology, immunology, neurobiology, and scientific instrumentation. These advances are moving toward the possibility that new drugs, with specific properties, can be targeted and perhaps one day even designed, in contrast to the randomized, expensive, and exhaustive screening methods that have characterized pharmaceutical research in the past (Gambardella l995). The new pattern of innovation is, by its very nature, highly multidisciplinary. Success requires close cooperation among an increasing number of specialists: chemists, biochemists, pharmacologists, computer scientists. What is most certain is that the biological sciences will play an increasingly pivotal role in drug discovery and development. This is also apparent in the emerging biotech industry, which is still in its infancy. This industry draws on many scientific disciplines, including cell biology, molecular biology, protein chemistry, and biochemistry.

This sort of close cooperation among specialists from different disciplines has already accounted for some of the most important breakthroughs of the last fifty years. The transistor was the product of cooperation among physicists, chemists, and metallurgists. The scientific breakthrough leading to the discovery of the structure of DNA was the work of chemists, biologists, biochemists and crystallographers. More productive rice varieties, that have transformed the population-carrying capabilities of the Asian continent, were originally developed at the International Rice Research Institute in the Philippines, through the combined efforts of geneticists, botanists, biochemists, entomologists, and soil agronomists.

The increasing value of interdisciplinary research creates serious organizational problems for the future. Such research often runs counter to the traditional arrangements, training, intellectual priorities, and incentive structures of the scientific professions, particularly in the academic world, where tremendous emphasis is placed upon working within well-recognized disciplinary boundary lines. Department-based disciplines have played a crucial role in teaching and research, and are certainly not to be discarded casually. Historically, disciplines emerged because, within their boundaries, there was a set of problems that could be solved by some common conceptualization, analytical framework or methodology. Workers within a discipline spoke a common language; and, not least important, the discipline provided a basis for forming judgments about the quality of research. In this respect, commitment to a particular discipline provided some standards for quality control.

Although great (and justifiable) concern is currently being expressed over the future financial support of universities, organizational issues may also become increasingly worrisome as a rigid departmentalism comes to confront a research frontier requiring more and more frequent crossing of traditional disciplinary boundaries. Such problems, it is worth observing, are not likely to be nearly so serious in private industry, where disciplinary boundaries do not loom nearly so large, and where the highest priorities are problem-solving, improving the performance of existing technology, and, ultimately, generating higher profits, regardless of the disciplinary sources through which these goals can be attained.

The persistence of uncertainty

There is a final issue that needs to be addressed, and that is the persistence of uncertainty, not only in the realm of science, where it is universally acknowledged, but in the realm of technology as well. We are accustomed to expect a high degree of uncertainty and unanticipated developments in the world of scientific research. It is widely assumed, however, that uncertainties decline as one moves across the spectrum of activities from basic research to applied research to product design and development and, finally, to the commercialization of the new product in the market place.

It is, of course, true that some uncertainties have been resolved after a new technological capability has been established, and even after its first acceptance in the market place, the questions change, and it is far from obvious that the new questions are any less complex than the old ones. The most fundamental of all questions is, to what social purposes will the new capability be put?

It appears that no one anticipated the invention of the Internet; rather, it simply "appeared" after a sufficient number of computers were in existence. As David Mowery observed in a fascinating article: "The Internet is the world's largest computer network—a steadily growing collection of more than 100 million computers that communicate with one another using a shared set of standards and protocols. Together with the World Wide Web, a complementary software innovation that has increased the accessibility and utility of the network, the Internet stimulated a communications revolution that has changed the way individuals and institutions use computers in a wide variety of activities." (Moweryand and Simcoe 2002.)

Consider the laser, an innovation that is certainly one of the most powerful and versatile advances in technology in the twentieth century, and one that may still be moving along a trajectory of new applications. Its range of uses in the fifty years since it was invented is truly breathtaking. This would include precision measurement, navigational instruments, and a prime instrument of chemical research. It is essential for the high quality reproduction of music in compact discs (CDs). It has become the instrument of choice in a range of surgical procedures, including extraordinarily delicate surgery upon the eye, where it has been used to repair detached retinas, and gynecological surgery where it now provides a simpler and less painful method for removal of certain tumors. It is extensively employed in gall bladder surgery. The pages of my manuscript were printed by an HP laser jet printer. It is widely used throughout industry, including textiles where it is employed to cut cloth to desired shapes, and metallurgy and composite materials where it performs similar functions. But perhaps no single application of the laser has been more profound than its impact on telecommunications where, together with optical fibers, it is revolutionizing transmission. The best transatlantic telephone cable in 1966 could carry only 138 simultaneous conversations between Europe and North America. The first fiber optic cable, installed in 1988, could carry 40,000. The fiber optic cables installed in the early 1990s can carry nearly 1.5 million conversations. And yet it is reported that the patent lawyers at Bell Labs were initially unwilling even to apply for a patent on the laser, on the grounds that such an invention, dealing with the realm of optics, had no possible relevance to the telephone industry. In the words of Charles Townes, who subsequently won a Nobel Prize for his research on the laser, "Bell's patent department at first refused to patent our amplifier or oscillator for optical frequencies because, it was explained, optical waves had never been of any importance to communications and hence the invention had little bearing on Bell System interests." (Townes 1968, 701.)

The transistor was, without doubt, one of the greatest achievements of the twentieth century—or, for that matter, any century. Consequently, one might expect to find the announcement of its invention, in December 1947, displayed prominently on the front page of the *New York Times*. Nothing of the sort. When it was finally mentioned in the *Times*, it appeared only as a small item buried deep in that newspaper's inside pages, in a regular weekly column titled "News of Radio." Hardly any future uses were mentioned beyond improved hearing aids.

This enumeration of failures to anticipate future uses and large markets for some of the most important inventions of the twentieth century—laser, computer, transistor—could be extended almost without limit. We could, if we liked, amuse ourselves indefinitely at the failure of earlier generations to see the obvious, as we see it today. But that would be, I believe, a mistaken conceit. I am not particularly optimistic that our ability to overcome the uncertainties connected with the uses of new technologies is likely to improve.

Similarly, a main reason for the modest future prospects that were being predicted for the computer in the late 1940s was that transistors had not yet been incorporated into the computers of the day. Introducing the transistor, and later integrated circuits, into computers were, of course, momentous events that transformed the computer industry. Indeed, in one of the most extraordinary technological achievements of the twentieth century, the integrated circuit eventually became a computer, with the advent of the microprocessor in 1970. The world would be a far different place today if computers were still operating with vacuum tubes.

Bibliography

Constant, Edward W. *The Origins of the Turbojet Revolution*. Baltimore: John Hopkins University Press, 1990, 240.

Dinn, Neil F. "Preparing for Future Satellite Systems," *Bell Laboratories Record*, October 1977, 236–242.

Fagen, M. D., ed. *Impact, Bell Telephone Laboratories*, 1972, 87.

Gambardella, Alfonso. *Science and Innovation in the US Pharmaceutical Industry*. Cambridge University Press, 1995.

Mowery, David and Timothy Simcoe. "Is the Internet a US Invention?" *Research Policy*. December 2002. Vol. 31:1369–1387.

Townes, Charles. "Quantum Mechanics and Surprise in the Development of Technology." *Science*, February 16, 1968, 701.

Whitehead, A. N. *Science and the Modern World*. Macmillan, 1925, 98.

why fighting poverty is hard

ABHIJIT V. BANERJEE

One reason anti-poverty policy has not worked better than it has is because we went into it naively, without enough of an understanding of what makes it hard.[1] This essay addresses what I have learnt about this question from my own research, most of which, is based in India.

Finding the poor

Who are the poor?

Suppose someone wants to help the poor. How would he find them? A part of the problem is inevitable: "poor" is an invented category, like tall or beautiful. While we often have a sense of what we mean when we talk about the poor, getting to an operational definition of poverty requires making many rather arbitrary choices. For example, even if we were prepared to bite the bullet and say that people who are below a certain level ("the poverty line") are the poor and the rest are not, we would not know how to set that critical level. For one, the level of what? Income, consumption, and wealth are the obvious candidates, but one could no doubt think of others. Of these income might seem the most natural, till one starts worrying about the challenges of measuring incomes: after all, incomes vary a lot, especially for the poor who tend not to have salaried jobs, and some of that day-to-day or month-to-month variation is expected or even deliberate (think of the vendor who takes a day off each week) and does not affect what they can buy or consume (because they spend out of their savings or borrow). In other words we run the danger of calling the vendor poor because we measured his income on his off day.

Averaging over longer periods of time obviously helps us here, but creates other problems. People are not very good at remembering what happened several weeks or months ago, especially if there is a lot of underlying variation. Moreover, it turns out people have a very hard time figuring out what their own incomes are (unless they are salary earners, and even then they may not know value of the benefits that come with the job). This is in part because they have both inflows and outflows (i.e. earnings as well as costs), and these do not happen at the same time (so you have to figure out how to make them comparable).

For these reasons many economists favor using measures of consumption, which clearly varies a lot less than income (reflecting people's inclination to avoid large swings in their consumption) and therefore is closely related to average income over the period. This comes with its own limitations: we systematically underestimate the well-being of those who are saving

[1] The case for this claim is made in Banerjee (2007).

a lot compared to those who do not save, even though the latter might have a better future facing them. Dealing with health spending poses yet another problem: should we exclude health expenditures when we calculate consumption on the grounds that this is a compulsion and not a choice, or include it because it shows that this family is able to deal with its health problems (whereas an even poorer family might have to resign itself to endure the ill-health).

Measuring consumption, though probably easier than measuring income (mainly because people tend to have relatively stable consumption patterns and therefore you get a reasonable idea by asking them how they spent money over the recent past) is also far from straightforward. For one it can be extremely time consuming: people have a hard time recalling what they consumed in the last week unless you prompt them by specifically going through the entire list of goods they could have consumed and asking them about each of them separately. Consumption decisions are also "gendered": Men usually know more about how much they spent on fixing up the house, while women are often much better informed about the price of onions. As a result you may need to poll more than one person in each household to get an accurate picture of its consumption spending.

The practice of identification

Given how time-consuming and painstaking one needs to be to do either income or consumption measurement right, it is perhaps no surprise that most governments in developing countries take a more rough and ready approach to the problem of identifying the poor. Instead of looking for direct measures of consumption or income, they typically use what are called proxy means tests. In a proxy means test, each family gets scored based on a relatively small number of what are believed to be good proxies for the family's standard of living. The identification of the BPL (Below Poverty Line) population in India, for example, is based on a scoring rule which puts some weight on measures of family wealth (ownership of land, kind of house, whether the house has indoor plumbing, etc.), some direct measures of well-being (such as whether you have two square meals a day), some measures of earning capacity (education of the adults, type of job they hold, etc.) and some indices as to what one might call behavioral responses to poverty (whether children are in school, working, etc.). Mexico's flagship welfare program, now called *Oportunidades*, uses a very similar index to identify potential beneficiaries: the index they use is a weighted mean of the number of people per room in a household, the age of the household head, the dependency ratio, the level of schooling and occupation of the household head, the number of children ages 5–15 not attending school, the number of children under 12 years, and some simple binary variables characterizing the housing and asset holdings of the household. Indonesia's various targeted public assistance programs use a similar, though somewhat more sophisticated rule.

The advantage of a rule like this is that the necessary data could be collected in half an hour or less; the disadvantage is that it may not always get us where we would like to be. Using data from Indonesia, Nepal, and Pakistan that has information about both consumption and assets, Filmer and Pritchett (2001) show that between 60–65% of those in the bottom 40% of the distribution based on consumption were in the bottom 40% based on asset ownership. In other words, something like 35–40% of the poor might be misclassified but probably less, since there is no reason to assume that the consumption always gets it right.

There is however another concern. Using specific forms of wealth as markers has the advantage of being easy to measure but the disadvantage of being easy to manipulate: if I think that building another room in my house will reduce my chances of a handout from the government I might choose to put my savings into gold. This becomes an even bigger concern when we base the choice on whether your child goes to school. Parents who are already unconvinced of the benefits of education (more on that later) may not hesitate too much before withdrawing their child from school in order to secure their position on the public assistance list.

The implementation challenge

Any method for identifying the poor is of course only as good as the people using it will allow it to be. As we already noted identifying the poor is hard work, even with simplified criteria and it is not clear that those responsible have a strong reason to get it right. Indeed it is not hard to imagine that the person who decides whether you get to be on the public assistance list or not might want to charge something for that favor, and if you are really poor and cannot afford the price, he may prefer to hand your card to someone, less deserving, who can. There is also a natural tendency to be generous in interpreting the rules: why deprive somebody just because he fails to meet the criteria, when there is very little risk that anyone will complain if you do.

Consistent with this, a recent study in India that compares the number of poor people in the country with the number of BPL cards issued concluded that there were 23 million extra BPL cardholders (NCAER 2007, reported in *Times of India* 12/22/07). Another study, conducted by the international NGO Transparency International in partnership with the Center for Media Studies in India focused more directly on mistargeting. They asked a random set of households both questions about their economic status and also whether they have a BPL card (TI-CMS 2007). The study concluded that about 2/3rds of households that were actually BPL had BPL cards, which is not too bad given that the measure of economic status they used was relatively crude and they still out-performed the Filmer-Pritchett study of targeting using wealth data, mentioned above. Of course there are also inclusion errors (the 23 million extra cards) but this could just reflect the fact that it is hard and perhaps pointless to make fine distinctions within a group that is generally poor.

However a more detailed study from Karnataka resists this more benign interpretation. In Atanassova, Bertrand, and Mullainathan (2007), the authors survey 21 households in each of 173 villages in the Raichur district in the state of Karnataka. In each of these households they collect the data used for the BPL classification by the government and based on that data they can construct their own BPL list. They find that while 57% of households in the control villages have a BPL card, only 22% of the households are actually eligible. Moreover, 48% of households are misclassified. The inclusion error, i.e. ineligible households who have a card, is 41% and the exclusion error, i.e. households who are eligible for BPL but don't have it, is close to 7%. This means that about one third of the eligible households don't have a BPL card, while about half of the ineligible households do have a BPL card. More worryingly, when they use income as a proxy for wealth, the poorest among all ineligible households are not the ones who have a BPL card. In particular those who are just above the eligibility cutoff for BPL, i.e. those with annual incomes between Rs. 12,000 and Rs. 20,000, are less likely to be included than those whose incomes are between Rs. 20,000 to Rs. 25,800 and 42% of the wealthiest people (with income above Rs. 38,000) have a BPL card. When they investigate the reasons for the inclusion of ineligible households, the fact of being socially connected to village officials turns out to be a good predictor.

A more participatory approach

The fact that the identification process can get captured by the village elite may be one reason why others have suggested a very different approach: why not make use of the fact that small communities (like villages) can probably identify those among them that are really poor? And while individual villagers might have reason to slant their information in specific ways, this ought to be mitigated if we brought together a large enough group of them.

Bandhan, one of India's largest Micro Finance Institutions, made use of this approach to identify beneficiaries for their Ultra-poor program. Under this program, families that were identified as being too poor to be able to brought under the microcredit umbrella were offered the "gift" of an asset (which could be a cow, a few goats, or a threshing machine) and some short term income assistance (for the period before the asset starts paying off) with the hope that this might permanently rescue them from dire poverty and put them in the mainstream of the village poor population. Following the methodology developed by the Bangladeshi NGO BRAC, which originally came up with this program, for identifying the ultra-poor, Bandhan carried out Participatory Rural Appraisals (PRAs) in the village.[2] In the PRA, a minimum of twelve villagers ideally drawn from various sections of village society sit together and come up with a map of the village where each household is assigned a location. Then they classify the households into six groups, from the poorest to the richest. Following the PRA, Bandhan selects about 30 households from the set of lowest ranked households.

Bandhan's process does not stop here. They then collect asset and other information about these 30 households and eventually 10 are picked to be part of the Ultra-poor program. We were however interested in the effectiveness of the PRA as a way to target the very poor and in some ways the results bear out the validity of this approach (Banerjee, Chattopadhyay, Duflo, and Shapiro 2008). Those who were assigned to the bottom two categories in the PRA have about 0.13 acres less land than the rest of the surveyed population which might not seem much until we consider the fact that the average land holding in this population is actually 0.11 acres. Similarly while 34% of the surveyed villagers report not always getting a square meal, that fraction is another 17 percentage points (i.e. 50%) higher among the poorest two groups in the PRA. Such households are also less likely to have much schooling and more likely to have a child out of school or a disabled family number.

The one place where the PRA does not help is in identifying those who are consumption poor, but then we also found that in these villages possession of a BPL card is uncorrelated with consumption. And unlike the BPL card, the PRA does predict being land scarce and not being able to get two square meals.

[2] Participatory Resource Appraisals (PRAs) are a standard technique for getting a group of villagers to map out their village together.

Villagers therefore do have information that they are able and willing to use in the public interest: in particular their information might make it possible to make distinctions within the population of the poor.

Unfortunately, at least in these villages, the PRA completely missed a quarter of those who showed up in our survey—their names never came up. And since our survey deliberately focused on the poor, it is not because these people were irrelevant to the question at hand. Basically it seems that even in a village of a few hundred people, "out of sight" might be "out of mind." The PRA classifies those it finds relatively well, but what about those it leaves out?

Another concern with the PRA approach is that it might work better as a way to identify the ultra-poor, than as a way to identify the average poor person. Most people probably feel that they are superior to the ultra-poor, and therefore a certain *noblesse oblige* takes over when they are thinking in terms of helping those unfortunate people. When it is the average poor person who is being identified, most villagers probably feel that they are just as deserving as anybody else, which is likely to lead to disagreements and conflict.

Nonetheless the results from this very small pilot were promising enough to encourage us to investigate this issue further. Perhaps one should combine the two approaches: begin by coming up with a list of the potentially poor based on wealth (or other) data and then have the village community edit the list (to reduce the risk of people being forgotten) based on their superior information. One could imagine many other hybrids as well. In some ongoing research, Rema Hanna, Ben Olken, Julia Tobias, and myself from MIT's Abdul Latif Jameel Poverty Action Lab, along with the Indonesian government and Vivi, Alatas and her team from the World Bank in Jakarta, have been designing experiments to rigorously compare the efficacy of the survey and PRA methodologies for identifying the poor, and to study some of these hybrids.

Self-targeting

The alternative to targeting is self-targeting. The idea of self-targeting is of course not new. The notorious Victorian poorhouses, which Scrooge commended and about which the compassionate gentleman in *A Christmas Carol* said "Many can't go there; and many would rather die," were exactly that: a place so miserable that only those who are so desperately poor that they had no recourse would want to go there. India's recently introduced National Rural Employment Guarantee Scheme (NREGS), under which every rural household is entitled to 100 days of unskilled public employment at the minimum wage on demand (i.e. within 15 days of asking for employment) in their village is probably the biggest single effort in this direction.

The theory behind such schemes is well-known: it does not need to be targeted, because only those who have no better alternatives would want the kind of work (digging ditches, carrying bricks) that it offers. The fact that it is work on demand also means that you don't need anyone's sanction to seek work. It also has the advantage of flexibility: a lot of extreme poverty is temporary and/or unpredictable. For example, when the income earner in your family is unexpectedly taken ill, it might take a long time to get your family reclassified as BPL, but the right to work is designed to be always there for the asking.

The disadvantages are also clear: what happens if there is no one in your family who is fit enough to do manual labor? Moreover, labor is a social resource: making people dig ditches in order to prove they are poor, is of course wasteful unless you want the ditch dug. If you never wanted the ditch and had some way of knowing who the poor were, you could have given them the money and let them do something productive with their time. A significant part of the original NREGS documents was therefore devoted to spelling out what the village needs to do to make sure that the labor is used to create useful (public) assets for the village.

Corruption is also a challenge. This is of course always an issue, but the fact that the NREGS is supposed to be driven and therefore there is no fixed budget, must make it particularly tempting to throw in a few extra names. This is the problem of fake muster rolls (a muster roll is where NREGS transactions are recorded) that critics of the program have talked about. For this reason, the program requires that all muster rolls be displayed in public and supporters of the program put a lot of emphasis on what are called social audits. During these audits, concerned volunteers try to find the people named in the muster rolls and ask them if they received the claim payments.

These audits do reveal a fair amount of corruption in the implementation of the NREGS. In the state of Jharkhand a social audit of five randomly chosen villages carried out by researchers from Allahabad University found that about one third of the money was lost (Dreze, Khera, and Siddhartha 2008). More frighteningly, one of the activists involved in a social audit somewhere in Jharkhand was murdered, and the presumption is it had something to do with what the audit had unearthed. On the other hand, in Chattisgarh an audit of nine randomly chosen projects suggest that about 95% of the claimed wage payments were actually made.

While 5% seems good and one third less so, it is not clear what the benchmark ought to be. This is also the problem with the other criticism one hears; that the program is not doing enough. The Comptroller and Accounts General of India, a government organization charged with the oversight of public programs, reported that 3.2% of those who had registered themselves for the program had actually worked for the full allowed 100 days and that, on average, a registered family got less 20 days of employment. In response the Ministry of Rural Development, which runs the program, pointed out that among the families that actually participated in the program (i.e. those who actually worked) the average number of days of employment was closer to 40 and 10% worked for all 100 days.

But how does one tell whether 40 days (or 10%) is too many or too few? If no one actually ends up taking these jobs, but the presence of NREGA employment at minimum wages pushes up earnings in the private sector and everyone still continues to work there, we would presumably call the program a success. We would also think it a success if almost no one takes the jobs, but the assurance that a job would be available if need be makes the populace less worried and/or more willing to take profitable risks. By contrast, if everyone wants an NREGA job, but only 50% get employment for 100 days a year, we would presumably be quite disappointed. The CAG report mentioned above, suggests that there is at least some unmet demand, and blames the fact that the program is understaffed, but we do not know how much.

