Introduction to Microeconomics by Luís Cabral is licensed under CC BY-ND 4.0. This license requires that reusers give credit to the creator. It allows reusers to copy and distribute the material in any medium or format in unadapted form only, even for commercial purposes. To view a complete copy of this license, visit https://creativecommons.org/licenses/by-nd/4.0

This book was typeset in LaTeX using the memoir style as well as multiple macros developed by the author. All images included in this book are in the public domain. Whenever appropriate, links to the license and author are included.

Cover design based on Arch (2005), oil on canvas (18x18in), by Luís Cabral.

Luís Cabral is Paganelli-Bull Professor of Economics at NYU’s Stern School of Business. Updated information about this book will appear at luiscabral.net/economics/books/micro/

Version 1.1 (September 2020)
CONTENTS

PREFACE

PART I: INTRODUCTION

1. THE ECONOMY
   1.1. Living standards and happiness ........................................ 5
   1.2. The capitalist revolution .................................................. 12
   1.3. The limits of the market economy ...................................... 26
   1.4. A sustainable economy .................................................... 32

2. ECONOMICS
   2.1. Scope and method ........................................................... 47
   2.2. Behavioral and social science .......................................... 57
   2.3. Central themes .............................................................. 68
   2.4. A force for the good ....................................................... 80

PART II: SCARCITY AND CHOICE

3. OPTIMAL CHOICE
   3.1. Feasible set ......................................................................... 88
   3.2. Preferences .......................................................................... 91
   3.3. The marginal rule ............................................................... 100

4. HOUSEHOLDS
   4.1. Consumption ....................................................................... 106
   4.2. Labor supply ....................................................................... 114
   4.3. Other household decisions .................................................. 123

5. FIRMS
   5.1. Production function ............................................................ 145
   5.2. Input mix ............................................................................ 154
   5.3. Output level and price ......................................................... 166

PART III: MARKETS

6. SUPPLY AND DEMAND
   6.1. Cost function and supply ..................................................... 185
   6.2. Willingness to pay and demand ............................................ 206
7. EQUILIBRIUM AND EFFICIENCY
   7.1. Competitive markets ......................................... 224
   7.2. Gains from trade and efficiency ............................ 244
   7.3. Price controls .................................................. 261

PART IV: MARKET FAILURE

8. MARKET POWER
   8.1. Sources and effects of market power ......................... 276
   8.2. Antitrust and competition policy .............................. 293
   8.3. The rise of the tech giants ................................. 296

9. PROPERTY RIGHTS
   9.1. Internalizing externalities ................................ 303
   9.2. Climate change ................................................. 316

10. INFORMATION
    10.1. Asymmetric information ...................................... 331
    10.2. Consumer protection .......................................... 342

PART V: SOCIAL JUSTICE

11. EQUITY
    11.1. Measuring and explaining inequality ....................... 352
    11.2. Discrimination ............................................. 371

12. SOLIDARITY
    12.1. Fairness ...................................................... 382
    12.2. Political economy ........................................... 388
    12.3. Taxation ...................................................... 398

13. OPPORTUNITY
    13.1. Migration ..................................................... 416
    13.2. Inter-generational mobility ................................ 425
    13.3. Housing, schooling and family ............................ 429

INDEX
We hear it from politicians, journalists, activists — anyone who’s paying attention. The clamor is near unanimous, or so it seems: We need to break with the capitalist system. Technology and economics have gotten us into a huge mess: (a) The planet is at great risk. (b) From France to Chile to the United States, social tension is at an all-time high, largely the result of increasing income inequality. (c) Stable jobs seem like a thing from the past. The list goes on and on. So, we need to get rid of a system that does not work and of the technocrats who designed it.

Economists have not always been at the forefront of policy. However, their influence has increased substantially in the past thirty or forty years. As such, we might say that the problem is not just the market economy but also economics as an academic discipline. Some swear by economics, some swear at economists. As I will argue in the next few hundred pages, both are right in some sense and both are wrong in a different sense.

This book is an attempt to correct the way economics is taught and at the same time an attempt to answer the critics’ wake-up call. The book’s motto might be

*Technology and the market economy got us into the present mess, but technology and the market economy are the only hope of getting us out of the present mess.*

The world and the world’s economy have changed a lot in the past half century or so. The microeconomics textbook, however, has really
not evolved that much since Samuelson’s Economics was first published in the 1940s. The graphs got better, new examples were added, but the core has remained the same for the longest time. One exception I am aware of is the core-econ project, in particular the eBook The Economy. I learned a lot from reading The Economy. In fact, I have adopted or adapted several features from that book. However, I have different ideas regarding the optimal organization of a textbook. To this I turn next.

STRUCTURE

Introduction to Microeconomics is divided into five parts. Parts II and III correspond to the “traditional” approach to microeconomics. I have no interest in throwing out the baby with the bathwater: The basic economics model of rational, optimizing behavior remains an important tool to understand economic agency (not so much how they should behave but rather how they actually behave). Part IV deals with market failure. While competitive markets represent an important reference point, most markets today deviate from that ideal model in fundamental ways. Part V, arguably the most innovative part of the book, is devoted to social justice: equity, solidarity, opportunity. Finally, Part I corresponds to a longish introduction to the economy (the real-world phenomenon) and to economics (the academic discipline).

NAVIGATING THE BOOK

Introduction to Microeconomics is primarily intended to be read as a pdf-based eBook. Other formats, such as ePub, have the great advantage of being easily scalable. However, they are very poor at handling graphs; and it’s nearly impossible to write an economics text without graphs. I chose font size and page size so as to allow for comfortable reading on a tablet and even on a large-screen smartphone. Naturally, you are free to print the pdf file if you are so inclined, but there is a loss in the process: One of the advantages of electronic publishing is the inclusion of hyperlinks. In the present context, a simple rule applies: if it’s blue, it will link for you. Data and photo sources, article and book references, individuals and events, and so on, are typeset in blue and accompanied by an external internet link. Chap-
ter, section, figure, table and page numbers, in turn, correspond to internal links and are also typeset in blue.

A special reference should be made to the page number (top right corner of each page), which functions as a “button” linking to the table of contents. From the table of contents, in turn, you can link to any chapter or section of the book by pressing the chapter or section name (typeset in blue).

**FEEDBACK**

*Introduction to Microeconomics* is a free, open access text primarily intended for college undergraduate courses. I plan to develop additional materials and will be happy to share them with instructors. I will also be very happy to receive any feedback you might have. The current version may be found on the first page after the cover. The latest version, as well as the version history, may be found on the book page. Please submit comments, corrections, typos, etc, to luis.cabral@nyu.edu.
PART I

INTRODUCTION
1.1. LIVING STANDARDS AND HAPPINESS

Popular wisdom has given us a number of jokes about money and happiness. Playwright George Bernard Shaw misquoted Paul of Tarsus by claiming that “lack of money is the root of all evil.” Alexander Hamilton, first US Secretary of the Treasury — and inspiration for the Broadway smash hit “Hamilton” — wisely remarked that “money isn’t everything, but it certainly keeps you in touch with your children.” Novelist Gertrude Stein assures us that “whoever said money can’t buy happiness didn’t know where to shop.” Addressing specifically the issue of happiness, actor Alan Alda teaches us that “it isn’t necessary to be rich and famous to be happy, it’s only necessary to be rich.” Spike Milligan, in turn, pleads, “all I ask is the chance to prove that money can’t make me happy.” And finally, a quote (of uncertain origin) that summarizes much of what this chapter is about:

I’ve been rich and I’ve been poor. And, believe me, rich is better.

All joking aside, a general principle that most if not all agree on is that a minimum living standard is a necessary condition for happiness. Not a sufficient condition, but certainly a necessary one. As Franklin D. Roosevelt famously stated, “We have come to a clear realization of the fact that true individual freedom cannot exist without
economic security and independence.” Economics is about managing — individually, as a household, as a country, as a planet — the limited resources we have so as to achieve a reasonable living standard for all.