In the survey mentioned above that we carried out in the West, we also found that at least in the villages that were part of our study the possession of a job card (which is what you get by registering for the program) does not predict being poor. Does that mean this program is seriously off-target, or is it that everyone wants to get a job card in order to be safe, but they actually plan to use it only if they run out of alternatives?

Most importantly, even if the targeting is reasonably good and the leakages are no worse than in other programs, how do we know that it was worth the hoops that people had to jump though in order to get the money? In other words, unless we are reasonably confident that the assets built by using program labor were worth the time and effort that went into them, how can we be sure that it made sense to go through all that to get better targeting?

Most of this could have been answered if the program had been subject to a rigorous evaluation (combined with a detailed survey of the various groups that end up not participating in the NRGS), but the current decision to extend it to the whole country means that there will not be such evaluation in India.[3] The question of whether self-targeting is worth the trouble remains an open question.

The performance of targeted programs

The government of India's largest targeted program is the Targeted Public Distribution Scheme under which BPL households are allowed to buy subsidized food-grains and other eatables from what is called a fair price shop in the village, which in turn gets supplies from the nearby government warehouse. This is the program that the government's own Finance Minister recently described in the following terms: "About 58 per cent of subsidized grains do not reach the target group, of which a little over 36 per cent is siphoned off the supply chain. I ask you, respectfully, do not the poor of India deserve a better PDS? How can we sit back and watch helplessly the poor being robbed of their meager entitlements?"

What is striking about the numbers he quotes (from the government's own Programme Evaluation Organization's recent report) is that the biggest source of leakage is not the mistargeting of BPL cards, discussed above; it is the direct theft of the grains along the way. Of this 36%, 20% is "lost" in transit, while the other 16% is distributed against "ghost" BPL cards (i.e. cards issued to people who don't exist).

The report also gives a measure of what it calls "exclusion error." According to its numbers only 57% of the BPL households are served by the TPDS. In other words, one cannot even argue that the massive leakages are the cost of reaching all the poor people.

While, as discussed above, targeting is problematic, it is hard to imagine that the government could not do more to prevent the theft if there was the political will. Indeed we do see that in at least two Indian states, Tamil Nadu and West Bengal, theft is less than 20%.

But if lack of political will is a big part of the problem and targeting is as inefficient as it seems to be, there may be a case for giving up on targeting. This would both eliminate exclusion error and bring the non-poor, with their greater influence on the political system, into the ambit of the program.

Helping them to help themselves

In the conventional view the government does this primarily by helping the children of the poor grow up with the health and education that would enable them

[3] There is still value in trying to evaluate the impact of varying some of the program details before the whole thing gets etched in stone. Would there be more jobs created if, for example, instead of assuming that all program beneficiaries have to work for the community, some of them were sent off to work for private business, even though the government continues to back-stop their wage and make sure that no one earnes less than the minimum wage? Would there be more energy in the program if the local elites start to feel that they too have something to gain from it? Or would it somehow encourage the elite to try to "capture" the program?

to be full participants in the economy. It might also provide healthcare for adults as a way to insure them against things that are largely out of their control.

Nutrition

India has, by a wide margin, the largest number of wasted and stunted children in the world. According to the recent National Family Health Survey (NFHS-3), 48% of children are stunted and 43% are wasted, which means that India, a much richer country, has roughly twice the stunting and wasting rates as sub-Saharan Africa.

However, while malnutrition is clearly a huge problem in India, it is not clear to what extent it is a matter of access to food rather than nutritional practices. The shocking levels of stunting and wasting rates we report above turn out to correspond to the average for the middle category among the five wealth categories reported in the NFHS. It is hard to imagine that this group cannot afford the levels of nutrition that children must be getting in an average family in an average country in sub-Saharan Africa.

Moreover, it is not obvious that the TPDS, as currently designed, does very much to fix problems of malnutrition. In part it is just an income transfer and most of the evidence suggests that extra money does not turn into very much extra nutrition (Strauss and Thomas 1998). The fact the extra income comes in the form of extra food might help, but only if the 20kg. of grain that a family gets from the TPDS is more than what it would have bought in any case, which, from all accounts, seems implausible.

Given this and the rather disastrous performance of the TPDS, it might make sense to entirely rethink the idea of providing food subsidies to people. Why not give people money rather than food and thereby avoid all the problems that come from the fair price shops? It is true that the price of food varies, but the amount of money could be tied to the consumer price index and in any case there is a strong suspicion that, under the current system, when the market price goes up relative to the TPDS price, leakages increase and the poor end up with less.

There is of course still the challenge of how to make sure the cash actually reaches those who it's meant for, but this is where information technology can help us. South Africa pioneered the technology of using ATM machines that can recognize fingerprints to deliver pensions and something similar might very well work in India. Certainly it seems worth an experiment or two.

However it is not clear that a cash transfer program, however well implemented, will do much for the problem of malnutrition. As pointed by recently by Deaton and Dreze (2008), the substantial increase in the incomes of the poor between 1983 and 2004 did not lead to a sharp increase in calorie or protein consumption, even in the group that lives on a low 1,600 calories a day. Both calorie and protein consumption went down for all the other (less-poor) groups.

This raises the concern that the poor may be under-investing in nutrition, either because they do not recognize its value or because they do not want to be left out entirely from the consumer paradise that middle-class India is becoming.[4] In either case, it suggests that informing and influencing people's consumption choices may be an important part of nutrition policy. This is reinforced by other evidence. For example, exclusive breastfeeding till the age of six months is one simple and widely recommended way to fight malnutrition and many childhood diseases. According to the NFHS, the average duration of exclusive breast-feeding is only two months. It is also recommended that breastfeeding be started right after childbirth, so that the child does not miss out on the colostrum, which contains many valuable nutrients. Only a quarter of the mothers in NFHS say that they started breastfeeding within an hour of child birth.

The challenge here is to change behavior, including behaviors that may be deeply embedded in tradition. The Government of India's current idea is that this will be the responsibility of a ASHA *Sahayogini*, a local woman with some schooling who will given 23 days of training and a stipend of about $25 a month. It is not entirely clear that the kind of people who will take this job will have the energy, the knowhow, or the charisma to the point of being able to persuade other women to change age-old practices. A credible evaluation of the impact of this program is however not on the horizon as far as we know.

Education

The poor performance of the Indian primary education sector has been in the news in recent years thanks to the Annual Survey of Education Reports brought out by the prominent educational NGO Pratham. The basic finding from these reports is well-known: 42% of fifth graders in India cannot read at second grade level, and 68% cannot do subtractions involving two digit numbers.

Yet while there are examples of schools where more than hundred children crowd into a single classroom, the Indian education sector is not underfunded by the standards of comparable countries. India spent 3.7% of its GDP on education in 2005, which is somewhat

[4] These are not the only possible explanations—Deaton and Dreze (2008) suggest that it may reflect that there is less demand for hard physical labor.

below the average for lower middle income countries (4.3%) but higher than the average for East Asia and the Pacific (2.9%) (World Bank 2007). According to a recent paper by Murgai and Pritchett (2007), government teachers in India are a paid more than other people with similar qualifications. Student to teacher ratios are high but below 40, the cut off that is used in Israel (a much richer country) for the maximum acceptable class size.

The problem, at least in part, seems to lie in the quality of teaching. According to the World Absenteeism Survey (Chaudhury et al. 2003), which sent surveyors to schools at random times to measure teacher presence, 25% of teachers are missing on any given day. Moreover, conditional on being present, they spend only 45% of their supposed teaching time in the classroom.[5]

However what is less emphasized, but equally striking, are child absence rates, which are comparable or higher than teacher absence rates. Given this, one might wonder if the teachers are not simply responding to the general climate of indifference they find among their pupils. Perhaps, at the margin, a few more days of attendance by teachers will not do much for child performance. A recent randomized experiment reported in Duflo, Hanna, and Ryan (2007) tests this hypothesis. Seva Mandir, a prominent NGO in the state of Rajasthan, was facing teacher absence rates of 40% in the single-teacher schools it ran in some of the more remote corners of the state. Under encouragement from Duflo they started monitoring the teacher's presence using a camera and paid the teacher based on the number of days present. It was introduced in a random set of schools, so that the impact could be evaluated.

Many people in the Seva Mandir community felt that while this might make teachers come to school more, it will not affect learning. In fact it raised test scores by a not inconsiderable 0.17 standard deviations, proving that if teachers put in more effort children do benefit.

The fact that better incentives for teachers can lead to better student results was also the conclusion of Muralidharan and Sundararaman (2006). They studied an experiment in Andhra Pradesh where government school teachers were promised a reward based on the improvement in the performance of their students and found a significant impact on test scores, including fields where the results did not count towards the incentive.

However the affect of incentives was, once again, not huge—0.15 standard deviation. Clearly a lot more would be needed to transform India's faltering education sector. How are we to generate the incentives needed for that to happen?

One answer, which was at the heart of Indian government's last major attempt to reform primary education—the *Sarva Shiksha Aviyan* (or SSA)—is that the community has to play a much more active role in demanding education from the system. However in a survey of 280 villages in Jaunpur district in the Indian state of UP, revealed that at least four years after the launching of the SSA 92% of parents did not seem to know about Village Education Committees, which is the main channel through which they can get involved in improving the local school and access SSA funds, while only 2% could name its members (Banerjee et al. 2006). At that time we had speculated that this could be because no one had taken the trouble to inform them about the VEC or the SSA. We therefore carried out a field experiment in the district aimed at informing and mobilizing parents around the state of education in their village and the possibilities opened up by SSA.[6] In this experiment volunteers from Pratham spent a day and a half in each village, holding numerous small and large meetings where they informed parents about their rights, including the right to complain about teachers who do not attend and the right to hire extra teaching assistants (*shikshakarmis*) for their overcrowded schools. They also told them about the (poor) performance of the children in the village and taught them how to test their child's reading skills.

None of this had any effect on any parent outcome except that a statistically significant but minuscule 2.6% more parents now knew about the VEC: there was no increase in the number of complaints, no additional visits to the school, no extra effort put into hiring teaching assistants. And not surprisingly, given that, it had absolutely no effect on test scores.

What we cannot yet tell is whether this indifference stems from a belief that the educational route to prosperity does not work for people like them (India, after all, has a long history of believing education is only for certain elite castes). Or is it that they believe that teachers are beyond their reach, politically and socially, and therefore are convinced that trying to make them change their ways is not really an option for people like them. However there is some recent evidence suggesting that it might be a bit of both: Jensen (2007) carried out an experiment in the Dominican Republic where he told poor parents about the returns of education, and found that their children do work harder in school when this happens. On the other hand, the only successful intervention—our UP study—was the one where trainers from Pratham

[5] This put India among the worst performers in the survey (only Uganda at 27% teacher absence rates did worse).

[6] For the experiment see Banerjee, Banerji, Duflo, Glennerster, and Khemani (2008).

trained village volunteers how to teach. One or more classes were started in every treatment village, children came and test scores increased substantially. The success of this experiment and the failure of the other interventions (the ones requiring some degree of social action) suggest that parents do care about education but shy away from confronting the teacher.

In either case it is hard to be confident that parental activism is going to solve the lack of incentives, at least in the near future. The alternative is to rely on market solutions, i.e. some kind of a program where parents are given publicly financed vouchers that will pay private school fees for their children. The usual arguments against vouchers seem relatively uncompelling in the Indian context: will it generate more segregation and inequality in the kinds of education children get? Perhaps, but given that the rural elite has already exited the public system in many areas, it is at least as plausible that it would reduce inequality, at least as long as the vouchers are set up in such a way that they cannot be used to subsidize sending children to really elite schools. Should one be concerned about parents colluding with school management to cash in their vouchers instead of sending their children to school? Unlikely, we think, now that parents care enough about education to reach nearly 100% school participation rates.

Moreover the veritable explosion of private schooling among relatively poor families in rural India in the last few years means that in many villages there are multiple private schools competing for students. According to ASER (2007) 19.3% of all children age 6–14 in rural India go to private school. Muralidharan (2006) reports on a nationally representative survey of rural private primary schools in India and observes that 50% of the schools in the 2003 survey were founded in the previous five years.

Muralidharan also observes that these schools are cheap (the median monthly fee is less than $2 at current exchange rates) despite the fact that the teachers in these schools are more likely to have college degrees and that they have a student-teacher ratio that is slightly more than half that found in public schools. This is because private school teachers are paid between 10–20% of what public teachers are paid. Andrabi, Khwaja, and Das (2003) who studied a very similar phenomenon in the province of Punjab in Pakistan, argue that the gap in performance between private and public schools is too large to be explained by plausible selection arguments: i.e. private schools are simply cheaper and better.

However we clearly need much more compelling evidence before such a radical shift would be warranted. Karthik Muralidharan and Michael Kremer are currently carrying out a randomized evaluation of school vouchers in the state of Andhra Pradesh; hopefully a number of other upcoming voucher programs in other states will also be evaluated. The challenge for all these evaluations is how to deal with the fact that supply of private schools will need to adjust to the expansion of demand that will happen when vouchers are universalized, but does not happen under experimental conditions. The fear is that fees will go up sharply as schools compete for teachers. In order to be able to answer this question, Muralidharan and Kremer, randomize both across and within villages. If the village is the relevant market for teachers, then the village level experiment will tell us about the impact on school fees and the supply of schools. If however teachers are willing to change villages in order to find work, as seems likely, this will not give us the complete answer and further research would be needed.[7] In the meanwhile, the education sector is clearly drifting.

Healthcare

Any problem that the education sector has, the healthcare sector shares in abundance. The 40% absentee rates for the Auxiliary Nurse Midwives (ANMs), the lowest level health practitioner in India's multi-tiered healthcare system, are substantially higher than that of teachers (Chaudhury et al. 2003). When a number of health sub-centers (where these nurses are based) were randomly chosen for a program of incentives based on attendance, the nurses and their immediate bosses colluded to completely undermine the incentives: nurse attendance after the experiment was actually lower than before (Banerjee, Duflo, and Glennerster 2008).

Even more worrying, though perhaps unsurprising given the absentee rates, is the fact that even very poor people have mostly stopped making use of these nurses. In the rural Udaipur district, where per capita daily expenditure for the average person is no more than a dollar a day, Banerjee, Deaton, and Duflo (2004) found that less than a quarter of visits to healthcare providers were to government facilities. Nearly 60% of all visits were to private practitioners and the rest were to traditional healers. This is despite the fact that private "doctors" are further away, more expensive, and less likely to have any medical qualifications.

When we asked potential patients why this is so, they cited quality of treatment. We know that this quality is often poor. We already talked about the very high rates of absenteeism. Das and Hammer (2007),

[7] The answer may also depend on whether the scaled up voucher program will be accompanied by shutting down a large number of public schools, in which case a lot of teachers would be available to be hired.

based on a survey of public and private doctors in urban Delhi, make the point that public doctors who deal with poorer patients more often than not prescribe medicine without ever touching the patient. But a part of what patients call quality is also what the government providers complain about—they say that private providers overuse injections, in particular injected antibiotics and steroids, and this is seen by the populace as good treatment. Our data does provide some support for this view. A remarkable 60% of all visits to private practitioners involve an injection being given, though we do not have the data to say whether these are actually dangerous for the patients. The consensus view among experts is that there is substantial over-medication.

A related concern is that the movement towards private healthcare means that people are no longer talking to people whose job it is to educate them in public health practices (rather than sell them a treatment). For example, in rural Udaipur district less than 5% of children are fully immunized according our data (Banerjee et al. 2008) and it is not clear that any of the private health providers are going to do anything about it.

More generally, what makes healthcare for the poor particularly hard is that the market solutions are not necessarily particularly attractive, precisely because of the tendency to underestimate the cheap but valuable preventive aspects of medicine, relative to expensive and potentially harmful curative procedures. Subsidized health insurance is the equivalent of vouchers for the case of healthcare, and there are a number of on-going experiments in India including one that we are evaluating. However almost all of these insurance policies only pay for inpatient services for the simple reason that they are much easier to verify. This means that check-ups, tests, and all other forms of preventive medicine are expenses carried by the individual and that the insurance system actually discourages.

At this point there are some doubts about whether even this very simple product can be made economically viable. A product that covers more outpatient services is likely to be much more costly because the utilization of these services is far harder to monitor, and the government may have to get involved. One advantage of a subsidized program is that it could be used as hook to get people more involved in early detection and prevention, as in order to get the insurance at a subsidized rate the individual would need to meet certain requirements.

That such incentives can work well is demonstrated by a recent experimental study where women were offered a kilo of lentils whenever they got their children immunized. This more than doubled the number of children who are fully immunized (Banerjee et al. 2008).

In some ways government policy in India is moving in this direction. There is now a scheme that gives financial incentives for women who give birth in hospital and as part of the scheme the woman is required to make a certain number of antenatal and postnatal visits to the clinic. While enforcement of these new rules seems relatively lax at this point, it has the potential to make a substantial contribution.

The way forward

The current trend in anti-poverty policy is a rather different approach to the idea that the poor need to take charge of their lives. Instead of thinking of the poor as workers who need to have the requisite skills, it thinks of them as potential entrepreneurs who need capital and property rights and the protection of the law: hence the emphasis, for example, on microcredit. It is not that investment in human capital is unimportant; it is more that the sponsors of this view are skeptical of the government's ability to deliver human capital and would rather see the poor earn extra income to pay for the human capital they want for their children.

The fact that it is not easy to get the government to deliver is, of course, entirely consistent with the argument we are making. The question is whether we can be confident that where the government will not deliver, the poor will; that they can and will pull themselves up by their bootstraps, with a little help from micro-credit organizations.

As we have argued elsewhere at length (Banerjee and Duflo 2007, 2008) there is no empirical warrant for this view. The basic point is that the poor neither have the skills, nor the knowledge of markets, nor the understanding of technology, to compete effectively in the marketplace. They are also, even after they get their microcredit loans, limited by their capital to the most primitive technologies and most crowded occupations, and hugely vulnerable to all the risks of entrepreneurship. The businesses that they actually run bear the imprint of all these constraints. They are tiny (the median firm owned by the poor has no employee) and heavily concentrated in the few industries that can be entered with minimal specialized skills.

What is more, the poor themselves do not expect their businesses to transform their lives. If they did they would put more effort into growing these businesses.

We argue that many could easily expand their businesses, make more money, and thereby slowly climb the ladder out of poverty; but they choose not to.

None of this is to deny that the poor are not resourceful or energetic; it is just that the playing field is so slanted entreat the point of entry that only those few with enormous resolve and/or talent can make it past the starting line. Nor is it to question that microcredit has probably made the lives of the poor more bearable and therefore deserves our support.

But in the end, the government must remain at the center of anti-poverty policy, because without some help and resources from the outside the poor face an utterly unfair challenge. It does not need to do all the things it currently does (badly) and it should certainly focus more on paying for things rather than making them. Income support and strategically targeted subsidies to key delivery agents (NGOs, Microfinance Institutions, private firms) can go a long way in making the lives of the poor better, without involving the government in delivery. But we should not forget that a very important part of what the government does are things that the market will not—behavior change, preventive healthcare, education for those who live in areas where there are no private schools, emergency relief, etc. Even in these cases, the government can work with implementing partners outside the government, as the example of BRAC in Bangladesh has shown, but realistically, the government will continue to be a major delivery agent in the economy. The challenge for those of us who are in what might be called the ideas sector is therefore to think of ways of redesigning what the government does to make it work better, both in terms of choosing between what it does and what it delegates, and in improving its effectiveness in what it must do.

Bibliography

Andrabi, Tahir, A. Khwaja and J. Das. "Learning and Achievement in Pakistani Education." Education Sector Feasibility *Report*. World Bank, 2003.

Annual Status of Education Report (ASER). "Rural." Pratham Resource Center, 2007. <http://www.pratham.org/aser07/ASER07.pdf>

Atanassova, Antonia, Marriane Bertrand and Sendhil Mullainathan. "Misclassification in targeted Programs: A Study of the Targeted Public Distribution System in Karnataka, India." Unpublished Harvard University mimeo.

Banerjee, Abhijit. *Making Aid Work*. Cambridge, Massachusetts: MIT Press, 2007.

Banerjee, Abhijit and Esther Duflo. "The Economic Lives of the Poor." *Journal of Economic Perspectives*, vol. 21(1), 2007, 141–167.

—, "What is Middle Class about the Middle Classes Around the World?" *Journal of Economic Perspectives*, vol. 22(2), 2008, 3–28.

Banerjee, Abhijit, Angus Deaton, and Esther Duflo "Health Care Delivery in Rural Rajasthan." *Economic and Political Weekly*, vol. 39(9), 2004, 944–949.

Banerjee, Abhijit, Esther Duflo, and Rachel Glennerster. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." European Economic Association, vol. 6(2–3), 2008, 487–500.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Dhruva Kothri. "Improving Immunization Cover in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives." mimeo, Massachusetts Institute of Technology, 2008.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther DUFLO, and Jeremy SHAPIRO. "Targeting Efficiency: How well can we identify the poor?" Unpublished Massachusetts Institute of Technology mimeo.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. "Can Information Campaigns Spark Local Participation and Improve Outcomes? A Study of Primary Education in Uttar Pradesh, India." *World Bank Policy Research Working Paper*, no. 3967, 2006.

—, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *World Bank Policy Research Working Paper*, no. 4584, Impact Evaluation Series no. 21, 2008.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Kartik Muralidharan, and Halsey Rogers. "Teachers and Health care providers Absenteeism: A multi-country study," mimeo, Development Research Group. World Bank, 2003.

Deaton, Angus and Jean Dreze. "Nutrition in India: facts and interpretations." Unpublished Princeton mimeo, 2008.

Dreze, Jean, Reetika Khera, and Siddhartha. "Corruption in NREGA: Myths and Reality." *The Hindu*, Jan 22, 2008 <http://www.hindu.com/2008/01/22/stories/2008012254901000.htm>

Duflo, Esther, Rema Hanna, and Stephan Ryan. "Monitoring Works: Getting Teachers to Come to Work." *Poverty Action Lab paper*, 2007.

Filmer, Deon, and Lant Pritchett. "Estimating Wealth Effects without Expenditure Data or Tears: An Application to Enrollments in States of India." *Demography*, vol. 38.1, 2001, 115–132.

Jensen, Robert. "The Perceived Return to Education and the Demand for Schooling." Unpublished Brown University, mimeo, 2007.

Leonard, Kenneth, Jeffrey Hammer and Jishnu Das. "The quality of medical advice in low-income countries." *World Bank Policy Research Working Paper*. Series 4501, 2008.

Muralidharan, Karthik. "Public-Private Partnerships for Quality Education in India." Unpublished draft, 2006.

Muralidharan, Karthik, and Venkatesh Sundararaman. "Teacher Incentives in Developing Countries: Experimental Evidence from India." *Job Market Paper*. Harvard University, 2006.

Pritchett, Lant, and Rinku Murgai. "Teacher Compensation: Can Decentralization to Local Bodies Take India From Perfect Storm Through Troubled Waters to Clear Sailing?" India Policy Forum, 2006/07.

Strauss, John, and Duncan Thomas. "Health, Nutrition and Economic Development." *Journal of Economic Literature*, vol. 36, no. 2, 1998, 766–817.

World Bank. "World Development Indicators, 2007: Table 2.9, Education Inputs." 2007. <http://siteresources.worldbank.org/DATASTATISTICS/Resources/table2_9.pdf>.

the artistic discovery of the future

FRANCISCO CALVO SERRALLER

Modern art emerged from a revolution about two-and-a-half centuries ago, and it is very difficult to prognosticate about its possible future when we are still trying to discover what, exactly, it is. What was traditionally understood to be art until the second half of the eighteenth century reflected the affirmations of the Greeks who invented it. Around the sixth century B.C., they defined art as a selective imitation of reality or nature—not a simple copy or indiscriminate replica of what they saw around them, but rather one that transcended appearances and captured the underlying order of things. The depiction of this order produced beauty, making it clear that the fundament and goal of art was, in fact, beauty. Not beauty as it could be interpreted subjectively, but rather an objective codification of beauty, with a mathematical nature. In that way, the aesthetic qualities that could be demanded of a classic art work always answered to mathematical concepts such as "harmony," "proportions," "symmetry," "rhythm," and so on. All of these were guidelines for order, but this orderly formal selection also corresponded to the content or subject being represented, because artists could not choose their subjects freely. They had to be socially edifying, that is, they were limited to the instructive actions of gods and heroes from a remote and mythical past. In sum, the Greeks defined art as, so to speak, a good representation of good, in which formal order and moral order were combined. Now this canon held sway at the heart of Western civilization for approximately 24 centuries, creating a historical tradition united by the term "classicism." Clearly, such a long tradition suffered successive crises, whose gravity sometimes opened broad parentheses, including what is very significantly referred to as the "Middle Ages." At such times, even though they did not completely disappear, classical artistic principles became highly corrupted. But for that very reason, at the beginning of the modern era, the restorative formula of the Renaissance was invoked, bearing with it the values and criteria of classical Greco-Roman Antiquity. Beginning in the fifteenth century, this restoration of the classical ideal of art, which at first was accepted with dogmatic enthusiasm, suffered innumerable polemical attacks as a result of the profound changes of all kinds brought on by the modern world. Still, art's doctrinal essence managed to survive right up to the dawn of our revolutionary epoch, which radically changed what had been thought of as art until then.

Perhaps this succinct, synthetic and reductive version of the artistic history of Western classicism will help us to better understand the subsequent revolution. The first hint of this artistic revolution was the insinuation that beauty not only could not be the exclusive fundament of art, but that it was even an unbearable barrier to art's potential development. Art demanded a much broader horizon. It was then, over the course of the eighteenth century, that art sought ideas that were different and opposed to the limitations imposed by beauty, such as the "sublime" or the "picturesque," which postulated that "disorder" and "irregularity" could offer humanity new aesthetic experiences that would be more intense and revelatory than traditional ones.

In the first paragraph of chapter three of *Laocoon: An Essay on the Limits of Painting and Poetry,* published in Berlin in 1766, G. Ephrain Lessing wrote: "But, as we have observed, in recent times, art has acquired incomparably broader domains. The field in which it exercises its imitation has spread to all of visible Nature, of which beauty is but a small part. Truth and expression, it is said, are the supreme expression of art; and just as Nature is continually sacrificing beauty to attain higher interests, so too, the artist must subordinate that beauty to his general plan, without looking beyond what truth and expression permit. In a word: truth and expression transform natural ugliness into artistic beauty." Among many other contemporary testimonies, the clear and forceful explanation offered by Lessing's text ineluctably confronts us with the problem of our own time's revolutionary art, which will not be an art based on beauty. Moreover, it can be affirmed that the aesthetic debates and the development of art itself in the second half of the eighteenth century were ruled by a common desire to wage a true war of liberation on beauty, whose classical canon was assaulted with the same fervor that characterized the taking of the Bastille.

Now then, on what could art be based, once it had been stripped of its traditional foundations? Neither the pleasure of the limitless sublime nor the irregularity of the picturesque are foundations in and of themselves. They are, instead, challenges to the traditional concept of beauty. At any rate, it is clear that they change our perspective, a little like Copernicus's effect on classical physics, because they rebuild, or shift, the horizon, which is no longer focused on what is well known and familiar, but instead on what is beyond the limits and unknown. But how can artists grasp that unknown aesthetic Cosmos without imposing constrictive limits like those of beauty? On the other hand, it is important to keep in mind that this rebellion against the hegemonic domination of classic beauty never signified that supporters of the artistic revolution denied the value of traditional art, nor that they sought to keep anyone wishing to continue submitting to its limits from doing so. What they were combating was the universal obligation to submit, that is, the negation of the artistic value of those who sought to explore new paths. In sum: they realized that the artistic horizon could be incommensurably vaster than classicism, but they had still not found precise limits to define its new dimension, so it is logical that the question would be settled in the most open and negative way possible: by seeking the foundations of art in the foundationless, that is, freedom. Schiller was already expounding this idea during the last and crucial decade of the eighteenth century when, in *On the Aesthetic Education of Man* he affirmed that it was characteristic of the aesthetic state to "give liberty by means of liberty." In other words, that the foundations of the new revolutionary art would be, in effect, freedom, whose only circumstantial limitation is not aesthetic, but "technical."

If the only foundation for art is that, through it, one can reach freedom, which necessarily negates any previous limitation, then how can the new map of art be drawn? The only way would be the obvious one that defines its territory as that which successive explorers discover, or to put it another way, that "art is whatever we call art," because that is what the artist says, and we accept it. Beyond these hurried generalizations, anyone who observes the material evolution of art from the second half of the eighteenth century to the present will realize that it is not that art has continually changed form or style, but rather that, each time it does so, it entirely redefines what art is. Significantly, if we analyze the reactions of art's public to each new change in the art of our time, we will see that most of the time, it doesn't deny the value of its form and content; it denies that it is art. Nor does the opposite attitude change anything, because when, as now, the public accepts all that is new, simply because it is new, then inevitably and symmetrically, it is recognizing that, in fact, anything can be art. In other words, that if nothing or everything is art, then what is constantly being questioned is art itself.

It is true to a certain degree that new art did not produce such a level of perplexity when it still used traditional supports and techniques, but we cannot make that the *quid* of the question, as has sometimes been done. I mean to say that neither the postulation of an industrially manufactured urinal signed by an artist, nor the development of new techniques like photography or cinema—regardless of whether they were rejected or accepted—resolves our perplexity, which stems from having to constantly discover what art is. This uncertainty has led to a social polarization between those who disqualify, and those who absolutely defend contemporary art. The former cling to the traditional concept of art as its only genuine manifestation, looking only to the past, as it were. The latter symmetrically postulate that only innovative art can truly respond to the current demands of humanity, which places all its expectations on the future. Now, is it indisputable that an irreparable breach has opened up between traditional art and the art of our time? Is the art of the past the only authentic art, and contemporary art mere fraud, or vice versa? Anxiety leads men to make sweeping statements whose crystal-clear theories grow cloudier when they come into contact with real life. Of course it is always easier to explain the past than the present, yet the fact that contemporary art does not answer to the classical canon does not mean that its practice is arbitrary. Instead, it can signify that its practice is not regulated in an objective way. Based on anti-canonical and elastic liberty, the art of our time is necessarily concerned with exploration, which will not stop until art finds a limit. That limit, in turn, will make it possible to determine the new frontiers, but it is clear that such a moment has not arrived. At any rate, two and a half centuries of contemporary art constitute a history that can be understood and ordered, indicating that uncertainty is always located in the present, but not in the past, which is also a modern past. In other words, there is now a solid avant-garde tradition. Successive new art styles have passed through a critical filter and, just as in the classical period, that filter has left out much of what has been done, selecting only the best—what remains relevant as time passes. So not knowing where art is going has not been enough to crush the capacity to discriminate, partially because leaving the beaten path does not automatically mean getting lost. Otherwise, there would never have been any new discoveries. We can close this matter by observing that there is also irrefutable historical proof: none of the most revolutionary modern artists, from Goya to Picasso, innovated without taking the past into account. And while the manner of doing so continues to change as well, it continues to be the case today, as can be seen in innumerable examples.

Moreover, isn't the same thing happening in all fields of contemporary knowledge and experience? It is certainly true that the currently hegemonic scientific knowledge, and even more so, its technological offspring, are ruled by the dynamics of progress, in which each new step invalidates the previous ones, but it is not clear that this ideology could be applied in the field of art, none of whose changes—even the revolutionary ones—has implied the invalidation of what came before. In reality, appreciating Duchamp in no way detracts from Phidias or Raphael, and their value should in no way disqualify those who do not continue to repeat their formulas. Thus, art has continued to change constantly, but that accumulation of innovations has not lessened the value of antiquities. So shouldn't we consider this idea that the art of today and that of yesterday are radically opposed to be a fallacy constantly debunked by events? And the fact that we have no guarantee that the same will happen in the unknown and unknowable future certainly does not authorize us to adopt an apocalyptic perspective. Were such an apocalypse to occur, it would certainly be much more serious in any other dimension than the artistic one, whose scope is comparatively innocuous. In general, art is actually the practice with the greatest propensity to build bridges with the past, not only because it does not answer to a monosemous mode of knowledge, but also because it is a constant reflection on, and investigation of experience, of what has been lived and of what has passed. And its main *raison d'être* is to become lasting, memorable, immortal.

But what mark has the art of our time made on history over its two-and-a-half century existence? It is hard to discern if one does not first understand the modernizing drive that characterizes it. Etymologically, the term "modern" comes from Latin and means "what is current," so modernizing art means bringing it up to date, that is, making it respond to the present as we see it, with its modes or fashions, novelties and innovations. Modernizing classical or classicist art meant, in the first place, modernizing its contents, its concept of history. For history was its subject matter until then, not the present, and certainly not the present taken in an indiscriminate manner. According to the aesthetic hierarchy established by the Greeks, the most elevated and profound aspect of a story was its tragic mode, the renewed celebration of an ancestral, founding sacrifice whose recreation would have a traumatic—cathartic—and instructive effect on the viewer. On the other hand, the least elevated was the comic mode, whose characteristic was to offer contemporary stories of mortal men. Its instructive effect was much less intense and could even be called "relaxing." If this scheme is applied to the visual arts, it will be understood that the superior genre was that of tragic narration, while the other, subordinate, and inferior ones represented comic episodes that sometimes bordered on the insignificant. With this in mind, modernizing the artistic representation of what is historical meant not only giving preference to the lesser stories of mortals—who themselves lacked social importance or deep symbolic significance—but sometimes even completely eliminating their direct presence, so that a landscape or any object by itself could attain the same dignity as human actions. In that sense, the modernization of art's content was similar to its "democratization," making the subject matter accessible to anyone, and its message immediate and relevant, like an on-the-spot codification of what simply occurs. Anyone who observes this modernizing process will notice how it progressively became an acceptance not only of common stories, but even of the most commonplace and banal of them all, until, around the middle of the nineteenth century, it blazed the way to insignificance.

Once avant-garde art reached the point of telling everything, the next step was almost forcibly not telling anything, that is, "denarrativizing" or "deliteraturizing" art, whose true identity was now considered to be its form, that is, all the materials that make up an art work besides what it symbolizes or represents. This second modernizing gesture, which concludes around the time of the World War I, was what concentrated its revolutionary powers on what had been considered formal elements until then. Elements that constituted what art was, in and of itself. The goal was first to eliminate the inherited manner of painting in perspective, and the second was, so to speak, dealing with the resultant "nothing." Because, what was left of painting after removing its traditional content and form? The surgical instrument used for this operation was Cubism, which led to non-representational or abstract art. Moreover, it almost immediately led to a complete reconsideration of what art is when it no longer needs a painting or a sculpture, but can instead use anything in any medium, as was somehow implied by the heteroclite and nihilistic movement called Dada. If we recall that, in that stormy second decade of the twentieth century, besides collages and readymades, photography and cinema also became overwhelmingly present and meaningful, we will understand that by then traditional art could be considered completely liquidated in both form and content. And if the modernization of content offered the possibility that new art could be free of any thematic restrictions, and if the modernization of form allowed anything to be art—and that has been the case ever since—then what limits could be assigned to the resulting artistic horizons, that "musical nothing," as Mallarmé liked to call it?

Eighty years have passed since the avant-garde eliminated the final residues of traditional art. So what has happened since? In a way, it could be said that, following the qualitative leap made by avant-garde art during the first quarter of the twentieth century, there have been no more essential innovations. But that does not mean that there have not continued to be incidences in the art world from a social, economic, and technological standpoint. Nor can we scorn the aesthetic phenomenon of the avant-garde's own self-destruction, for it failed to survive the final quarter of the twentieth century. That period unceasingly proclaimed itself to be the "postmodern era," which can be understood as a period characterized by the definitive social triumph of modernity and the inanity of continuing to use a strategy of linear struggle for its affirmation. And today, there is no need for an organized phalanx to support the consumption of novelties. That seems to be proven by the fact that the current art market has become hegemonic, and institutional platforms for legitimization, such as contemporary art museums, are not only multiplying exponentially, but have also become artworks themselves.

And now that we've mentioned it, this matter deserves a certain focused consideration. First of all, we must recall that public museums were a creation of "our" time, beginning in the transition between the eighteenth and nineteenth centuries. But so-called "modern art museums" emerged at the end of the nineteenth century and developed during the twentieth. The latter originally emerged as a political response to society's rejection of successive waves of new art entering historical museums. Indeed, that a Manet or a Monet could be hung alongside consecrated works by the Old Masters produced authentic public perplexity. In order to protect new art from social rejection, the Luxembourg was created in Paris, and it soon served as an example for other countries affected by the same problem, including Spain, which founded its first museum of modern art in 1895. That is when public museums were split into three categories—archeological, historical, and contemporary. Significantly, the first and last of the three owed their identities to the fact that they handled, so to speak, "what was not clearly art." Around the end of the nineteenth century, the "discovery" of Paleolithic cave paintings was just one of a string of anthropological surprises that occurred around that same time, so it is hardly strange that a place should be sought out for the conservation, study, and classification of an entire series of strange objects whose formal beauty sometimes led them to be considered artistic, but whose original use was completely unknown. In short, archeological museums became the home for everything that awoke interest more because of its age than because of its artistic quality. On the other hand, museums of contemporary art—an art in a continual state of redefinition—constituted a temporary parenthesis while waiting to see whether their contents could be included in historical museums. In any case, the first "positive" affirmation of a museum for contemporary art was the one that led to the Museum of Modern Art in New York in 1928. It was the result of the brilliant mind of a young art historian and critic, Alfred H. Barr, and two audacious patrons who underwrote that adventure, which we must remember was a private initiative. For Barr, it was not a matter of dedicating a space to art rejected for its strangeness, but rather to art that had its own personality and was generating a singular history. We in no way seek to lessen the unquestionable merit of this undertaking by pointing out that it took place between the two world wars, at a time when the idea of the avant-garde was entering a crisis for the first time. It had already passed the two periods of modern agitation discussed above, but also, for the first time, avant-garde art was receiving social interest, although only among a small elite. The fact is that Picasso, for example, was already considered the fascinating and fearful epitome of artistic modernity after the World War I. Not only did he become internationally known outside artistic circles, he obtained an income worthy of a multimillionaire by the 1920s. This speaks eloquently of the avant-garde in that sense.

Still, between 1930 and 1960, although modern art had generated an increasingly important market and growing social credibility, it was still practically unknown outside minority circles, and was still not the product of mass consumption it was to become from the 1970s onward. Significantly, over the last twenty-five years or so, not only has the number of modern and contemporary art museums multiplied in an overwhelming manner; all sorts of institutional and commercial platforms have arisen to support not just the art of our time, or of the twentieth century, but increasingly, the "latest thing." This anxious polarization towards newness not only represents the triumph of the modern, but also a rethinking of the contemporary art museum, which has overcome its two initial periods, first as a temporary refuge for a polemical product, and then as a place for its exultant affirmation.

But what problems can a massively appreciated contemporary art recognized by all major authorities create for an institution such as a museum? Some think it is a merely functional problem, limiting the question to the idea that the new "type" of artwork using non-traditional supports and materials demands a new sort of building and complementary services. But one would have to be blind not to notice the aesthetic questions involved. A museum is a cultural archive that not only orders the memory of the past, but also periodically adds new and current products to that sequence. That is what happened during the first century of public museums' existence and, as we mentioned above, the process was only provisionally interrupted at the end of the nineteenth century in order to calm social disquiet. During the twentieth century, the idea took hold that it was better to keep that environment separate from the contemporary one, whose generation of innovations was broad and complex enough to constitute its own history. Nevertheless, since the 1970s a growing polarization towards the latest art has posed the need to recycle any past that, merely by being past, is unwieldy. In that sense, we have seen how many of the "new museums" of contemporary art decided at first to do without the first half of the twentieth century and later, to begin removing decades from the second half until they have practically become provisional homes to what is fashionable and new, as long as it remains so.

But what does it mean for a museum to be nothing more than a home for changing novelty, as if, like what is now called a *Kunsthalle,* it were exclusively a platform for temporary exhibitions? And if that is the case, why should we continue to call it a museum? This second question has an obvious answer: no matter what we decide a museum should or should not be, if what now goes by that name has nothing to do with what has defined such institutions until now, it would be better to find a new name, thus avoiding misunderstandings. We must not forget that art has lived almost all its long history outside of museums or anything like them, and so-called public museums, as mentioned above, are barely two centuries old. So it would not be any tragedy if they were to disappear altogether. On the other hand, while an anxious drift towards the latest art by "new museums" or "museums of the latest generation," is increasingly contradictory to the traditional concept of a museum as a historical archive of the past, or of the present's constant entry into the past; there is no real reason to lament it. If these new institutions continue along that path, not only can a new and more adequate name be found for their function; that very function can be formulated in another way. It occurs to me, for example, that the art space that houses new work considered artistic at the time, is a space that is, itself, a work of art, rather than only being so when art is shown in it. At any rate, no matter what one's opinion might be, it seems clear that the question of museums of contemporary art is, so to speak, "all the rage." The debate will definitely continue into the future, which is exactly what concerns us in the present essay.

I believe the moment has arrived to review what has been written up to this point, because if we are to conjecture about the possible future of art, we must not only ask whether, in fact, it *has* a future—something both uncertain and impossible to demonstrate at present—but also, to give our conjecture a certain solidity, we must focus on the only thing we know: what art has been until now—its past—and even more so, what we believe is happening now. This is not easily grasped at first glance, because what is actually happening now has practically nothing to do with what is being publicized as "current reality," for that is nothing more than what is being delivered by the most efficient media powers. In other words, "current reality" is nothing more that what is publicized in the media, that is, publicity.

Still, none of these uncertainties should discourage us, because not knowing everything about something is a characteristic of human knowledge, which is partial or limited no matter what its method or content. By this I mean that what is revolutionary in our time is in no way constricted by the limits of what we call, or used to call, art, nor by the far broader and more diffuse limits of what we seek to include in the field of culture, which is now the subject of much more intense political debates. It is not even restricted by the limits of science, which, with dogmatic ingenuity, we consider the only reliable handhold for us as advanced explorers of chaos. For chaos is merely the overwhelming immensity of what we have not yet managed to discern, unable, as yet, to even minimally sense its real presence. So why should we single out a small, almost insignificant part of this terrifying "vacuum" of knowledge to be attacked simply because it contains not a shred of certainty about anything? After all, unlike all other socially esteemed knowledge and experiences, art has never—neither today nor in the past—presumed to offer us answers. In any case, it serves to formulate questions by bearing witness to what we have been and are, without our yet being able to apprehend its sense. So we could say that art has been, and is, a "memorizer" of those reflections we radiate without really knowing how or why, even though we intuit that their loss—their forgetting—would strip us and our descendants of essential keys.

This perspective should help to attenuate our poorly projected anxieties about art's "guilt" at being unable to guide our fatal irresponsibility. The antithetical path seems much more fecund to me: becoming a part of the questions current art poses, delving into them, rather than disqualifying or trivializing them. In that sense, the legitimate concern we now feel at not being anywhere near capable of grasping the limits of the territory we continue to call art is accompanied by the even vaster and more complex territory constituted by the increasingly indiscriminate social extension of art's "consumers." This "public" accepts anyone who wants to join with no questions about their depth of knowledge in this field, nor their tastes, nor anything else. And yet, it seems clear that the overwhelming vastness of contemporary art's horizons is commensurate with that of its flock of followers, so any revelations about one will also shed light on the other.