MEASURING LIVING STANDARDS

Economics is largely a quantitative discipline. Of the many measures used to describe the level of economic activity and living standards, the most important is probably that of Gross Domestic Product (GDP). GDP measures the total goods and services produced in the economy at market prices and in a given period. If you get a haircut and pay $20 for it, then an extra $20 is added to GDP. The same applies to all other final-consumer sales of products and services.

A related measure of living standards is per-capita GDP. This is simply given by GDP divided by population. For example, in 2018 the US GDP was $20.50 trillion (that is, $20.50 \times 10^9$ dollars). The US population, in turn, was 327.2 million (source: Bureau of Economic Analysis). This implies that per-capita income was $20.50 trillion divided by 327.2 million, or simply $62,652 per capita. In addition to the levels of GDP and per-capita GDP, we frequently measure their respective growth rates. For example, US GDP in 2018 was 5.2% higher than in 2017.

There are a number of issues with using GDP and per-capita GDP when making comparisons across time. First, we must take into account that prices change over time. Even more important, the range of products and services available is very different at different moments in time: How do you compare per-capita GDP in 2000 and 1970, considering that back then there were no smartphones? (Recommended video: The Numbers Game: Let’s Party Like It’s 1973!)

Comparisons across countries are also difficult. For one, prices vary considerably from country to country, as I recently found when I ordered wine in Sweden. Equally important, some products and services — e.g., health or education — are available “for free” in some countries but not in others. For example, a visit to the hospital may be free in Denmark, whereas an American patient must pay a high price. This may lead to overestimating the value of GDP in the US relative to Denmark. Along similar lines, In Algeria (and in 2019) only 15 percent of women were part of the labor force (source: World
Bank). In the same year and in the Republic of Congo, that number was 68 percent. While Congo women earn market wages for their labor, most women in Algeria work at home, earning no market wage. This does not mean their work is less important, simply that it is not accounted by GDP measures.

So far, the issues with GDP measurement are largely technical problems which can be “solved” with appropriate corrections. For example, it is customary to make cross-country comparisons by correcting for price differences, that is, by computing per-capita GDP in purchasing power parity (PPP) units. However, the limitations of GDP as a measure of wellbeing run deeper than the above technicalities. How do we account for the quality of the social and physical environment? Do people have access to clean air and clean water? Do people trust each other? Is the social environment conducive to friendship? Do individuals and families have sufficient free time and vacation time? More generally, the argument can be made that the market price we pay is not always a good measure of how valuable things are. For example, many people spend more on diamonds than on water, but it would be easier for them to live without diamonds than without water. (We will return to these issues in a later chapter, namely in reference to the so called water-and-diamonds paradox.)

For all of these reasons, many question the legitimacy of GDP, and over the years there have been repeated calls for a different measure of living standards. For example, Robert F Kennedy’s 1968 speech at the University of Kansas included a deep criticism of the concept of Gross National Product (a similar measure to GDP). More recently and more specifically, New Zealand Prime Minister Jacinda Ardern, at the 2019 World Economic Forum, proposed a “well-being budget” as an alternative to simple growth (as measured by GDP). At about the same time, Cambridge’s Diane Coyle claimed that

Over eight decades after its introduction, there is a widespread consensus that GDP is no longer a useful measure of economic progress.

“Widespread consensus” is a bit of a stretch. I for one am not part of it. Here’s why: Referring to democracy, Winston Churchill famously wrote in 1947 that
No one pretends that democracy is perfect or all-wise. Indeed it has been said that democracy is the worst form of Government except for all those other forms that have been tried from time to time.