But what does all this have to do with what is periodically trumpeted as the newest of the new, based on banalities such as how unusual its support is, or its renewed panoply of techniques, or any of its symbolic reformulations? I do not believe, for example, that the questions a current video-artist, or any member of that tribe of so-called "immaterialists," asks himself are any different than those that Francisco de Goya, Michelangelo, or Phidias asked themselves. And I do not believe they have changed because, among other things, we have yet to find a plausible explanation for the existence of Paleolithic cave art, even though from the caves of Altamira to Demien Hirst we continue to radiate luminous creations that we feel are essential to knowing ourselves and living more, and better. So what, then, is the future of art? A good way of dealing with this unanswerable question is to reconsider everything that art has been until now, and is thus part of its past. Maybe none of it will have anything to do with what happens tomorrow, but the only sure way to find out is to survive until then and live through it. So, how will the art of the future and the future of art be? Allow me the truism, "we will soon know," for that knowing is nothing but noticing how the future "presents" itself to us continually by becoming the present.

the architecture of the new century: around the world in ten stages

LUIS FERNÁNDEZ-GALIANO

Opposite page:

The Guggenheim Museum in Bilbao, by Frank Gehry.

Both technique and art, architecture is also a constructed expression of society. As technique bordering on engineering, it has experienced the impact of new materials and innovation in the areas of construction, structures, or installations, facing the historic challenge of sustainability. As public art, it has been a participant—and sometimes a protagonist—in the renewal of visual language and the aesthetic mutations of a period marked by the spectacle. Lastly, as constructed sociology, it has given form to the colossal transformation that has urbanized our planet, replacing traditional landscapes with sleepless megacities.

The classical treatises at the root of Western architecture already speak of these three complementary facets when theorizing on a discipline that overlaps so many others. Ever since Vitruvius, in Roman times, architecture has been assigned the task of reconciling technique and art with the social use of its spaces, and the motto, *firmitas, utilitas, venustas* (solidity, usefulness, beauty) has been shorthand for this approach. But those three facets are so impossibly intertwined in concrete works of architecture that it is difficult to consider them separately, and here we have sought out a different strategy.

Instead of describing technical, functional, and formal innovations that characterize architecture at the beginning of the twenty-first century, we have preferred to select ten episodes in different cities on the planet that offer both a sequence of significant achievements in the last two decades, and an illustration of tendencies or phenomena of a more general nature. Those episodes, which are presented in more-or-less chronological order—from Berlin following the Fall of the Wall, Bilbao and the Guggenheim, or New York and 9/11; to Olympic Beijing and the titanic works of the petroleum autocracies of the Persian Gulf or Russia—are also organized so that the their consideration herein resembles the stages of a trip around the world.

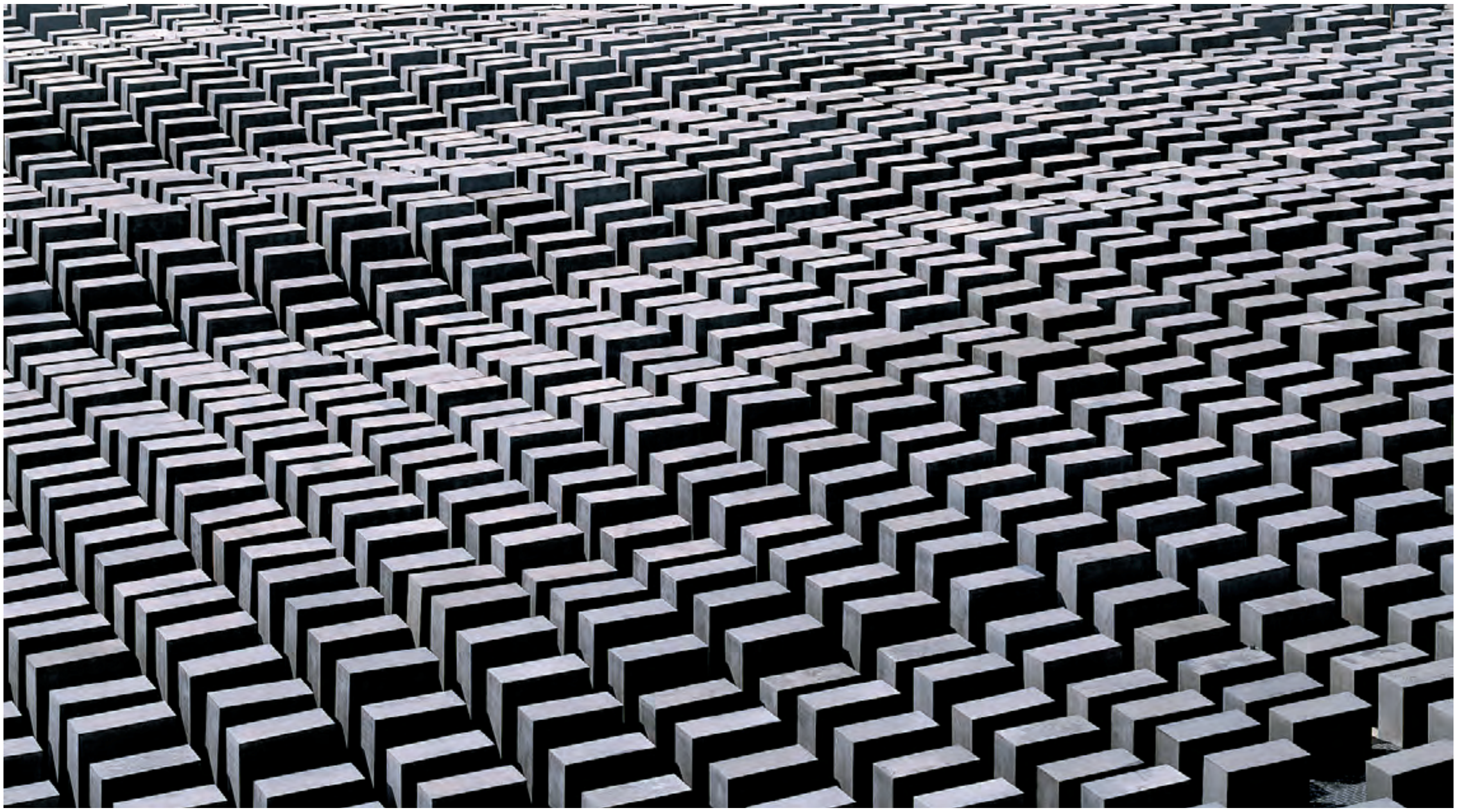
Ever westward, and always in the Northern hemisphere—which leaves an enormous amount of geography out of the picture—our journey begins in Europe at the close of the twentieth century and of the Cold War, marked by the demolition of an urban border. It then travels to the United States, which saw the destruction of the Twin Towers as the parting shot for its “War on Terror.” Next is Asia, which builds energetic signs of its economic force, and finally, Russia, astride two continents. It, too, is using architecture to affirm something, namely its recovery following the dissolution of the Soviet Union. After ten stages, the circle closes with another political ice age that coincides with an economic cooling, and financial and social convulsions, in a cumulus of fractures and trembling that architecture registers with the exactitude of a seismograph needle.

Berlin without the Wall: the architecture of memory in the face of ideological struggles

Our journey begins in the city where architecture most faithfully reflects ideas, capital of a totalitarian empire defeated in 1945 and frontier for four decades between the democratic West and the Communist bloc. Since the demolition of the Wall in 1989, Berlin has continued to be an urban laboratory where architecture is subjected to the demanding filter of ideology and memory. Such is the case of the Jewish Museum by the United States architect of Polish origin Daniel Libeskind—a group of fractured and unstable volumes added to a baroque building—and the new Reichstag by British architect, Norman Foster, which is a critical restoration that transforms the character of a legendary headquarters; as well as the Holocaust Memorial by New Yorker, Peter Eisenman, an extension studded with concrete stellae that turn this monument into an urban landscape.

The Zigzag shape of the Jewish Museum alludes to German history's dramatic changes of direction and the tragic interruption of the Jewish presence in that city, but it is also of singular architectural importance. With the IBA of 1985—an exhibition whose objects were buildings made on a scale of 1:1 in different neighborhoods of the city—Berlin became the main stage for the postmodern movement that foreshadowed a return to classicist architecture in opposition to the abstractions of modernity. And with Libeskind's project, announced around the same time as the Fall of the Wall, Berlin was to construct an icon of deconstruction, a rival tendency launched with a show at MoMA in New York in the summer of 1988 that defended fractured and catastrophic architecture as an expression of a convulsed world.

No city can better personify convulsion than Berlin, the epicenter of two World Wars that left the ruins of its former parliament as a mute witness to the collapse of democracy and the Wagnerian defeat of German expansionism. When Gorbachov capitulated to Reagan and Thatcher at the end of the Cold War, allowing Germany to reunite and Berlin to recover its status as capital, Foster rehabilitated the old Reichstag, making it



Holocaust Memorial in Berlin, by Peter Eisenman.

The Tate Modern in London, by Herzog & de Meuron.

the new Parliament of a nation determined to impede the return of the specters of an ominous past. To do so, he crowned the massive Wilhelmine structure with a glass dome enlivened by spiral ramps. That dome serves as a place to observe the city and symbolically places its citizens above their political representatives, overseeing their assembly to prevent new historical derailments.

Near this regenerated Reichstag—even the artist, Christo, exorcised it by covering it with canvas before the renovation began—Eisenman built a colossal and lyrical memorial to the murdered Jews: a screen of concrete prisms that is simultaneously the waving landscape of planted fields and a disquieting labyrinth running among exact tombs. Originally conceived with the sculptor, Richard Serra, this commemorative monument—so different in its jagged abstraction from most of the Holocaust museums that have sprung up in recent times—is a gesture of penitence in the heart of horror. At the same time, it is an effective example of architecture's capacity to express ideas through form.

Rotterdam or Basel: new landscapes and old cities in an indecisive Europe

Following the exhaustion of the postmodern style—which had its showroom in Berlin, and its think tanks in Milan and New York—architecture's debate did not move to Paris, where Mitterrand's grand presidential projects combined geometric monumentality and the glamour of celebrities. Nor to Blair's London, which feted the millennium with a technological and *cool* architectural third stream. Instead, the debate took hold in two medium-sized European cities: Rotterdam in Holland, and Basel in Switzerland. In the first, numerous young architects inspired by the abrasive talent of Rem Koolhaas—especially those who worked under the initials of MVRDV—exacerbated modern language with accents of Russian constructivist utopias from the nineteen twenties, applying them to the urban landscape. In the second, a new generation of German Swiss architects emerged. There, the creative energy of Jacques Herzog and Pierre de Meuron quickly stood out—with the permanent rural and essentialist counterpoint of Peter Zumthor—as they created a stronghold of constructive excellence, demanding art, and sensitivity to the material heritage of ancestral territories.

Dutch hypermodernity was fed by the *tabula rasa* of a city devastated by the war, on the artificial territory of a country of *polders.* But it was also fed by Koolhaas' futurist fascination with the metropolitan crowding of a New York that was, for many years, his adopted home, object of study, and intellectual laboratory, especially in the IAUS (Institute for Architecture and Urban Studies) directed by Peter Eisenman. Combining the formal grammar of Le Corbusier and audacious Russian diagonals with American pragmatism, those architects created an optimistic, playful school in the Netherlands. Soon, they were flirting with the fragmentation and droopiness of Anglo-Saxon deconstructivism—drawing largely on the extreme ductility offered by new systems of computer representation. But their finest manifestation was artificial landscaping in which a building is surreally penetrated by the topography of its surroundings, creating a "free section" that puts a new spin on the "open floor plants" of the historical avant-garde.

The Swiss Germans, on the other hand, developed a "degree zero" of architecture with elemental and exquisitely constructed prisms deeply rooted in the traditions and territory of their Alpine country, but also influenced by the rigorist teachings of Aldo Rossi, who was Herzog and De Meuron's teacher at the ETH in Zurich. Defiantly archaic but tightly linked to the art scene—initially through Joseph Beuys, and later with multiple collaborators from the art world—the two partners in Basel became the leaders of their generation with a series of decorated boxes characterized by great material and tactile refinement, and a series of interventions in industrial buildings—especially their conversion of a power plant into the new home of Tate Gallery in London—that showed the relevance of an architecture of continuity.

In a Europe characterized by economic and political fatigue—hesitating between the modern messianism of building a contemporary city *ex novo,* and its cultural and emotional ties to its heterogeneous urban heritage—the Dutch and Swiss supplied opposing architectural and urban models, establishing a fertile disciplinary dialog between Rotterdam and Basel. In time, this resulted in a cautious convergence of the two schools.

Bilbao and the Guggenheim: the spectacle of a museum as urban motor

In 1997, the inauguration of the Bilbao branch of the Guggenheim Museum—an undulating sculptural accumulation of titanium sheeting designed by the Californian, Frank Gehry—was a media event that changed the course of both architecture and museums. Of course, that New York institution already had original premises of great architectural singularity and beauty—the famous spiral ramp built by Frank Lloyd Wright on Fifth Avenue emblematic building beside the estuary in Bilbao had such significant iconic precedents as the Sydney Opera House, where the Dane, Jørn Utzon, designed concrete sails that made it the symbol of Australia, or—in terms of museums—the Pompidou Center in Paris, in which the Italian, Renzo Piano, and the Englishman, Richard Rogers, interpreted the countercultural spirit of Paris' 1968 youth demonstrations with a joyful, colorist, and technological futurism.

Bilbao's Guggenheim took one step further, because it entirely subordinated art to the spectacle of architecture, turning the latter into a gigantic sculpture with delicately matt reflections that is reckless in its detained stormy movement. A critical and popular success, the museum attracted numerous visitors to a rough city of obsolete industry that had, until then, been far removed from artistic and tourist circuits. It became a powerful motor for urban regeneration and showed the capacity of cultural infrastructures to contribute in the transition towards a service economy. What became known in Spain as the "Guggenheim effect," and outside the country as the "Bilbao effect," spread like wildfire, and mayors of every decaying city in decadence sought to obtain an emblematic building that would beckon to tourists and investors, improving self-esteem, and acting as the logo for a change of image.

This use of architecture for the modernization of identity and urban *rebranding*—which went so far as to affect cities of the dimensions and character of London or Rome—accentuated the discipline's drift towards sculpture, as each new cultural center or sports stadium had to be unmistakable and surprising. That was the case with museums, of course, but also with libraries, auditoriums, and stadiums, all of which had to reconcile their specific functions with their symbolic role. Even buildings with such organizational demands as stations and airports—in Bilbao itself, the subway stations were designed by Norman Foster and the airport by Santiago Calatrava—became a part of urban identity, following a path blazed by large corporations that promote singular skyscrapers as the image of their brand on the city skyline.

In 1967, Guy Debord theorized *The Society of the Spectacle,* but four decades later, his intuition remains fully applicable. The absorption of architecture by show business has a bittersweet taste. On one hand, it brings greater visibility to those works, making them the object of social debate in the media, as can be seen in recent works by such demanding and secretive masters as Álvaro Siza or Rafael Moneo. On the other, it turns architects into *glamorous,* stylish celebrities. Gehry designs jewelry for Tiffany's and Koolhaas or Herzog & de Meuron design stores for Prada, while the Anglo-Iraqi, Zaha Hadid, designs a sinous portable

Opposite page:

The headquarters of the *New York Times* in New York, by Renzo Piano.

pavilion for Chanel. And all of them frequently appear in advertisements for luxury consumer items, as lofty representatives of discriminating aesthetics and avant-garde elegance.

New York after 9/11: the future of skyscrapers and the future of the empire

The fourth stage of our trip around the world takes us away from Europe, where so many expectations were raised by the end of the Cold War and the hedonistic enjoyment of the dividends of peace, across the Atlantic to New York, the setting for a titanic attack that produced a tragic massacre and turned the tides of contemporary history. The group of young, suicidal Islamic militants directed by architect and urban planner, Mohamed Atta, demolished two Manhattan towers designed by the American architect of Japanese origin, Minoru Yamasaki, that symbolized the financial power of that city and the global leadership of its nation. Their atrocious act provoked an unprecedented geopolitical crisis, and, in passing, it raised questions about the future of skyscrapers, the buildings that best represent the twentieth century's architectural challenges.

In effect, the destruction of the Twin Towers redrew the planetary borders of conflict, and the by-then extinct rivalry between capitalism and communism was replaced by the confrontation between the West and Islamic fundamentalism. At the same time, the prestige of the superpower erratically led by George W. Bush suffered a devastating blow that was worsened by the errors of the posterior "wars on terror" in Afghanistan and Iraq. Its economy experienced the drag of military spending and financial baroquism, and New York suffered a wound that has yet to heal. The intellectual, aesthetic, and administrative fiasco of the architecture competitions launched to rebuild the ominous vacuum of Ground Zero is one more sign of a loss of touch that leads us to fear a foretold decadence.

And yet, the predicted end to skyscrapers—to whose complexity and cost was now added an extreme vulnerability—has never arrived, and towers continue to spring up everywhere. Safety measures have been revised and budgets have inevitably increased, but major public and private protagonists of power continue to build skyscrapers that manifest their strength in the form of height. Many corporations have turned their eyes to office parks, and towers over 200 meters high are hardly justifiable in economic terms, yet the drive to break planetary or regional records continues to feed competition among cities or countries, garnering media attention and awakening popular curiosity.

Even New York, which directly suffered 9/11, has not renounced its traditional designation as "the city of skyscrapers." It continues to build and design new towers, often linked to its persistent cultural and artistic heritage, such as the headquarters of the Hearst and New York Times news groups (designed by Norman Foster and Renzo Piano, respectively), the small stacking of the New Museum (by the Japanese architects, Sejima and Nishizawa), or the residential skyscraper designed by French creator, Jean Nouvel, alongside the MoMA. That sector—luxury living spaces designed by great architects—has certainly prospered in New York. And in Chicago—birthplace of the skyscraper and home to legendary buildings by Sullivan, Wright and Mies van der Rohe—it has led to a spectacular project by Santiago Calatrava, the same Spanish architect who is constructing the only relevant work in Manhattan's afflicted Ground Zero, a monumental subway station.

Las Vegas as a paradigm: the urban design of leisure and the world as a theme park

America gave birth to the skyscraper, which takes urban density to its most hyperbolic extreme; but it also paved the way for the most scattered urban sprawl. With the help of the automobile, it rolled out the city's territory like a thin carpet of houses and gardens. Such unanimous suburbanization, wasting space, time, materials, water, energy, and land—not to mention the transit infrastructure—has very successfully spread around the world. In that setting, the collective domain is relegated to large commercial agglomerations that are often presented with the trappings of traditional urbanity, figuratively interpreted with the same scenographic resources as Disney's amusement parks or the thematic casinos of Las Vegas—so admired by Warhol's Pop-art gaze, and Venturi and Scott-Brown's approach to architecture.

Las Vegas, Nevada, the fastest-growing city in the United states, is also a fine paradigm of postmodern city planning, whose tendencies it exacerbates to the point of paroxysm. This city is an effective metaphor for the contemporary rise of "casino capitalism"—as brilliant, noisy, and massive as the gaming rooms that stretch unbroken from the lobbies of endless hotels offering the illusion of leisure in this neon city. The Egyptian or Venetian motives of casinos in Las Vegas—like the Wild-West towns or Snow White's castles in countless Disneylands scattered around the world—return like stubborn echoes in malls throughout the United States and the world, and the urban design of consumerism clumsily apes the traces of a long-gone urbanity.

Opposite page:

The Guggenheim Museum in Las Vegas, by OMA/Rem Koolhaas.

The central role of commerce in these new ways of occupying territory—admirably analyzed by Koolhaas in his description of contemporary *Junkspace*—is unarguable, and the morphology of the shopping center—enormous sales areas with built-in *food courts*—has infiltrated the remaining infrastructures of transportation, leisure, sports, culture, health, and work whose activity nodes articulate the indiscriminate spread of residential construction. Airports and stations, amusement parks, stadiums, museums, and even hospitals, college campuses, research and business compounds—all suffer the invasive penetration of the mall, whose stores and restaurants end up as the protagonists of meeting and social areas in the theme-park suburbanization of the world.

Even compact cities from the European tradition, extended with anonymous and indistinct low-density peripheries, reformulate their historic centers to include leisure and tourist spaces, capacious, open-air shopping centers where boutiques, fashion stores, bars, and cafés rub elbows with the occasional palace, church, or museum. Thus, cities like Bohigas and Miralles' Barcelona—a showcase for the 1992 Olympic Games and an exemplary model of urban transformation, equally concerned to "clean up the central district" and to "monumentalize the periphery"—illustrate this contemporary drift that creates an urban setting to service occasional visitors, far removed from the modern, avant-garde, and even utopian fundaments of the initial project.

Tokyo in cartoons: tradition and modernity in Japanese density

The mid point of our trip is actually the meridian that marks the dateline. On the other side of the Pacific Ocean, the seventh stage of this journey brings us to Tokyo, a metropolis whose form has lost all memory, where surviving traditional habits coexist with a futurist urban landscape, packed with multicolored signs and the spasmodic animation of cartoons. At the beginning of the twentieth century, the fascination with an exotic "Japan-ism" tinted the language of the artistic avant-garde, while, for architectural modernity, the "Empire of the Sun" was the source of the extreme rationality of wooden construction, the modular lightness of houses divided with tatamis and rice paper, and the laconic and ceremonious refinement of objects. From Frank Lloyd Wright and his Viennese disciples in California, to Berlin architect Bruno Taut's round trip, or the discovery of the Far East by Alvar Aalto and his colleagues from the school of Scandinavian organicism, Japan and modernity have been architectural synonyms.

Nowadays, though, Japanese hyper-urbanity offers a model far removed from the shaded introversion of the ageless home. Were Tanizaki to rewrite *In Praise of Shadows*—a fundamental text for the Zen sensibility of Western minimalism—he would now be writing *In Praise of Neon,* the emblematic exponent of a juvenile, ultracommercial pop culture as jangling as that of Las Vegas, although here it is adorned with the infantilism of *manga* and the cybernetic autism of the *otaku,* and fully given over to the worship of luxury labels that dot the urban landscape with their exquisite and hermetic shops.

Beyond immaculate gardens and geometrically-exact museums—many of concrete and glass, in which Tadao Ando successfully combined the formal languages of Le Corbusier and Louis Kahn—it is fashion stores that best reflect Japan's current social climate. Some are made by foreign architects—the extraordinary faceted crystal designed for Prada by Herzog & de Meuron, or the lyrical translucent prism erected for Hermés by Renzo Piano—but more often they are examples of the most refined local architecture. Sejima and Nishizawa's aleatory overlapping for Dior or Toyo Ito's arborescent blinds for Tod's, and the branches of luxury firms in Omotesando or Ginza—Tokyo's two fashion districts—bear witness to a hyperbolic exacerbation of luxury consumerism that surpasses its forebears in Europe or the United States.

The great transportation infrastructure in the rest of the country is overshadowed by this innocent ostentation, nevertheless, it has such outstanding examples as the colossal airport at Osaka, built by Piano on an artificial island, or the delicate maritime terminal in Yokohama, designed by Zaera and Moussavi with undulating wooden platforms. There are also quite singular cultural works, such as the media center in Sendai, which Ito holds up with tangled strands of metal pillars; or the museum of Kanazawa, whose contours were delimited by Sejima and Nishizawa with an evanescent circular perimeter. And all of these are set in a light, fluid public domain that is reflective and streamlined, as limpid as it is frigid. In any case, it lacks the magnetic and centripetal magic of the private and exclusive strongholds of the most sophisticated, empty luxury in the heart of the "Empire of Signs" that is Tokyo.

Olympic Beijing: the central role of China's icons in the rise of Asia

If Tokyo is fashion, Beijing is spectacle. The inauguration and unfolding of the Olympic Games in the summer of 2008 allowed China to be proud of its economic and social achievements, offering the world a formidable

Opposite page:

The headquarters of Tod's in Tokyo, by Toyo Ito.

Previous page:
The Olympic Stadium in Beijing, by Herzog & de Meuron.

example of its organizational capacity with an event in which architecture was rather more than a mere mute stage for the ceremonies and competitions. The new airport terminal where athletes, spectators, and journalists arrived, the television headquarters where the Games were transmitted, and at the top of the list, the stadium and swimming pools—all these great works carried out for the occasion—spoke of China's quest for excellence. Even though they were almost all designed by foreign architects, they bore witness to the ground that has been covered by the "Middle Kingdom" over the thirty years that have passed since the political changes of 1978, when the chaotic Maoism of the Cultural Revolution was replaced by the single-party capitalism propelled by Deng Xiaoping.

The new terminal, which is also the largest building on Earth, was designed by Norman Foster, who also designed Hong Kong airport—on an artificial island, like that of Osaka. With characteristically British technological refinement, he knew how to interpret the red columns and floating roofs of traditional construction, using the steel and glass of advanced engineering to create an endless and luminous enclosure that protects passengers from the airplanes under a roof as light as a festival dragon or a paper kite. The installation was inaugurated a year before the Games, as was another great work promoted for the event: the National Theater erected by French architect, Paul Andreu—curiously, he is also known for his airports—beside Tiananmen Square. It is a gigantic titanium dome that emerges from the quiet water of a vast pond.

The sports events had a liquid protagonist in the indoor pools. Designed by the Australian team PTW, what soon came to be called the "Water Cube" was a large prism whose bubbling façade is made with pillows of translucent ETFE (ethyltetrafluorethylene) plastic. Most of all, though, the games enjoyed a formidable setting in the Olympic Stadium, a titanic steel tangle thought up by the Swiss architects, Herzog & de Meuron, with the aid of the Chinese artist, Ai Weiwei. This, too, received a fond nickname from the public—"The Bird's Nest"—and its extraordinary formal singularity has made it an icon of the Games and a symbol of China's drive, which reached its zenith when its spectral and trilling nocturnal appearance was complemented by the spectacular opening and closing ceremonies, replete with choreography and fireworks.

Inevitably, the television headquarters—two towers linked at the top to create the bent frame of a colossal urban gate, designed by the Dutchman, Rem Koolhaas—was the most polemical building. This was not so much because it was not finished in time for the Games as because the governmental character of information is one of the most polemical aspects of this country, which combines economic success with a stricter state control than in the West.

The Pyramid of Peace and Reconciliation in Astana, by Norman Foster.

Astana on the steppes: a new capital in the land of the Great Game

Our eighth stop is undoubtedly the most exotic because we associate the steppes of central Asia less with architectural achievements than with the music of Borodin or Kipling's writing about the geostrategic Great Game of the Eurasian empires. In this land of crossroads and nomads, it wasn't long ago that many would have listed the *yurt*—a circular tent of exquisitely defined construction—as the steppes' most original contribution to the history of human lodgings. With the disappearance of the Soviet Union, however, a new actor appeared on the international stage: Kazakhstan. With oil reserves, its charismatic president decided to leave his mark on architecture with a new capital: the existing Almaty—the legendary Alma Ata—was to be replaced by Astana, a city created *ex novo* on the trans-Siberian railway line, and many of the world's most important architects would be called to design it.

In the tradition of Pandit Nehru's Chandigarh or Juscelino Kubitschek's Brasilia (developed by Le Corbusier, and Lucio Costa and Oscar Niemeyer, respectively), the Astana of Kazak president, Nursultán Nazarbayev, was laid out by the Japanese architect, Kisho Kurokawa. Its most significant buildings are by the Englishman, Norman Foster. Thus, Kazakhstan is no longer just the country associated with the British comedian Sacha Baron Cohen—the polemical Borat—and Astana is no longer just the name of a

Ecocity designed by OMA/Rem Koolhaas in Abu Dhabi.

Ecocity designed by Norman Foster in Ras al Khaimah.

cycling team. The country and its new capital have become a new and audacious chapter in the story of contemporary architecture.

Of course Foster is not the only Westerner with important commissions in Kazakhstan. Despite the administrative transfer of the capital, petroleum income continues to foster a singular building boom in old Almaty, where many US and European studios—including Rem Koolhaas' OMA, which is building a large technology campus on the outskirts of the city—express the country's economic vigor in that territory. In Astana, though, Foster's London firm is the absolute protagonist of emblematic architecture. It has already finished a colossal pyramid and is raising an enormous transparent tent-like structure that will be the city's ceiling when finished.

The pyramid or "Palace of Peace and Reconciliation"—inevitably called the "Pyramid of Peace" by public and media alike—is home to periodic interfaith conferences and seeks to reconcile the country's different races, cultures, and religions with its archaic and exact geometry, crowned with a translucent vertex of innocent stained glass with doves. The "tent," which houses 100,000 square meters of leisure space under a surface of ETFE held by masts and cables, is well over twice as tall as the pyramid and is practically its symbolic opposite. Establishing an unexpected dialog between the steel points of the ideological temple, and the plastic warps of the titanic tent dedicated to spectacles and consumerism, it links old tribal and religious identities with the new sense of belonging to a global tribe that worships only prosperity and entertainment.

Dubai and the Gulf: oil cities and the challenge of sustainability

Our next stop takes us to another real-estate boom driven by petroleum, but in this case the dimensions and speed are such that theoreticians of the contemporary city such as Rem Koolhaas have no qualms in calling it "a new urbanity," an until-now unknown way of producing urban tissue. Bordering on science fiction, construction in the Emirates of the Persian Gulf was fed initially by the exploitation of oil wells, but is increasingly linked to financial and tourist flows. It extends from a surreal landscape of skyscrapers that emerge from the desert sands like a rosary of artificial islands in the form of continents or palm trees, including innumerable educational and cultural infrastructures that house franchises from the United States and Europe's leading museums and universities.

In many ways, Dubai was the pioneer. With far less petroleum than the other emirates, it quickly redefined itself as a regional financial center for the Middle East—capable of replacing Beirut, which was devastated by war and political conflicts—and as a destination for luxury tourism for new millionaires from Russian and Europe. Built with the expertise of Anglo-Saxon project managers and the effort of an army of immigrant workers from India, Pakistan, and Southeast Asia, who have almost no civil or labor rights, this forest of skyscrapers with a ribbon of thematic islands boasts the most luxurious hotel in the world—Burj al Arab, by the British firm, Atkins—and the highest building on the planet—Burj Dubai, by the US firm, SOM. These are economic and technological records, and they are also undoubtedly social indicators, but sadly they say little about the quality of the architecture, in which the accumulation of important names has not yet generated any masterpieces.

Qatar has a different strategy. It seeks to become an intellectual center, with an ambitious city of education designed by global architects such as the Japanese,

The Crystal Island Complex in Moscow, designed by Norman Foster.

Arata Isozaki, the Mexican, Ricardo Legorreta, the North American of Argentinean origin, César Pelli, and the Dutchmen of OMA. And two other emirates also have different political and urban objectives. Ras al Khaimah seeks to promote sustainable tourism in a setting of superb natural beauty, while Abu Dhabi, capital of the United Arab Emirates, has begun work on a spectacular cultural district, with branches of the Guggenheim and the Louvre.

The most visionary projects in Ras al Khaimah—including a dreamlike tourist center high in the mountains and an ecological city on the coast, with an emblematic spherical convention center—are all by Koolhaas, the selfsame theoretician of the Gulf's urban boom. In Abu Dhabi, though, the participation of great names is more choral: Frank Gehry, Jean Nouvel, Zaha Hadid, and Tadao Ando handle the museums and theater in the cultural district, while the powerful studio of the ubiquitous Norman Foster carries out everything from an exemplary sustainable city *(carbon neutral)* with collective transit and energetic self-sufficiency, to a lyrical interpretation of the traditional bazaar, in the city's new Central Market. Unexpectedly, the place in the world with the greatest energy reserves does not simply promote ostentation and consumerism. As Koolhaas and Foster's eco-cities demonstrate, abundance does not exclude testing future forms of austerity or scarcity.

From Moscow to Saint Petersburg: the titanic works of the Russian autocracy

Our final stop is quite close to where we began, in the same Russia that placed the physical and symbolic border of the Cold War in Berlin. Stimulated by its control of the oil and gas needed by much of Europe, it has recovered the imperial pride of the Czarist autocracy and the implacable self-esteem of Soviet Stalinism. In tune with the Eastern authoritarianism of Beijing, Astana or Dubai and enjoying the same impulsive throb of sudden prosperity, Moscow has unleashed a whirlwind of megaprojects, employing the eloquence of architecture to define the renewed ambitions of the Eurasian colossus. Inevitably, this building boom is centered in the capital, but it touches many other cities, especially historic Saint Petersburg.

Both cities have a very significant presence of British architects, but in Moscow we must underline the material and media presence of the same Foster who designed Beijing Airport, the "Pyramid of Peace" in Astana and the sustainable city in Abu Dhabi. With both the Russia Tower, whose 612 meters make it the highest skyscraper in Europe, and the Crystal Island on the banks of the Moscova River—a true city under a gigantic spiral roof that not only improves its climate but also makes it the largest construction on the planet, surpassing Foster's own record-holding

Terminal 3 in Beijing—this British architect adequately represents the regenerated vigor of this country. A nation that, as the Georgia crisis showed, will no longer allow itself to be treated with the commiserative disdain that followed the dismemberment of the Soviet Union and the subsequent decline of Russian power.

Saint Petersburg deserves separate mention. It is Russia's cultural capital and the birthplace of Vladimir Putin, who made it the headquarters of Gazprom, the Russian energy giant. Following a polemical contest in which the leading lights of international architecture were invited to compete—the Scottish studio of RMJM will build a colossal skyscraper that will dwarf Smolny Cathedral, on the other side of the Neva, eloquently manifesting the role of fossil fuels in Russia's rebirth. This country, which intimidates the governments of Eastern Europe with its gas pipelines, permits itself the luxury of having a former chancellor of Germany on the payroll of its energy company. As we end our journey here, we do not know whether the Cold War really ended two decades ago, but we are certain that architecture will continue to express the ambitions and conflicts, achievements, and disappointments of countries and regimes, companies, and people.

Closing what is more a vicious than virtuous circle, the emblematic architecture that today expresses the power of Russia, China, and the Arab Emirates is by the London architectural studio whose remodeling of the Reichstag retained obscene graffiti written in Cyrillic characters by the Russian soldiers who took Berlin. It is no coincidence that this firm is mentioned in seven of the ten sections of this text, for it is undoubtedly the most aggressively global of them all. A historical cycle has been completed and the end of the bipolar world that allowed the reunification of Germany following the Fall of the Wall in 1989 has given way—after a brief interval in which the only remaining superpower has failed in its efforts at global government—to a multi-polar scenario that architecture emphasizes with a proliferation of concentration points.

A provisional epilog: the dawn or dusk of a mutating discipline

In this ever-westward journey, it is difficult to avoid a melancholy tone as our story ends. The itinerary of architecture over the last two decades has transformed a modest craft based on technical knowledge, functional pragmatism, and aesthetic discrimination into an activity bordering on the clamor of publicity, the avidity of consumerism, and the whirlwind of fashion. The humility, perseverance, and silence that used to characterize it has been replaced by boasting self-confidence, capricious invention, and a loquacious justification of nonsensical proposals that can only be explained by the insatiable appetite for novelty of pupils and palates fatigued by an overly-prosperous society.

The great challenges for a species that is now mostly urban—from climate change and sustainable construction to the material orchestration of life in megacities like Mexico, São Paulo, Lagos, or Calcutta—seem to be outside the realm of this introspective practice so capable of creating emblematic or iconic works and tragically incapable of significantly improving the habitability and beauty of contemporary cities. As has so often been said, these are fine times for architecture (in the restrictive sense of erecting singular buildings) but bad times for cities, that is, for that setting that belongs to all of us and represents all of us.

Never in recent history have architects been so famous, but they may never have been so incapable of shaping the environment we live in, either. Just half a century ago, anonymous architects—known only to their colleagues and other specialists—worked in their studios, laying out urban plans and large collective housing projects that decisively affected everyday life for the majority. Today, media-star architects have become arbiters of fashion and dictators of taste, but they hardly have the capacity to participate in major decisions that shape cities and land. Those decisions are now made almost exclusively by economic forces and movement flows crystallized in transport infrastructures.

At any rate, architecture is an archaic and tenacious discipline that may have suffered a disconcerting process of change to fit into the society of the spectacle, but it has never abandoned its essential core of technical-constructive intelligence, orchestration of changing social needs, and symbolic expression of the times: the venerable *firmitas, utilitas* and *venustas* of Vitruvius. That is why the elegiac tone of these conclusions may be mistaken—incompatible with the headstrong confidence needed to carry out this demanding profession that so expertly reconciles the pessimism of intelligence with the optimism of willfulness. By traveling west, we gain a day along the way, and perhaps that uncertain light we take for dusk is actually a dawning for this useful, worldly art.

frontiers and knowledge in music? a few notes

LUIS DE PABLO

It is tempting to think that the subject of this work falls outside the capacities, or even the interest, of an artist—in my case, a composer, which only makes things worse, given the elusive nature of music as a language.

And that may even be the case. It all depends on the meaning we assign to the term "knowledge."

I'll avoid the useless trap of enumerating the avatars that the verb "to know" may have assumed over the course of its history, limiting myself to the most immediate dictionary definitions. The *María Moliner* dictionary of the Spanish language (second edition, 1998) offers the following primary definition of "knowledge": "The act of knowing." This is followed by, "The effect of knowing or the presence in the mind of ideas about something… things one knows with certainty, art… the capacity of knowing what is or is not convenient and of behaving in accordance with that knowledge… prudence, sensibility…" and so on, forming a wise and considerable list.

I believe the question we are being asked, at least as I understand it, is not so broad. If I have understood it correctly, we are being asked whether our "knowledge" —"act of knowing," "presence in the mind of ideas about something"—of our field of activity—in my case, music composition—could or should have limits, either because of the incapacity of our sensory organs—and their man-made aids—or because of the risks that "knowledge" might entail if used in an irresponsible or harmful manner.

Off the cuff, I can only answer in one way: music doesn't involve any knowledge of that sort. It is neither a question of incapacity nor of risk. Quite simply, music occupies a different place, as human as that of science —and maybe even more necessary to humanity's inner equilibrium—but it responds to different needs or, if you prefer, fulfills different functions.

In the interest of greater understanding, allow me to formulate this question from a composer's point of view: something like "the frontiers of artistic expression." It would be interesting to pose that question to my admired colleagues. I am certain that I, at least, would learn much from doing so. Before continuing, I would like to clarify something. There are other "frontiers," which I will mention in the interests of disclosure, but will not actually discuss. Music involves ordering—or disordering—sounds and silences in time. We already know there are divergent opinions about this, but I will not enter that debate. There is also the "frontier" of music's meaning in present-day society. After all, music is not merchandise, even though composers and performers try to make a living from it. But some musicians are born *to be* merchandise. Even worse, there is excellent music that has been used in such a way that it has become merchandise. But I will not enter that jungle, either—Pascal Quignard quite rightly speaks of *La haine de la Musique* ("Hate of Music")—for I am not a sociologist. Let us return to the matter at hand.

If anything is clear after consulting a definition of "knowledge" it is that it "lies"—so to speak—in the conscious, that is, in the supposedly rational and conscious world. To many people, "unconscious" knowledge may seem little more than a play on words, a contradiction and almost an oxymoron.

And yet, given what is involved in carrying out certain practices that imply the acquisition of knowledge, this may not be so clear-cut. Learning a language, for example, involves two clearly differentiated phases. The first is a conscious, constant, and deliberate labor of memorization. The second could be called assimilation, when what has been learned with conscious effort enters an individual's unconscious—such are the benefits of practice. When that happens, language no longer requires any deliberation. It simply comes out, for better or worse, with the spontaneity of what is supposedly known and assimilated.

Music, like any other language, is learned that way in its most elemental phase. We could even say, without exaggerating, that *everything* is learned that way, when it does not reach beyond that level of *practical* knowledge.

In that phase, music is a craft like any other—and I use the term "craft" it its most noble sense, because it is, indeed, noble. As such, its knowledge has frontiers defined by efficiency. They reach from mere sufficiency to levels of prodigy.

But behind its "craft," music holds many surprises, other kinds of knowledge that are not so much learned as invented. A trained composer—even before he is trained—finds himself facing the door of the enigma that led him to choose that profession. He is, in effect, face-to-face with his craft. He has learned many things. Some he considers useless, others he would rather forget—at least that is what he thinks—and still others he keeps just in case… But now, he must find his voice, and that, by its very definition, calls for "another" craft that has yet to be defined. Its definition—as in Kafka's short story, "The Trial"—is absolutely personal: only he can define it. Before, he had to assimilate pre-established teachings without objecting, but now he must be judge and jury; he must seek, find, use, and evaluate. The possible "knowledge" such work can offer him will not be a code of inherited rules; it will be one derived from his expressive objectives. If, as the years pass, he meets those objectives, they will become the heritage of a more-or-less significant collective: a profound and always faithful, though partial, reflection—there has yet to be one that spans all of humanity—of what it meant to be a man at a specific time and place (I will return to this idea, below).

With what I have already said, I believe we can accept the thesis that music is not about transmitting solid, invariable or even evolving "knowledge" about anything. It shares this with the other arts and also, I suspect, with the so-called "sciences of man," although the latter do so in a different way and for different reasons. In music, and in art in general, the experience produced by this "knowledge beyond craft," is emotional, even for the scholar, not to mention composer, performer, and audience. A French expression which, out of decency, I will refrain from translating —it was one of Stravinsky's catch phrases—is crudely illustrative: *"ça ne me fait pas bander."*

Knowledge based on emotion is always suspect for a scientist—even when he, himself, is moved by it—and so it must be. Scientific knowledge is transmitted through words and formulas, but is the word the only medium for communicating or transmitting knowledge? As a composer, it is not my job to address the questions of "word and thing," "word and way of knowing," and so on, so I will only say that excessive verbalization destroys entire areas of human expression, and music is undoubtedly an eloquent demonstration of this.

Earlier, I insinuated that artistic creation—and music as well, though there are hardly any remains of its earliest manifestations—began incommensurably earlier than science. Let this embarrassingly obvious observation serve, here, to emphasize that pre-scientific forms of "knowledge" were the only ones possessed by human beings for a very, very long time. Of course that is not a justification, but then, I am not trying to justify anything. I simply seek to point out realities of our human nature that it would not be proper to forget, let alone combat. Here, there are certainly "frontiers" but they are not those of a musician.

As I understand it, the type of "knowing"—perhaps a better term than "knowledge" when dealing with art—offered by music moves between the conscious and the unconscious: a constant back-and-forth that enriches both spheres. Perhaps that is one of its greatest charms, or even the reason why it is so absolutely necessary.

Let us, then, accept the possibility of a "knowing" of these characteristics, even if only hypothetically. Let us accept—and this is the true leap of faith—that it can be spoken and written about.

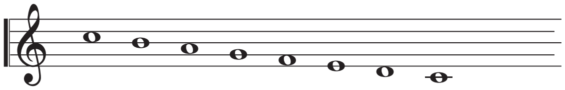
The questions that arise multiply like rabbits. How can we speak about an art that does not employ words, except when discussing its technique or craft? How can we calibrate or measure the emotion it produces? Is it possible to conceive of a repertoire of expressive means that can be used to deliberately provoke certain states of mind? And if such means exist, can they be cataloged? What can be done with the countless forms of expression that have been accumulated by different cultures and epochs? Are they interchangeable or, instead, mutually incomprehensible? How can a musical language be made collective? Would such a thing be desirable? What should we think of the old adage "music is a universal language?" And so on and so on… It is impossible to conceive of answering them in detail, and offering an overall answer is the same as avoiding them altogether, or simply lying.

As for me, I can only jot down a few comments and, if Erato—or Euterpe if you prefer, but not Melpomene, please!—inspires me, I can opine with prudence and discretion on this confusing labyrinth.