Many economists think similarly about GDP: it’s the worst measure except for all the other ones! GDP is not a single number that tells you unequivocally how well an economy or a society is doing. In fact, that single magic number will likely never exist. So, rather than looking for such magic number, we’re much better off by complementing GDP with other measures of living standards and, more broadly, measures of happiness. Here’s an analogy: When you want to know what the weather will be like, you look for (objective) measures such as temperature and humidity. Many weather sites in the US also provide a “feel like” index, which corrects for temperature by also taking humidity and wind into consideration. But I much rather have the values of temperature, humidity and wind than a “feel like” index. What does “feel like” mean anyway? Whose feeling is it? To summarize,
1.1. LIVING STANDARDS AND HAPPINESS

GDP AND HAPPINESS

As we have just seen, GDP is an index with many limitations. But does it at all help understand how living standards contribute to happiness? Figure 1.1 shows the relation between per-capita GDP and self-reported life satisfaction in the 144 countries for which data was available in 2017. The blue line (a logarithmic function) describes a possible approximation to the relation between the two variables. Clearly, the relation is positive. In particular, for low values of per-capita GDP — say, lower than $10,000 a year — the relation is particularly steep. The estimated curve implies, for example, that an increase in per-capita GDP from one to two thousand dollars is associated with an increase in “happiness” (self-reported satisfaction) from 3.786 to 4.282, a 13 percent increase. By contrast, an increase in per-capita GDP from $50,000 to $51,000 is associated with an increase in “happiness” from 6.584 to 6.599, a mere .02 percent increase. (As we will see at various points throughout this book this is an instance of an important “law” in economics, the law of diminishing marginal returns: the gain from getting more of a certain good tends to decrease as we get more of it.)

Three important notes regarding the relation between per-capita GDP and self-reported happiness. First, Figure 1.1 only shows a correlation between the two variables. No claim is made regarding the causal relation between the two. It could well be that there is a third variable — for example, social cohesion or health or political stability — that influences both GDP and happiness, to the point that we observe a correlation between GDP and happiness that does not correspond to causality. (The distinction between correlation and causality will be a constant throughout this book.)

A second note is that, notwithstanding the positive relation between per-capita GDP and happiness, we also observe considerable variation in the levels of self-reported satisfaction. For example, Botswana and Costa Rica had similar levels of per-capita GDP ($15,807 and $15,524, respectively). However, the happiness indicator is considerably higher in Costa Rica (7.225) than in Botswana (3.505). Clearly, there is more to happiness than just GDP. The World Happiness Report considers a list of factors including — in addition to per-capita GDP — measures of social support, life expectancy, freedom to make life choices, generosity, and the perception of corrup-
tion. I encourage you to read their report, which is published annually.

Notwithstanding these qualifications, the fact remains that a country like Afghanistan has both a very low per-capita GDP and a very low level of self-reported happiness, whereas a country like Luxembourg has high values of both. It’s possible — it’s likely — that economic conditions have something to do with it.

Economic living standards are an important component of human welfare and happiness. Per-capita GDP is a limited but helpful indicator of living standards.

To conclude this subsection, a third feature of note in Figure 1.1 is that the observations (the dots) are very clustered near the left axis. In other words, most countries in the world have relatively low levels of per-capita GDP (say, lower than $10,000), with a few countries enjoying substantially higher values (say, higher than $60,000). In addition to the levels of world inequality this reflects (more on this in Section 1.3), such variation makes it difficult to represent the data. Whenever we have observations with highly dispersed values, one trick economists use is to represent variables on a logarithmic scale. This we do on Figure 1.2, where the horizontal axis is drawn on a logarithmic scale (and everything else is as in Figure 1.1).

On a linear axis — the one we normally use — each tick increment corresponds to a constant increase in value. For example, each tick on the horizontal axis of Figure 1.1 corresponds to a $2,000 increase in per-capita GDP. By contrast, on a logarithmic graph, each equidistant tick corresponds to multiplying the variable in question by 10. For example, each labeled tick on the horizontal axis of Figure 1.2 corresponds to $1k, $10k and $100k, respectively. Also, so as to avoid printing multiple zeros, frequently we simply use exponential notation. For example, $10^0$ is the same as 1, $10^1$ is the same as 1 followed by 1 zero, and $10^2$ is the same as 1 followed by two zeros.