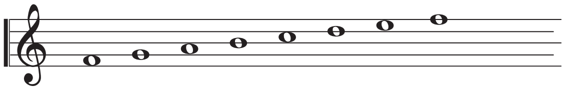
The aspiration of defining a repertory of musical means capable of provoking precise states of mind is as ancient as the first known documents. This would seem to indicate that it is actually as old as music itself, with or without documentation. It may not be entirely useless to glance at some chapters of its history—particularly the oldest ones—from the standpoint of what materials they employed. That excursion will reveal the successive "frontiers" that musical "knowledge" has experienced. I confess a certain reluctance to present this bird's eye view of music—no less—in what may be too personal a manner. On the other hand, much of what I will say is common knowledge—or so I believe. A thousand pardons, then, for doing so, but I see no other way to speak clearly, and with some meaning—usefully, in other words—about the delicate subject I am attempting to address.

The first documents—Hindu, Greek, Chinese, Tibetan, and many, many, more—are abundant. Almost all of them aspire to the same thing: either to awaken emotional states in the listener, or to serve as prayer. The musical means employed are highly varied.

The Greeks were convinced of the expressive value, and most of all, the ethical value of Music: valor, cowardice, faithfulness or softness could all be provoked, increased or diminished by it—one need only read the classics. There are abundant technical texts. Perhaps the most ample of these is by Aristoxenus of Tarentum (4th century BC), and I will limit myself to him here. The central musical idea was "monody." It consisted of notes based on the natural resonance of bodies: natural intervals of the fifth and its inversion. Monody was accompanied by the octave and, beginning in the fourth century, fourths and fifths as well, which corresponds to his concept of "scale" organized in tetrachords, that is, four notes together. The outermost notes of a tetrachord are "fixed notes," and the ones in between vary according to which kind of tetrachord it is: diatonic, chromatic, or enharmonic. There is no diapason—in other words, no absolute pitch, for that is exclusively Western and arrived much later on. The Greek word, *diapason* designates two successive tetrachords—that is, an octave. The notes that make up this double tetrachord already bear their name, indicating their position in the tuning of the lyre. There are seven types of octave, each with its name and mood. Those names are almost identical to the ecclesiastical modes of Christianity, but the actual scales do not correspond. For example the Greek *Lydian* mode is:



(from C to C), while the ecclesiastical *Lydian* is:



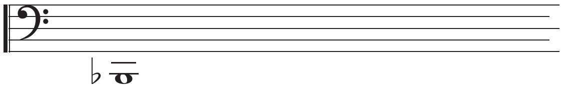
(from F to F).

All of this theory—which is highly detailed and would be bothersome to completely explain here—does not seem to correspond very well with practice, which was steeped in an oral tradition subject—as they all are—to constant and not especially foreseeable changes. This has undoubtedly kept us from knowing how the music of ancient Greece "really" sounded.

In the secular music of northern India, which is also very rich in theory, *ragas* (modes, more or less) and *talas* (rhythmic structures) are meticulously precise: the time of day or night, the season, mood, and so on. The material of *ragas,* that is, their pitches or intervals, is extracted from the 22 pitch-classes that make up the complete scale, of which the corresponding *raga* almost always uses seven. A *raga* needs a constant reference point in order to achieve the desired effect. Therefore, there is always an instrument—the *tampura* or *shruti*—whose function is similar to what would be called a "pedal" in Europe, that is, a continuous note. Naturally, there are numerous *ragas:* there is the *raga* (father), and the *ragini* (mother), which are mated to produce innumerable descendents: *putra* (sons) and *putri* (daughters). The *tala* use two *tablas*—hand and finger drums—and are equally subject to rigorous codification. Once the apprentice has mastered this complex mechanism, he is free to use it with imagination—Hindu compositional forms stimulate the capacity to improvise. The music of northern India surpasses that of ancient Greece in its effects: it can cure (or cause) illnesses, stop (or start) storms, and so on.

Vedic psalmody, the most archaic religious music from Northern India, and possibly the oldest known musical documentation, may be one of the most intricate manners imaginable for establishing contact with the divine. The Brahman's voice plays with the text in a thousand ways, all codified: the order of the phrases and syllables, changing accentuation, subtle changes of pitch, and so on. The result is incomprehensible to the uninitiated. Perhaps that is deliberate: these songs of praise—not prayers—must be accessible only to the superior caste.

To keep this introduction from being too long, I will only allude to the vocal techniques of Tibetan religious chanting—they chant on texts from Tantric Buddhism—which seek the same goal, that is, making the sacred text incomprehensible by projecting the male voice into an unusual register. This is done with very precise techniques and serves to keep "knowledge" considered "dangerous" out of the hands of the uninitiated.
Lower limit of the voice, around:



Without generalizing then, it is possible to say that what is known about ancient musical cultures indicates a belief that music has an unquestionable capacity to generate a great variety of emotional states that are determinable and determined, as well as the capacity to contact the divine (both can go together). It also shows that those powers and capacities have taken innumerable forms and have, in turn, constituted a first approach to discovering and using the nature of sound as a physical phenomenon. Behind all this there is a clear and shared necessity to use sound as an expressive vehicle, a guide or compass in a spiritualized universe. This necessity is as urgent as the need for food, sexual relations, and protection from the violence of nature. Music, in its origins—as we know them today—was more a matter of mystery than of knowledge. And it would be fairer to say that that knowledge, rigorously cataloged and learned, allows us access to the mystery of our existence. That is why I have used the word "knowing," which seems most apt to me. I will, however, continue to use the term "knowledge" in order not to stray too far from our subject, reserving the right to use "knowing" when I think it helps to clarify matters.

In the West, the Christian church used Greek modes in a peculiar way, through what they called *octoechos.* These were: *Dorian* or *protus* (d–d), *Phyrgian* or *deuterus* (e–e), *Lydian* or *tritus* (f–f), *mixolydian* or *tetrardus* (g–g), and the corresponding *plagal* modes a perfect fourth lower. The modes can be told apart by the location of the semitone. Each mode has two key notes: the *finalis* on which it ends, and the *repercussio* (or *corda di recita*) around which the melody is constructed. Those were the denominations—erroneous with respect to the original Greek modes—that predominated and later became the basis for polyphony and the studies of our harmony and counterpoint. Several modes were excluded because they were considered too upsetting for the dignity required of sacred service. Music served the sacred text and, unlike music from Northern India, Tibet, and so on, it had to be understood by the faithful. The delicate metric inflections of the "neuma" must have favored the text and its understanding. This rigidity did not, however, impede variety: Ambrosian, Mozarabe, and Gallican chant, among others, were eventually unified as Gregorian chant, following considerable struggle by known figures and the risks of excommunication.

Around the twelfth century, European polyphony was born in Northern France—every day, more people share the opinion that polyphony originated in Africa, but that is another story—and from that moment, the relation between music and text began to change. The words began to be incomprehensible, as they were being sung by more than one voice, but the expressive capacity of the music expanded enormously, threatening the liturgy's monopoly on music. The troubadours may have been the first to make deliberately secular music. In the eleventh century that is already true "accompanied melody." Later, we find it explicit in Dante ("Purgatory," Canto II), when the poet encounters a musician friend, Casella, who had put some of his poems to music when he was still alive. Dante asks him: "*…ti piaccia consolare alquanto / l'anima mia, che, con la mia persona / menendo qui, è affanta tanto!"* The musician sings Dante's poem, *Amor che nella mente mi ragiona,* whose music is now lost.

I regret only being able to mention the fascinating fact that music's independence from liturgy—that is, a secular text set to music with the accompaniment of an instrument—is not European. In fact, it came to Europe as a result of the Crusades on one hand, and the Muslim presence in Spain on the other. In both cases, its provenance was from ancient Persia—including the instruments—and it was brought by religious dissidents (the Cathars, etc.), with the glorious exception of the "Canticles," where there was no persecution at all, perhaps because of their explicitly sacred contents.

Music's awakening as an autonomous art—that is, free of liturgy, and even of text—was relatively late in the West, even though, thanks to its subtle polyphonic techniques, it was always considered the most emotive of the arts, essential to, and inseparable from, both poetry and religious uses.

This interdependence of text and music, both religious and secular, is not the only one that has characterized our art. Music has had many adventures in other cultures, and I would like to mention a few to demonstrate its ductility and its volatile—that is, protean—character. Here are two examples:

In *Genji* (Heian dynasty, tenth century) Lady Murasaki (Murasaki Shikibu) tells us how there was an entire repertoire of outdoor pieces for the *fue* transverse flute. This repertoire varied with the seasons, taking into account the natural sonic setting brought about by each one. This music is still played and I had the good fortune to hear the admirable Suiho Tosha perform it in Kyoto. Obviously, this idea is rooted in the Shinto religion.

There are other musical forms that are not satisfied to be played outdoors; they aspire to a meticulous description of certain aspects of reality: rain, a spider balancing on her web, a child crying, a young man attempting to sneak into the "house of women," and so on. I am referring to the traditional music of the Aré-aré (Hugo Zemp, "Melanesian Pan Flutes," Musée de l'Homme 1971) with its pan flute ensembles (*au tahana* and *au paina*).

This is a precious example of how a descriptive music, understood as such by its creators, signifies absolutely nothing to those outside its culture, except as an exquisitely beautiful object lacking any meaning. And we find this divorce of aesthetics and original meaning innumerable times.

I would not want these divagations—which I do not consider as such—to be mistaken for useless erudition. With them, I am attempting to show, on the one hand, that it is impossible to find the "frontiers" of music—other than purely physical ones—and on the other, that the word "knowledge" suits neither its nature nor the effects it produces on listeners. Its reality is too varied—not contradictory—for anyone to seek a unity that could only mutilate it. And that is enough about this matter.

When instrumental music was fully established in Europe (more or less in the fifteenth century), the rules of composition were defined with exquisite care. At the same time, new genres were being invented that favored emotional expression over everything else.

From then on, something uniquely European began to take shape in the West, so much so that, for many enthusiasts, that "something" is synonymous with music. As you will have guessed, I am referring to harmony, which replaced "modes." Harmony established the precise functions of intervals, measuring their capacity for movement or repose. The origins of this movement are in the polyphony that preceded it, but now it serves tonality and its fluctuations—what are called "modulations." Harmony has been compared to perspective, though the latter concerns space, while the former concerns time. Undoubtedly, it is the most specifically European musical technique ever created by the West, although it meant abandoning a great deal of popular music, which was frequently based on the old modes.

As harmony adapted to the rules of the former counterpoint—and vice versa—a series of procedures emerged that became the very core of European compositional teaching. One of the milestones in this teaching was Johann Joseph Fux's *Gradus ad Parnassum* (1725). For a very long time indeed, almost every later composer was respectful of the old master, even if they took certain liberties. Just a few years ago, György Ligeti (1978) told Péter Várnai (*Ligeti in conversation,* Eulenburg, London, 1983) that he asked his advanced composition students to be "familiarized" with that venerable text, after which they could do whatever they wanted. That piece of advice, I might add, ran throughout the nineteen century and, as we have seen, much of the twentieth as well. One might ask why, and the answer has to do with the "craft," I have so often mentioned. With *Fux* in hand, the solidity of the building is assured, though it offers no guarantee that the building will be interesting. But neither Fux, nor any other theoretician, has to address that issue: it is a matter for the artist… who learns discipline from him, even when that artist is undisciplined by nature, as was the young Debussy, for example. To put it clearly, once and for all: *Fux* and his innumerable offspring were, and may still be, essential for their role as a negative pole.

The attacks began only about a hundred years after it was published: late Beethoven, Romanticism, Wagnerian opera, French harmonic, and timbral subtleties, the nationalists—especially the Russians—the first central European atonalities (though not what followed them, as we will see below), as well as—from a different standpoint—the Universal Exhibitions, which brought colonial art to the metropolises, and so on. All of this produced an unstoppable fermentation that didn't eliminate it, but did change its meaning. Instead of a body of essential norms, Fux became a reference point for our old identity—something worth knowing if we want to avoid any missteps. Quite simply, it became a source of practical "knowledge" (which is no small matter).

The last hundred years or so of music's history is already present in what I have just written. The structures of classical harmony—and its formal consequences—collapse, and the causes are both internal—its developmental dynamics—and external —agents from outside.

Let us consider each for a moment. The internal cause stems from the same cultural setting as the model: the Germanic world. Naturally, I am referring to the School of Vienna and its "trinity" of Arnold Schoenberg, Anton Webern, and Alban Berg. Their primary objective was to dissolve tonality into what Schoenberg called "pantonality"—also known as "atonality," a term he considered inexact. From

1906–1909 (with his *Kammersymphonie,* op. 9, or even his Second Quartet, op. 7) to 1921, when he defined his compositional system (his "technique for composing with the twelve tones" generally called "serialism" in English), the three Viennese composers made an enormous number of technical innovations. But Schoenberg was also an extraordinary teacher, with an absolutely foolproof classical training, despite the fact that he was practically self-taught. His invention of twelve-tone technique was a way of prolonging classical techniques in a new context. His *Pierrot Lunaire* (1912), which precedes serial technique, is a fine example of how he was already combining the most risky new techniques with the oldest ones. The School of Vienna had a very clear "mission" that consisted of two complementary facets: first, using new means to make advances in musical expression; and second, recovering and resuscitating classical procedures from the great Germanic school, using up-to-date contents. Thus, we see that their serial, and even pre-serial, compositions return to severe counterpoint, classical forms, and so on, in order to avoid a break with the past, prolonging what they considered German-Austrian music's supremacy over all others. Despite Nazism, which prohibited their music as "degenerate," "Bolshevik," and so on, and forced Schoenberg into exile; none of them renounced the idea of considering themselves the only true bearers of the central-European musical heritage. The School of Vienna's attitude was of a very high ethical quality, colored by a rather ingenuous nationalism, clearly not very folkloric, but absolutely demanding of itself. Very much in the sense of Beethoven, their music was their moral—intransigently so—in the face of a hostile world. Sad to say, Karl Popper's aggressive reflections on the music of the School of Vienna are a model of incomprehension and arrogance.

Among the external causes, none was as powerful as the knowledge of other cultures and the changing attitude toward them. The disdainful attacks of Berlioz, Gianbattista Vico, and many others, at first gave way to curiosity and later enthusiasm. It may be difficult to find a European composer of a certain stature from the last seventy years—though not the German-speaking ones, as clarified above, except for Stockhausen—who has not been influenced by music outside our traditional Western teachings. It is essential to reflect on this fact.

1. These musical forms have been absorbed by the West. Some non-Western countries—Japan, Korea, Morocco, China, and so on—have undoubtedly assimilated certain aspects of Western music. But composers from those countries have done so from the perspective of our tradition. Let me explain: Toru Takemitsu, for example, uses Japanese instruments, scales, and even "time sense." But the results of his work are part of the Western symphonic tradition. His music is not included in the Japanese tradition; it is not a part of *gagaku, kabuki,* or the repertoire of the beggar monks with their *shakuhachi*—the end-blown flute he uses in his works. Allow me to furnish an illustrative anecdote: when Toru Takemitsu began studying music in Tokyo, his attitude was a total rejection of the Japanese tradition—those were the post-war years. He moved to Paris to broaden his studies, and it was there, in the sixties, that he discovered the musical tradition of his own country and his conversion took place. This is a vivid anecdote, and something very similar happened with Ahmed Essyad and Morocco… and with leading current composers from Korea, China, and so on.

2. Taken one by one, these musical forms cannot be translated into each other. For example, it is difficult to imagine how the vocal polyphony of the *Aka* Pigmies of Central Africa could be enriched by listening to and studying the drums that accompany the action of *Kathakali* theater from Kerala (southern India), or vice versa. As I say, these traditions are untranslatable.

On the other hand, this is not the case for Western music. And we can see and hear this every day in both theater and music. The reason is clear: Westerners have abandoned the extramusical contents that, like all others, those non-Western musical forms include. This should come as no surprise, given that we have done the same thing with our own music. Consider the ever-more-frequent use of medieval techniques in current music. And even more so, consider the neoclassical music of the nineteen twenties—the towering figure of Stravinsky—and you will see that this in not a new phenomenon in the West. This should not, however, be seen as a criticism on my part, but rather as an effort to understand the direction of our music's evolution.

3. Our "consumer music," with its exceptional means of distribution—which makes it an indicator of our wellbeing and power—is invading the planet and may well be the only part of our musical history known by all peoples. Moreover, the mutual incompatibility between musical forms mentioned above also occurs in the West, between our so-called "classical" or "straight music"—horrid labels—and our "consumer music." It is enough to know what the greater part of Western youth considers "its" music, "its" means of expression. Allow me to eschew considerations as to the provenance, possible consequences, and musical characteristics of this fact (at the beginning of this text I asked permission to do so).

In the early twentieth century, at the onset of this positive encounter between our musical tradition and music from other cultures, something else began that is only, or almost only, known by specialists, although in my opinion it is very interesting because of its direct relation to the interpretation of music and its frontiers.

In Germany and France, musical research centers were founded that drew on the possibilities offered by the advent of recording techniques—rudimentary at the time, but now efficient—and comparatively safe travel to remote areas. The object of those centers was to study music as a global phenomenon and to draw conclusions, where possible. In Berlin (1919), Curt Sachs founded the Institute of Comparative Musicology. In Paris, André Schaeffner created the Department of Ethnomusicology at the Museum of Man (1929). In Barcelona, the exiled musicologist, Marius Schneider, did the same in 1944. Eleven years later, he was to found the School of Cologne for the same purpose. In my opinion, those are the three names that best define this search, this effort to find a "general part" of music, collecting an enormous quantity of data—before computers even existed—that was not treated merely as isolated facts; contexts were taken into account and common links or significant contrasts between different materials were sought out. Some of the questions addressed include: what does the different use of the voice in different epochs and cultures signify? How has the cadential value of the descending perfect fourth been used? How do isochronies and heterochronies compare? And so on. Some researchers went even farther. Marius Schneider tried to draw links between, on one the hand, rhythms, intervals, cadences, and so on, and on the other, signs of the zodiac, animals, constellations, iconography, etc. André Schaeffner did the same with musical instruments, studying their origins, evolution, symbols, tuning, social significance—religious or otherwise—and so on.

This group of musicians has carried out a very worthy job of conserving, studying, and publicizing music that is rapidly disappearing. They have struggled to create a "science of Music" that aspires to give global meaning to our musical history. And they are not alone in their aspirations. A fervent *Sufi* reached the same conclusion from the perspective of his tradition. That perspective has not kept Karlheinz Stockhausen and numerous *rock* musicians from devotedly studying the writings of Hazrat Inayat Kahn, who was born in the early twentieth century in what is now Pakistan. His underlying idea is the same—music is the only vehicle for wisdom—but his starting point is not. European musicologists sought to establish a scientific basis—rather *sui generis,* perhaps—for their work. But Kahn calls simply for common religious faith. And we had best avoid the world of Pythagoras, on one hand, and ancient China, on the other…

The procedures and results of this admirable group of musicologists have frequently been branded as "esoteric" or "magic," and even dilettante. I think they have been attacked because their work on interpreting the phenomena of culture through music has been disconcerting to many, due to its unusual nature. To me, as a musician, they seem neither more nor less magic than Roman Jakobson's approach to linguistics or Claude Lévi-Strauss's work on family structures. Are these mere questions of affinities? No. Any theory that seeks to be holistic will overlook certain things, and may even be ridiculous. Let the guiltless party throw the first stone…

The explosion of science in the nineteenth century had important consequences for music, its meaning, and, most of all, knowledge of its materiality as sound. One of many examples is the contribution of the German physicist, Hermann von Helmholz (d. 1894), which was extremely important in stimulating the imagination of many composers more than fifty years after he published his discoveries. After analyzing hearing and studying timbres and pitches, he hypothesized that they are all related and that there could conceivably be a music in which this relationship had a function. Moreover, he set the bases for psychoacoustics, which is a landmark for composers in their treatment of the materials they use. It does not take much insight to realize that one of the origins of Stockhausen's music—French Spectral composition—and its countless consequences lies in the work of Helmholz and his followers (many of whom do not even realize they are his followers).

Paralleling this scientific research, musicians themselves needed to enrich their expressive media. A simple enumeration of the inventions they thought up and the results they achieved would already take up too much space here, but I will mention some of the most relevant. The first consists of the thousand transformations of musical instrumentation, with an emphasis on percussion. This led to an endless number of masterpieces, too long to list here, that are already a part of common compositional practice.

We cannot overlook the *intonarumori* of the Italian Futurists (Luigi Russolo, 1913), which were little more than curiosities, but indicate a renovating stance.

The arrival of technology generated an authentic revolution. First, in Paris, the GRM (the Musical Research Group led by Pierre Schaeffer [1910—1995]) and Pierre Henry (1922-) ushered in *musique concrete* during the second half of the forties. Initially considered an "other art" halfway between cinema, radio reporting, and music, it finally became a part of the latter. Its raw materials were real sounds/noises, recorded and treated using electro-acoustic devices: recorders, filters, etc. The GRM even invented one device, the *phonogène,* which is now a museum piece, but played a key part in its day. Soon thereafter, in 1951, Herbert Eimert (1897–

1972) and Karlheinz Stockhausen (1928–2007) founded the Electronic Music Studio at Radio Cologne, using oscillators, filters, and many other devices.

Together, *musique concrete* and *elektronische Musik* merged to form "electronic" or "electro-acoustic music," which rapidly spread throughout the technologically developed world: from Stockholm to Milan, from Lisbon to Warsaw, Montreal to Buenos Aires, Tokyo to New York, Sidney to Johannesburg, and so on. Its technical advances were extremely rapid and what had begun as a craft quickly became a world of discoveries in constant evolution. Masterpieces soon followed. The best known —and rightly so—is Karlheinz Stockhausen's *Gesang der Jünglinge* ("Song of the Adolescents," 1955, Cologne).

In 1960, the sound engineer, Robert Moog, designed the first Moog Synthesizer in Buffalo, New York. This flexible device made it possible to use a synthesizer "live," that is, like any other instrument. It could be connected in multiple ways, including to a computer, and it was so simple to control that it made electro-acoustic media available to an unlimited public. Suddenly, what had been considered the cutting edge of the avant-garde grew vulgar, falling directly into the hands of "consumer music." Free electro-acoustic creation was still possible, but grew increasingly difficult.

In 1955, at the University of Urbana (Illinois), Lejaren Hiller (New York, 1922) programmed a computer, the ILLIAC IV, to reconstruct traditional musical structures: harmony, form, and so on. The result was the *Illiac Suite* for string quartet. Hiller privately confessed that he was not trying to make music, but rather to demonstrate the unknown capacities of the machine. Shortly thereafter, though, the digital-analog converter was designed, allowing the machine to make sounds. Around 1965, John Chowning at Stanford University (California) designed a computer with frequency modulation, allowing the machine to broaden its possibilities to include any timbre—instrumental or invented—meter, number of voices, and so on.

And here it is prudent for me to stop, because what follows—and we are in the midst of it right now—is the computer invasion, which gives anyone access not only to any sound material, but also to anything related to music: publishing, listening, mixing, combinatoriality, meter of any desired degree of complexity, and so on. A "thing" that reminds us of music—because of its aural content—can be generated today by anyone who can handle a machine with a certain degree of skill. I am not speaking of a "work" but simply of the possibility of making "something." Nevertheless, that possibility does not seem to be of much help to the layman when trying to achieve what, with all due respect, could be called "quality."

Obviously, the machine has made it possible to do things that could only be dreamed of before: transforming an instrumental or vocal sound at the very moment of its production or doubling it; the interaction of pitches at different listening levels; results never heard before—literally, as there had never been any way to produce such sounds—either totally unknown, or of known origin, but unrecognizable; transformation of any constituent element of sound a varying speeds, and so on. The list is endless. Such was the enthusiasm for the "new music" that in the fifties, an illustrious composer—I've forgotten his name—stated, with moving forcefulness and faith, that "in ten or twenty years, no one will listen to the music of the past anymore." This never came to pass. On the contrary, "classical" music from the world over—from Bach to Coptic liturgy—is being heard more than ever, so the *facultas eligendi* is safely installed. I believe that in time the use of computers for music will lead to something else: the presence of a medium so powerful doesn't seem likely to impede the existence and development of music whose means of expression are more traditional. As always, the most probable outcome will be an unpredictable mixture.

This "bird's eye view" of music, though hurried and necessarily partial, has, I believe, been sufficient for the job at hand. One thing is clear: the limits of musical knowledge—in the sense of what can be learned—has expanded so much that is difficult to decide what is necessary and what is accessory. We should also point out something important, though rarely mentioned: techniques—in plural—of musical composition have proliferated to an unbelievable degree in the last 40 years, and I am speaking only of the West. Beginning with an illusory unity: the most restrictive version of the *serialism* that was the son or grandson of the School of Vienna—cultivated by the so-called Darmstadt School (Germany) and its courses founded by Dr. Wolfgang Steinecke in 1946. This severe technique simply self-destructed in the late fifties. Since then, it would be no exaggeration at all to say that there are as many compositional techniques as there are significant composers.

I will not attempt to explain why things exploded in the sixties—that would be a different article altogether. Moreover, that was a decade filled with revolutionary events of every sort. Those storms have calmed and in the last twenty years the waters seem to be running more smoothly. But this was not due to any *return to order*—which would have been as pretentious and extemporaneous as Jean Cocteau's—but rather to a healthy *systole* after the Pantagruelian *diastole.*

I mentioned above that Western composers of recent years have done away with the extramusical content of music. That is, music does not "represent" anything except itself. Is that true? In my opinion, yes. And I would have to add that it has always been true, especially in the most illustrious cases, where music has the most power to move us. Thus, when non-European musical forms began to be appreciated and assimilated by Westerners, they were not appreciated because of their non-musical contents, but for the beauty and interest of their sonic contents: our ears were already prepared for them.

Our composers have always known this—how could they not know, or at least intuit it?—even when they wrote descriptive music. Sixteenth- and seventeenth-century madrigals and opera are paradigms of what I say. The word stimulates the composer's imagination, but the music born of that stimulus is not a translation of those words. That would be both impossible and frivolous. The past, perhaps wisely, did not argue about this at all. It was clear that an adequate sonic order was expressive per se when done with imagination, freshness, and mastery (I am deliberately ignoring the untimely arguments about *prima la musica, dopo le parole*—or vice versa. They did not affect composers).

In our time, such arguments have arisen, however. Some composers thought music had to renounce any sort of expressivity, including its own (remember the rather comic indignation of Franco Donatoni at the emotive power of *Senza Mamma* from Puccini's *Suor Angelica*). That position was also taken by young Boulez in his *Structures I* for two pianos (1952), though certainly not in his following work, *Le Marteau san maître.* But in my opinion, even in the line of composition in which the author does not, in principle, take any interest in the expressive dimension of a sonic order, if that order is successful (in terms of originality, perfection, and depth) it will transmit a sort of emotion to us. Perhaps this is unwanted, but it is inherent to the material and, of course, impossible to put into words. Maybe the error of *Structures I* (Boulez always considered them an experiment) is that the sound is ordered in a purely numerical-combinatorial—rather than musical—manner, which makes the results unintelligible. With all due respect, certain works by Xenakis have the same problem.

Moreover, it is more than probable that the blossoming of musical techniques mentioned above has its origins in the emphasis that recent—and not so recent—composers have put on *pure* musical material as the main axis of creative impulse (Debussy must be turning in his grave…).

And so, we return to our starting point: what can be known with, or through, music? What does "Frontiers of Knowledge" mean when speaking of music? The "Frontiers of [Musical] Knowledge" would be those that define the capacity of scholars and creators on one the hand, and the receptive capacity of listeners on the other. In the latter case, the idea of knowledge and its frontiers will be identical to the capacity to experience deep aesthetic feelings, which enrich us personally, though rarely collectively (watch out for mass emotions!), despite the apocalyptic halls in which music that was intended to be heard by no more than ten people is so often listened to by crowds of ten thousand…

It remains to be seen whether this "emotion" could be a form of knowledge. If it is—and I believe it is: see what I wrote at the beginning of this chapter—it belongs to a different sort of cognizance. It does not seek objective, verifiable truth quite simply because that kind of truth does not exist in art. And I dare say—forgive me for talking on so much—that it is knowledge that stems from life experience, not from studying. As I said, studying—and its delights, for it has them—are for professionals. The lived, enjoyed, and instructive experience of sensitivity—without excluding scholars, needless to say!—is mostly for others: those to whom the artist offers his work. And with the passing years—almost always, many of them—a musician's contribution becomes part of a collective identity, part of that collective's "knowledge." The collective recognizes itself in that contribution, and that is the "truth" of art and music. Whether that truth is valuable or simply mediocre and trivial will depend on the collective's education. Moreover, in a healthily pluralistic society there are always many collectives. What is more, collectives united in shared recognition do not have to coincide with state, religious, linguistic, or any other sort of frontiers. They are linked by a shared emotion, a shared "knowing" of emotions.

I am finishing, and I must do so with a question. Nowadays, the frontier of this knowledge (recognition, knowing) is undergoing sudden, unforeseeable, and possibly uncontrolled changes. Musicians—most of all, composers—live in a constant short-circuit (as I already insinuated). This is not comfortable, but there is not time to get bored. Thus, I point it out without fear… but with some uneasiness.

biographies

JANET ABBATE is an assistant professor in the department of Science and Technology in Society at Virginia Tech. She holds a B.A. in History and Literature from Harvard-Radcliffe and a Ph.D. in American Civilization from the University of Pennsylvania. She is the author of *Inventing the Internet* (MIT Press, 1999) and numerous articles on the history of the Internet. As a Research Associate with the Information Infrastructure Project at the Kennedy School of Government at Harvard, she coedited (with Brian Kahin) the volume *Standards Policy for Information Infrastructure* (MIT Press, 1995). Other publications include "Privatizing the Internet: Competing Visions and Chaotic Events, 1987-1995" (*Annals of the History of Computing*, forthcoming); "Women and Gender in the History of Computing" (*Annals of the History of Computing*, 2003); "Computer Networks" (in Atsushi Akera and Frederik Nebeker, eds., *From 0 to 1: An Authoritative History of Modern Computing*, Oxford, 2002); and "Government, Business, and the Making of the Internet" (*Business History Review,* Harvard, 2001). Her current research focuses on women's professional identities as programmers and computer scientists in the US and Britain since the World War II.

SERGIO ALONSO OROZA is Senior Professor of Meteorology at the Universitat de les Illes Balears and Academician at the Royal Academy of Sciences and Arts of Barcelona. He received his undergraduate degree and his Doctor of Physics degree from the Universitat de Barcelona, where he was Vice-Rector from 2003 to 2006. In 2007 he became President of the Science Commission for National Accreditation of University Professors. His research focuses mainly on Mediterranean meteorology and climate, with publications in the most prestigious reviews of Meteorology, Oceanography, and Climate. He is a member of various scientific associations, including the Royal Spanish Society of Physics, President of the Specialized Group on Physics of the Atmosphere and Ocean; board member of the Royal Meteorological Society, the American Meteorological Society and the American Geophysical Union. Among his other activities, he is author and manager of the National Climate Program of the National Research and Development Plan, member of the Spanish delegation to the United Nations Framework Program on Climate Change, member of the National Council on Climate and its Permanent Commission, president of the Meteorology and Atmospheric Sciences Section of the Spanish Geodesic and Geophysics Commission, member of the Intergovernmental Panel on Climate Change, and of its Executive Commission, as well as one of its reviewers.

JESÚS AVILA is Research Professor at the CSIC in the Severo Ochoa Molecular Biology Center at the Autonomous University of Madrid, of which we was previously the director. He received his Doctorate in Chemical Sciences from the Complutense University of Madrid and is an Academician of the Royal Academy of the Exact, Physical and Natural Sciences, and of the European Academy. His work focuses on the study of the function of microtubular proteins in the determination of neuronal shape, the function of the tau protein in neurodegenerative processes and the search for axonal regeneration processes. He has been President, and is a member of the Spanish Society for Biochemistry, Cellular Biology, and Molecular Biology, a project evaluator for the National Science Foundation (USA), the Swiss National Foundation, NFR (Switzerland), and the Welcome Trust grants (United Kingdom). He has received numerous prizes, including those of the Royal Academy of Exact, Physical and Natural Sciences, the Medal of the University of Helsinki, the Prize of the Carmen and Severo Ochoa Foundation, the Medal of the Community of Madrid, the Lilly Foundation's Prize for Preclinical Biomedical Research, and in 2004 the Ramón y Cajal National Research Prize.

ABHIJIT VINAYAK BANERJEE was educated in the University of Calcutta, Jawaharlal Nehru University, and Harvard University, where he received his Ph.D. in 1988. He is currently the Ford Foundation International Professor of Economics at the Massachusetts Institute of Technology. In 2003 he founded the Abdul Latif Jameel Poverty Action Lab, along with Esther Duflo and Sendhil Mullainathan, and remains one of its directors. He is a past president of the Bureau for the Research in the Economic Analysis of Development, a Research Associate of the NBER, a CEPR Research Fellow, International Research Fellow of the Kiel Institute, a fellow of the American Academy of Arts and Sciences and the Econometric Society, and has been a Guggenheim Fellow and an Alfred. P. Sloan Fellow. His areas of research are development economics and economic theory. He is the author of two books and the editor of a third, as well as a large number of articles. He finished his first documentary film, *The name of the disease*, in 2006.

FRANCISCO CALVO SERRALLER is a Doctor of Philosophy and Letters from the Complutense University of Madrid, where he has been Senior Professor of Contemporary Art History since 1989. He has been an Academician at the Royal Academy of Fine Arts of San Fernando since 2001. In 1988, he created the Contemporary Art Database, ARCO-DATA, which he directed until 1994. He was director of the Museo del Prado. Dr. Calvo combines his research and teaching activity with work as an art critic in various press media, including the newspaper, *El País,* with which he has been collaborating since it was founded in 1976. He has directed and participated in a multitude of courses, congresses and international seminars on art. He has been part of scientific committees for important exhibitions and numerous international contests and has curated important shows in leading museums and art centers in Europe and America. Scientific advisor to various institutions, founding member and board member of the Friends of the Museo del Prado Foundation, and vocal member of the Board of Directors of the IVAM. His research focuses on the study of sources of the history of modern and contemporary art, on historiography, and artistic methodology of the contemporary era and on contemporary art history.

PAUL E. CERUZZI is Curator of Aerospace Electronics and Computing at the Smithsonian's National Air and Space Museum in Washington, D.C. Dr. Ceruzzi received a B.A. from Yale University and Ph.D. from the University of Kansas, both in American Studies. Before joining the National Air and Space Museum, he was a Fulbright scholar in Hamburg, Germany, and taught History of Technology at Clemson University in Clemson, South Carolina. He is the co/author of several books on the history of computing and aerospace technology: *Reckoners: The Prehistory of The Digital Computer* (1983); *Beyond the Limits: Flight Enters the Computer Age* (1989); *Smithsonian Landmarks in the History of Digital Computing* (1994); *A History of Modern Computing* (1998); and *Internet Alley: High Technology in Tysons Corner* (2008). His current research and exhibition work concerns the use of computers for long-range space missions. Dr. Ceruzzi has curated or assisted in the mounting of several exhibitions at NASM, including "Beyond the Limits: Flight Enters the Computer Age," "The Global Positioning System: A New Constellation," "Space Race," "How Things Fly," and the James McDonnell Space Hangar of the Museum's Steven F. Udvar-Hazy Center, at Dulles Airport. He is currently working on a new exhibit on space navigation, scheduled to open at the National Air and Space Museum in 2010.

CARLOS M. DUARTE QUESADA is a Research Professor at the Superior Council of Scientific Research (CSIC) in the Mediterranean Institute of Advanced Studies (IMEDEA). He completed undergraduate studies in Environmental Biology at the Autonomous University of Madrid, then, after two years of research in Portugal, he attained his Doctorate in the ecology of lake macrophytes at MacGill University in Montreal. Following a brief postdoctoral stay at the University of Florida, he began research on marine ecosystems at the Institute of Oceanographic Sciences (CSIC), worked at the Center for Advanced Studies of Blanes (CSIC), and finally, at the Mediterranean Institute for Advanced Studies (CSIC-Univ. Illes Balears). His research spans a broad spectrum of aquatic ecosystems and he has published over 300 scientific articles in international reviews, and a dozen chapters in diverse texts, and two books. He is President of the American Society of Limnology and Oceanography, editor-in-chief of the review, *Estuaries and Coasts*, and a member of the Scientific Board of Advisors to the CSIC and of the European Academy. In 2001, he received the G. Evelyn Hutchinson Medal for Scientific Excellence from the American Society of Limnology and Oceanography, and in 2007, the Alejandro Malaspina National Research Prize.

JOAN ESTEBAN has been member of the Institute of Economic Analysis of the CSIC since its beginning in 1985, and was its director from 1989 to 1991 and from 2001 to 2006. He is the current President of the Association for the Study of Economic Inequality, and member of the executive council of the International Economic Association (2005-2011). His research has been published in the *American Economic Review, American Political Science Review, Econometrica, Economics of Governance, Economics Letters, European Economic Review, International Economic Review, Journal of Economic Behavior and Organization, Journal of Economic Theory, Journal of Income Inequality, Journal of Peace Research, Regional Science and Urban Economics, Social Choice and Welfare*, and *Theory and Decision*. At present, his main field of interest is the study of conflict and polarization. This topic covers many aspects ranging from individual and group behavior to the role of political mechanisms in preventing open conflict. Other topics of interest are public finance, inequality, OLG models, behavior under uncertainty, and regional economics.

LUIS FERNÁNDEZ-GALIANO is an architect and Senior Professor of Projects at the School of Architecture at the Polytechnical University of Madrid, as well as director of the review, *AV/Arquitectura Viva*. Member of the Royal Academy of Doctors, he has been Cullinan Professor at Rice University, Franke Fellow at Yale University, visiting researcher at the Getty Center in Los Angeles and visiting critic at Harvard and Princeton, as well as at the Berlage Institute. He has directed courses at the Menéndez-Pelayo and Complutense Universities. President of the jury at the 9th Venice Biennale of Architecture and the XV Chilean Biennale of Architecture, expert and jury-member at the European Mies van der Rohe Prize, he has curated the shows, "El espacio privado" in Madrid and "Eurasia Extreme" in Tokyo and has been part of the jury for numerous international contests. His books include *La quimera moderna, El fuego y la memoria, Spain Builds* (in collaboration with MoMA) and *Atlas, arquitectura global circa 2000* (with the BBVA Foundation).

JOHN B. HEYWOOD has been a faculty member at MIT since 1968, where he is Sun Jae Professor of Mechanical Engineering and Director of the Sloan Automotive Laboratory. His interests are focused on internal combustion engines, their fuels, and broader studies of future transportation technology, fuel supply options, and air pollutant and greenhouse gas emissions. He has published over 200 papers in the technical literature, and is the author of several books including a major text and professional reference "Internal Combustion Engine Fundamentals." He is a Fellow of the Society of Automotive Engineers. He has received many awards for his work including the 1996 US Department of Transportation Award for the Advancement of Motor Vehicle Research and Development, and the Society of Automotive Engineers 2008 Award for his contributions to Automotive Policy. He is a member of the National Academy of Engineering and a Fellow of the American Academy of Arts and Sciences. He has a Ph.D. from MIT, a D.Sc. from Cambridge University, and honorary degrees from Chalmers University of Technology, Sweden, and City University, London.

GERALD HOLTON is Research Professor of Physics and of History of Science at Harvard University, a member of the American Physical Society, the American Philosophical Society, the American Academy of Arts and Sciences, and several European learned societies. He served as President of the History of Science Society and on several US National Commissions. His book publications include *Thematic Origins of Scientific Thought*; *Einstein, History and Other Passions*; *Science and Anti-Science*; and *Victory and Vexation in Science*. He was the founding editor of the quarterly journal *Daedalus* and member of the Editorial Committee of *Collected Papers of Albert Einstein*. His honors include the Sarton Medal, the Abraham Pais Prize of the American Physical Society, selection as the Herbert Spencer Lecturer at Oxford University, as the Jefferson Lecturer by the National Endowment for the Humanities, and the Ehrenkreuz 1.Klasse by the Republic of Austria.

CAYETANO LÓPEZ studied Physics at the Complutense University of Madrid and at the University of Paris VII, receiving his Doctorate at the Autonomous University of Madrid, where he has been Senior Professor of Theoretical Physics since 1983. He is Doctor Honoris Causa from the University of Buenos Aires. He has worked at CNRS (Centre National de la Recherche Scientifique) and at CERN (Centre Européen de la Recherche Nucléaire) in fields related to unified theories of elemental particles and nuclear transmutation for energy production and the elimination of radioactive residues. As well as numerous articles in specialized reviews and contributions on the spread of scientific knowledge, science, and energy policies in newspapers and philosophy magazines, he is the author of two books for the general public, *El ogro rehabilitado* (El País-Aguilar, 1995) and *Universo sin fin* (Taurus, 1999), as well as the textbook, *Física de los procesos biológicos* (Ariel, 2004). He has been Rector at the Autonomous University of Madrid, Vice-President of the Board of CERN and Director of the Scientific Park of Madrid. Since September 2004, he is Adjunct General Director of CIEMAT (Center for Energy, Environmental and Technological Research), where he directs the Energy Department.

JOAN MASSAGUÉ received a Ph.D. degree in Biochemistry and Pharmacy from the University of Barcelona in 1978. He joined the University of Massachusetts Medical School as a faculty member in 1982 and became Program Chairman at the Memorial Sloan-Kettering Cancer Center in 1989. He is an Investigator of the Howard Hughes Medical Institute, and the Adjunct Director of the Barcelona Institute for Research in Biomedicine. Dr. Massagué's work has revealed how external signals block mammalian cell division. These mechanisms are now known to be crucial in embryonic development, and their disruption causes congenital disorders and cancer. Building on his ability to dissect complex biological processes, Dr. Massagué has recently identified genes and functions that mediate tumor metastasis, which is the main cause of death in from cancer. Ranked among the 50 top cited scientists of the past two decades in all fields of science, Dr. Massagué is a member of the National Academy of Sciences, the Institute of Medicine, the American Academy of Arts and Sciences, the Spanish Royal Academies of Medicine and of Pharmacy, and the European Molecular Biology Organization. His honors include the Prince of Asturias Prize, the Vilcek Prize, and the Passano Prize.

JOSÉ M. MATO is General Director of CIC bioGUNE and CIC biomaGUNE in the Basque Country since 2003. He received his doctorate from the University of Leiden and before settling in the Basque Country, he was a researcher at the Jiménez Díaz Foundation in Madrid, Research Professor at the Superior Council for Scientific Research (CESIC), President of the CESIC and Senior Professor at the University of Navarre. He has been guest researcher at the National Institute of Health, at the University of North Carolina, Chapel Hill, at Pennsylvania University and Thomas Jefferson University in Philadelphia. He has received numerous prizes, including the Kok prize, the Morgagni Medal, the Lennox K. Black prize and, in 2004, the Gregorio Marañón National Prize for Medical Research, which is awarded by the Spanish State. Professor Mato's work focuses on the study of fatty-liver disease, liver regeneration, and liver cancer. His laboratory has pioneered in the identification of S-adenosylmethionine as a fundamental regulator of metabolism and hepatic proliferation, as well as the creation of new animal models for the study of liver diseases.

ROBERT McGINN is Professor of Management Science and Engineering and Director of the Science, Technology, and Society (STS) Program at Stanford University. He received a B.S. in Unified Science and Engineering at Stevens Institute of Technology, an M.S. in mathematics at Stanford, and a Ph.D. in philosophy and the history of ideas at Stanford. Apart from a year at Bell Laboratories in 1978–79, McGinn has been at Stanford since 1971. His academic specialties are technology and society; ethics, science, and technology; engineering ethics; and ethics and public policy. McGinn's publications include *Science, Technology, and Society* (1991) and articles in scholarly journals such as *Professional Ethics*; *Technology and Culture*; *Science, Technology, and Human Values; Science and Engineering Ethics*; and *Nanoethics*. From 2004–07, he conducted a large-scale study of the views about nanotechnology-related ethical issues held by researchers in the US National Nanotechnology Infrastructure Network. Recipient of several awards for teaching excellence at Stanford, McGinn has received research grants from the Mellon Foundation, the Marshall Fund, and the National Endowment for the Humanities. He received the Dinkelspiel Award for distinguished contributions to undergraduate education at Stanford.

GINÉS MORATA completed undergraduate studies at the Complutense University of Madrid in 1968, receiving his doctorate at the same university in 1973. Between 1973 and 1977, he did research at the English universities of Oxford and Cambridge. In 1977, he joined the Molecular Biology Center of the Superior Council for Scientific Research and the Autonomous University of Madrid, where he continues work as Research Professor. In 1990 and 1991, he was director of that center. Professor Morata is a specialist in the Development Genetics, a scientific field in which he has been working for thirty years. To date, he has published 115 research articles in international science magazines and has been invited to lecture about his scientific work at numerous universities and research centers in Europe, America, and Asia. He has also directed supervised ten doctoral dissertations. He has received numerous prizes and distinctions, including the Rey Jaime I Prize for Genetic Research in 1996, the Ramon y Cajal National Research Prize in 2002, the Mexico Prize for Science and Technology in 2004 and the Príncipe de Asturias Prize for Scientific and Technical Research in 2007. He is also Doctor Honoris Causa from the universities of Alcalá (2007) and Almería (2008).

LUIS DE PABLO studied composition with Max Deutsch (Paris) and attended courses in Darmstadt (Germany). He has a Doctor Honoris Causa from the Complutense University of Madrid. Over 150 of his works have been performed by, among others, the Arditti Quartet, Pierre Boulez, Bruno Maderna, ONE, the Orchestra of Paris, the Metropolitan Orchestra of Tokyo, Claude Helffer, José Ramón Encinar, Rafael Frühbeck, Massimiliano Damerini, the SWF Baden-Baden Orchestra, NDR of Hamburg, the Berlin Philharmonic, and the Arbós Trio. He has been professor of composition in Buffalo (New York), Ottawa, Montreal, Madrid, Milan, and Strasbourg, and academician at the Academies of fine arts of Madrid and Granada, the Academy of Santa Cecilia in Rome and the Royal Academy of Belgium, as well as Officer of Arts and Letters in France and member of the European Society of Culture since 1966. In 1964 he founded the first Electronic Music Laboratory in Spain and the ALEA center in 1965. Among his numerous prizes are the Gold Medal of the King, those of the Spanish Red Cross and the Circle of Fine Arts (Madrid), the Guerrero Foundation Prize, the CEOE Prize for the Arts, the Rennes and Lille Medals (France), the Honegger Prize (Paris), and the Pierre Prize of Monaco. His works are taught and studied in Paris, Madrid, Ueno Gakuen (Tokyo), Instituto di Tella (Buenos Aires), Academy of Santa Cecilia (Rome), Elisabeth School of Music (Hiroshima), and UCLA (Los Angeles). The INA (National Audiovisual Institute) has interviewed him for its sound and image archives. His music is published by TONOS (Darmstadt), Salabert (Paris), and Suvini Zerboni (Milan).

NATHAN ROSENBERG is the Fairleigh S. Dickinson, Jr., Professor of Public Policy (Emeritus) in the Department of Economics at Stanford University and Senior Fellow, Stanford Institute for Economic Policy Research. His primary research activities have been in the economics of technological change. His publications have addressed both the questions of the determinants and the consequences of technological change. His research has examined the diversity of the forces generating technological change across industrial boundary lines, as well as the mutual influences between scientific research and technological innovation. His books include *The American System of Manufactures, Perspectives on Technology*, *Inside the Black Box*, *Technology and the Pursuit of Economic Growth and Paths of Innovation* (with David Mowery), *How the West Grew Rich* (with L.E. Birdzell, Jr.), *Exploring the Black Box*, *The Emergence of Economic Ideas*, and *Schumpeter and the Endogeneity of Technology.* He is the recipient of honorary doctoral degrees from the University of Lund, the University of Bologna, and Northwestern University. And was awarded the Leonardo da Vinci Prize for his contributions to the history of technology.

VICENTE SALAS FUMÁS has a Ph.D. in Management from Purdue University and an MBA from ESADE. He is currently Senior Professor of Business Administration at the University of Zaragoza and previously held that post at the Autonomous University of Barcelona. His research areas include the economic analysis of organizations and empirical studies of Spanish companies. He has published articles in national and international academic reviews and books such as *Economía de la empresa: decisiones y organización*, *El gobierno de la empresa*, *La empresa familiar en España* y *El siglo de la empresa*. In 1992, he received the Rey Jaime I Economics Prize.

FRANCISCO SÁNCHEZ MARTÍNEZ is a pioneer and promoter of astrophysics in Spain and first Senior Professor of Astrophysics at a Spanish university. He founded and directs the Astrophysics Institute of the Canary Islands (IAC), through which he develops and spreads knowledge of this science and related technologies around the country. He made it possible to construct the largest and most advanced telescope at this time: the Large Canary Islands Telescope, which has segmented mirrors and a diameter of over ten meters. He belongs to various international scientific societies and has multiple distinctions, including the following awards: Alfonso X el Sabio, the award for Civil Merit and the award of Isabel la Católica, the Cross of Naval Merit, the Grand Cross of the Order of the Canary Islands, Commander of the Royal Order of the Polar Star (Sweden), Commandeur dans l'Ordre des Palmes Académiques (France), Knight of the Royal Order of Orange-Nassau (Netherlands), and Commendatore della Stella della Solidarietà (Italy). He is Doctor Honoris Causa at the universities of Copenhagen and Florida, was awarded the Medal of Honor for Fostering Invention from the García Cabrerizo Foundation, and recipient of the CEOE Science Prize, the Gold Medal of the Canary Islands, and Canary Islands Research Prize.

JOSÉ MANUEL SÁNCHEZ RON received his undergraduate degree in Physical Sciences from the Complutense University of Madrid and his Ph.D. in Physics from the University of London. He is currently Senior Professor of Science History at the Department of Theoretical Physics of the Autonomous University of Madrid, where he was previously professor of Theoretical Physics. Since 2003, he has been a member of the Royal Academy of Spain, where he holds the "G" chair. He is also a member of the European Academy of Science and the Arts (Academia Scientiarum et Artium Europaea) and a corresponding academician at the Royal Academy of Exact, Physical and Natural Sciences, and the Académie Internationale d'Histoire des Sciences (Paris). He is the author of over 300 publications in the fields of theoretical physics and the history and philosophy of science, including some thirty books, including: *El origen y desarrollo de la relatividad* (1983), *Diccionario de la ciencia* (1996), *Cincel, martillo y piedra* (1999), *El siglo de la ciencia* (2000), for which he received the José Ortega y Gasset Prize for Essay and Humanities from the City of Madrid, *Historia de la física cuántica. I* (2001), *El jardín de Newton* (2001), *El poder de la ciencia* (2007), and *¡Viva la ciencia!* (with Antonio Mingote; 2008).

ANGELIKA SCHNIEKE & ALEXANDER KIND have worked together since 1984. Angelika Schnieke worked with Rudolf Jaenisch from 1978 to 1987 first at the Heinrich-Pette Institute, Hamburg, and subsequently at the Massachusetts Institute of Technology, investigating retroviral transgenesis in mice. Alexander Kind gained his Ph.D. in 1984 from the University of Cambridge, working with Sir Martin Evans on the early characterization of embryonic stem cells, later joining Rudolf Jaenisch's group developing animal models of connective tissue disorders. Both subsequently extended their research to the production of transgenic livestock at Colorado State University. From 1992–2003 they worked with the biotechnology company PPL Therapeutics in Edinburgh, Scotland, where Angelika Schnieke gained her Ph.D. from the University of Edinburgh for the development of somatic cell nuclear transfer as a means of generating transgenic sheep. Angelika Schnieke currently holds the chair in livestock biotechnology at the Technische Universität München. Their current research focuses on animal stem cells and the genetic engineering of livestock for bio-medical applications.

SANDIP TIWARI was born in India and educated at IIT Kanpur, RPI, and Cornell, and after working at IBM Research joined Cornell in 1999. He has been a visiting faculty at Michigan, Columbia, and Harvard, the founding editor-in-chief of *Transactions on Nanotechnology* and authored a popular textbook of device physics. He is currently the Charles N. Mellowes Professor of Engineering and the director of USA's National Nanotechnology infrastructure Network. His research has spanned the engineering and science of semiconductor electronics and optics, and has been honored with the Cledo Brunetti Award of the Institution of Electronic and Electrical Engineers (IEEE), the Distinguished Alumnus Award from IIT Kanpur, the Young Scientist Award from Institute of Physics, and the Fellowships of the American Physical Society and IEEE. His current research interests are in the challenging questions that arise when connecting large scales, such as those of massively integrated electronic systems, to small scales, such as those of small devices and structures that arise from the use of nanoscale.

photo credits

Cover: Nasa, Esa and T.M. Brown (STScI). Young and old stars found in Andromeda's halo.

p. 6: Nasa Solarsystem Collection. Stereo ultraviolet 3D image of the Sun.

p. 10: Candida Höfer/VG Bild-Kunst, Bonn 2008. *Biblioteca Geral da Universidade de Coimbra III*, 2006, 260 x 200 cm.

pp. 20–21: Collart Herve/Corbis Sygma. Atmospheric scenes in the Amazonian jungle at nightfall, Bazin, Brazil.

pp. 26–27: Nasa/The Hubble Heritage Team (STScI/Aura/Nasa). Southern Ring Nebula.

p. 30: Oliver Meckes and Nicole Ottawa/Science Photo Library/AGE. Electrical hair-dryer hermographie hermogram of the heat distribution over the diffuser attachment of a hair dryer, showing variation in temperature over its surface.

pp. 40–41: Harry Gruyaert/Magnum Photos. BBC II tv shots Apollo XII, London, 1969.

p. 62: Professor Harold Edgerton/Science Photo Library/Age. *Milkdrop Coronet*, 1957.

pp. 72–73: Professor Harold Edgerton/Science Photo Library/Age. *30 Bullet Piercing an Apple*, 1964.

pp. 84, 85, 86, 87, 70, 75, 81, 82, 84, 85 and 87: AIP Emilio Segrè Visual Archives.

p. 92: Oliver Meckes and Nicole Ottawa/SciencePhoto Library/Age. Microgears.

p. 97: Oliver Meckes and Nicole Ottawa/Science Photo Library/Age. Microsubmarine in artery.

pp. 104–105: Image 100/Corbis. A jumble of cables with a transparent connector.

p. 108: Gustoimages/SciencePhoto/Age. Laptop colored X-ray.

pp. 118–119: José Azel/Aurora/Asa. Computer monitors emit a green glow of light.

p. 128: Peter Menzel/Asa. Night view of a dish antenna at the VLA near Socorro, New Mexico.

pp. 134–135: Peter Menzel/Asa. Observatory in Mount Hamilton, San José, California.

p. 142: Jupiter images/Brand X/Corbis. Circuit window to outer space.

pp. 148–149: Nina Berman/Aurora Pictures/Asa. Earthlink NOC in Atlanta.

pp. 156–157: Adam Fuss. *Invocation*, 1992. Unique cibachrome photogram, 101.6 x 76.2 cm. Courtesy Cheim & Read, New York.

p. 160: Andrew Brooks/Corbis. DNA bands on a computer screen, June, 2001.

p. 166: Peter Menzel/ASA. Flourescence micrograph of human chromosomes.

p. 170: Frans Lanting/Corbis. Blood vessels of the hand.

p. 177: Roger Ressmeyer/Corbis. Cultures of photobacterium NZ-11 glowing in petri dishes. September 1983.

p. 184: Najlah Feanny/Corbis. Dolly, February 1997.

pp. 190–191: Visuals Unlimited/Corbis. Human egg and sperm, magnification of x 2500.

p. 194: RBM online/epa/Corbis. First embryo cloned in Britain.

p. 202: Microdiscoveries/Corbis. Colon cancer cell, magnification of x 2000.

pp. 208–209: Visuals Unlimited/Corbis. Confocal micrograph of inmortal cultured human cells (HeLa) stained to reveal the distribution of the cytoskeleton protein b-tubulin (green) and f-actinby phalloidin (red), as well as DNA in cell nuclei (blue). Two-photon fluorescence microscopy at a magnification of x1800.

pp. 222–223: Ingo Arndt/Minden Pictures/Asa. Diverse animal feet.

p. 224: Sanna Kannisto. *Heliconias*, 1998, c-print, 105 x 130 cm.

p. 230: Sanna Kannisto. *Rothschildia*, 2001, c-print, 74 x 90 cm.

p. 231: Sanna Kannisto. *Dasypus Novemcinctus*, 2001/2008, c-print, 74 x 94 cm.

p. 234: Lester Leftkowitz/Corbis. Elevated highway at rush tour.

pp. 248–249: Harry Gruyaert/Magmum Photos/Contacto. Renault industry in Turkey.

p. 256: Peter Marlow/Magnum Photos/Contacto. The Nuclear Power Station on the coast of Kent, Dungeness.

p. 263: Ian Berry/Magnum/Contacto. British Gas off-shore gas, Thailand.

pp. 272–273: Nasa/Earth Observatory Collection. Lancaster Sound in Northwest Passage between Canada and Greenland captured by the Moderate-resolution Imaging Spectroradiometer (MODIS) on the Terra satellite, September, 2007.

p. 276: Gideon Mendel/Corbis. Floods in Bihar, India, August, 2007.

pp. 296–297: Pablo Zuleta Zahr. *Men in Black*, *Baquedano* project, 2006.

p. 300: Stuart Franklin/Magnum/Contacto. Lloyds, London, 1999.

pp. 304–305: Jacqueline Hassink. *The Meeting Table of the Board of Directors of BMW*, Stuttgart, Germany, 15 December 1994. *The Meeting Table of the Board of Directors of Daimler Benz*, Stuttgart, Germany, 25 August 1994. *The Meeting Table of the Board of Directors of EDF Electricité de France*, Paris, France, 30 November 1994. *The Meeting Table of the Board of Directors of Philips Electronics*, Eindhoven, the Netherlands, 15 July 1993.

p. 308: Gueorgui Pinkhassov/Magnum Photos/Contacto. Office.

p. 314: Peter Marlow/Magnum/Contacto. The Global Powerhouse, the City of London, the London Metal Exchange, 1999.

pp. 320–321: Erich Hartman/Magnum Photos/Contacto. Security boxes, Zurich.

p. 328: Peter Marlow/Magnum/Contacto. London,1999.

pp. 332–333: Harry Gruyaert/Magnum/Contacto. *The Stock Market*, 1998.

p. 336: Hiruji Kubota/Magnum/Contacto. The Head Office Building of the Hong Kong and Shanghai Banking Corporation (now the Hongkong Bank), designed by British architect Sir Norman Foster, 1996.

pp. 342–343: Tyler Hicks/The New York Times/Contacto. Afganistan.

p. 346: Bruno Barbey/Magnum/Contacto. *The Amazon River*, 1966.

pp. 352–353: Steve Mc Curry/Magnum Photos/Contact. Man pulling cart past burning destroyed house, Kabul, Afganistan, 1985.

pp. 362–363: Reuters/Carlos Barria/Cordon Press. A man walks in front of a painting from the Contemporary Fine Arts of Berlin during the Art Basel international art show in Miami Beach, 2005.

p. 366: Eva-Lotta Jansson/Corbis. Visitors take in the huge Sun installation by Olafur Eliasson, part of the *Weather Project* at the Tate Modern, London, November 2003.

p. 370: Bruno Barbey/Magnum Photos/Contacto. *The horse*, Maurizio Cattelan.

p. 374: Jeff Goldberg/Esto.

p. 376: Roland Halbe.

p. 377: Margherita Spiluttini.

p. 379: Michel Denancé.

p. 381: Michael Moran.

p. 383: Nácasa & Partner.

pp. 384–385: Nigel Young/Foster and Partners

p. 386: Iwan Baan.

p. 390: Colin McPherson/Corbis. Damaged violins on display at Andrew Hutchinson's workshop at Hoylake, north west England, April 2008.

pp. 396–397: Candida Höfer/VG Bild-Kunst, 2008. *Stadtcasino Basel I*, 2002, 152 x 204 cm.

Edition
BBVA

Project Direction and Coordination
Office of the Chairman BBVA
Communication and Image Area BBVA
Department of Cultural Activities

Editorial Production
Turner

Editorial Coordination
Cecilia Gandarias

Concept, Introductions and Definitions
Carlos Wert

Texts
Janet Abbate, Sergio Alonso, Jesús Avila, Abhijit V. Banerjee, Francisco Calvo Serraller, Paul E. Ceruzzi, Carlo M. Duarte, Joan Esteban, Luis Fernández-Galiano, John B. Heywood, Gerard Holton, Alexander Kind, Cayetano López, Joan Massagué, José M. Mato, Robert McGinn, Ginés Morata, Luis de Pablo, Nathan Rosenberg, Vicente Salas Fumás, Francisco Sánchez Martínez, José Manuel Sánchez Ron, Angelika Schnieke and Sandip Tiwari

Translations
Wade Matthews

Text Editing
Keith Patrick

Graphic Editing
Lala Herrero

Design
Estudio Joaquín Gallego

Layout
Pablo Nanclares
Concha Pérez
Nicolás García Marque

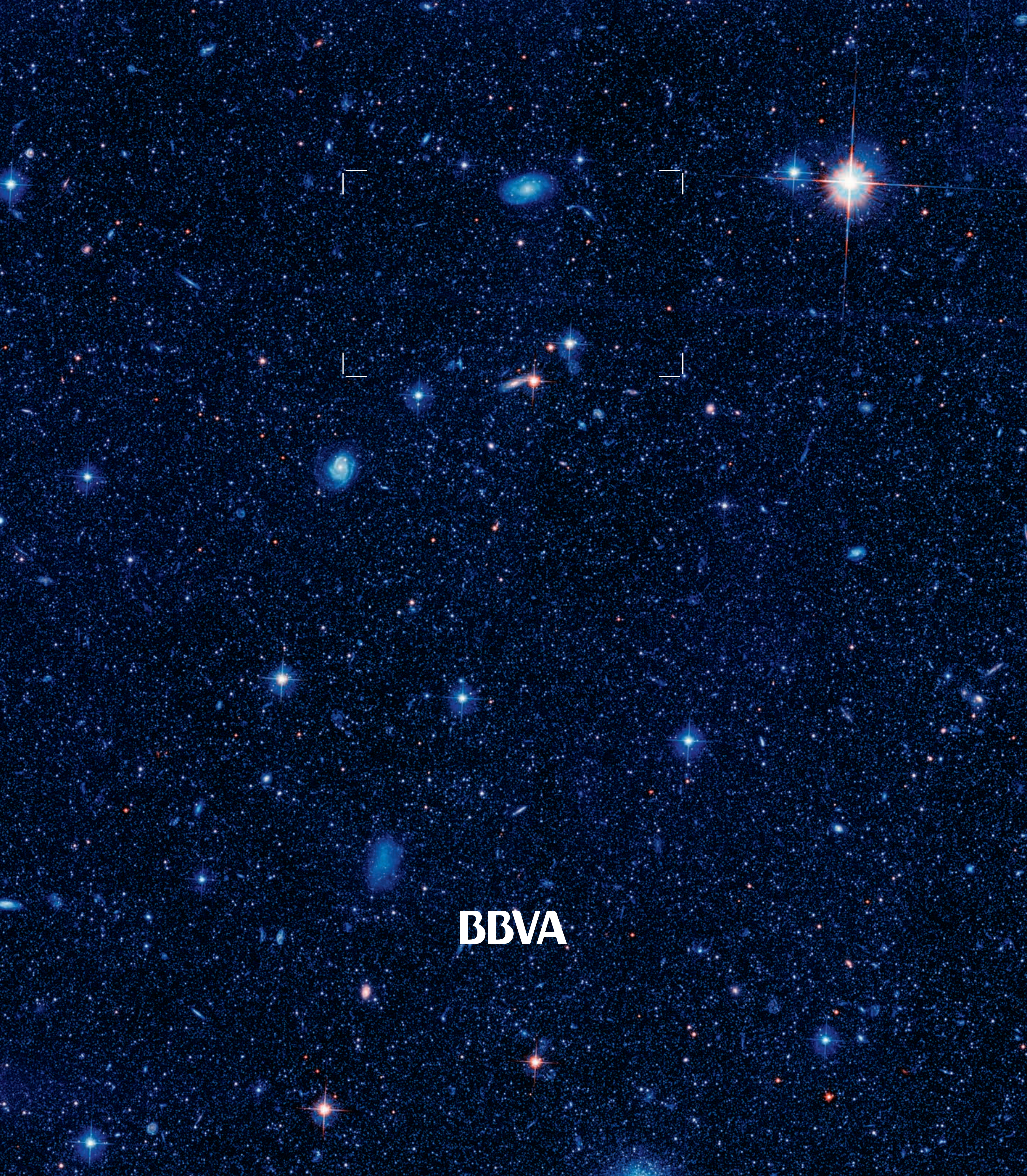
Colour Separation and Printing
Brizzolis

Binding
Ramos

DL: M-54481-2008
ISBN: 978-84-935133-7-5

Frontiers of knowledge was printed
in November 2008 in Madrid.
The typefaces used in its composition
were Rotis and Square.

The edition consisted of 4,500 copies
printed on 150 gr Gardapat paper.


BBVA