As can be seen, the dots in Figure 1.2 are more evenly spaced than in Figure 1.1. You will also notice that the relation between per-capita GDP and happiness looks like a linear relation. Don’t be fooled by this: the relation looks linear because the $x$ variable — per-capita GDP — is plotted on a logarithmic scale, not on a linear scale. As shown in Figure 1.1, the relation is very non-linear, specifically very
1.1. LIVING STANDARDS AND HAPPINESS

concave, with the effect of an increase in GDP much stronger at low values of GDP than at higher values of GDP.

In other words, when reading a graph on a logarithmic scale one must account for the fact that distances don’t always mean the same. For example, in Figure 1.2 the horizontal distance — that is, the difference in per-capita GDP — between Costa Rica and Luxembourg seems lower than the difference between Costa Rica and Afghanistan. However, as Figure 1.1 shows, in dollar terms the GDP gap between Costa Rica and Luxembourg is much higher than the GDP gap between Costa Rica and Afghanistan.

Another way of appreciating how distances don’t mean the same on a logarithmic scale as they do on a linear one is to examine the various ticks on the x axis between $10^1$ (that is, 10) and $10^2$ (that is, 100). They correspond to the values 20, 30, 40, ..., 70, 80 and 90. As we get closer to 100, the distance corresponding to a $10k$ increase becomes shorter and shorter.

As mentioned earlier, given the non-linearity of logarithmic scales we need to be careful about interpreting a straight line on a graph like Figure 1.2. A given shift in the x variable — in terms of horizontal distance — always leads to the same increase in the y variable. However, a given shift in x at low values of x corresponds to a much smaller change in per-capita income than the same shift in x at high

![Graph showing the relation between per-capita GDP and happiness](source: ourworldindata.org)
1.2. THE CAPITALIST REVOLUTION

Having established that per-capita GDP is an imperfect but helpful measure of living standards, we now turn to the question of what happened to living standards over the past few centuries, specifically as measured by per-capita GDP. Figure 1.3 presents estimates of the world average per-capita GDP from year 0 to year 2000. Economic historians actually have estimates going back to 1 million BC. However, I will spare you the boredom of the full graph. Let it just be noted that it is essentially flat all the way until we hit the 1700s. After that we observe a rapid inflection from a few hundred dollars to more than 7,000 in the 20th century.

The change in the time path of per-capita GDP in the 18th century is so sharp that economists refer to it as **history’s hockey stick**: a relatively constant per-capita level (near-zero growth) suddenly turning...
into a high-growth pattern, a seemingly sudden take-off of economic growth. Figure 1.3 also includes a reference to two important historical events closely related to the inflection point in history’s hockey stick. We will return to these later.

Figure 1.4 does two things. First, if zooms in on the past three centuries, starting in 1700. Second, it focuses on the evolution of per-capita GDP in six specific countries: China, Japan, India, Italy, UK, and US. Once again, we observe a series of “hockey sticks”. They appear less steep than the left panel for the simple reason that the time frame is three centuries rather than twenty centuries. The main differences across countries correspond to the time when take-off occurs. For Great Britain (now the UK) and for the US, it happened during the 18th century, whereas for China and India the take-off only occurs during the latter part of the 20th century.

To the extent that the time of take-off varies from country to country, we also observe significant variations in world inequality during the past three centuries or so. As can be seen from Figure 1.4, by 1700 there wasn’t much world inequality (that is, inequality across countries). For example, per-capita GDP in India and Great Britain were very similar. By contrast, in 1900 the UK-India gap was very significant. However, this does not mean that the gap is always increasing: once China takes off, for example, the level of world inequality declines (China grows faster than Europe or North America). We will return to this later.

The term capitalist revolution usually refers to the extraordinary pace of economic growth experienced by the countries that first adopted the capitalist system of economic organization. Capitalism is a loaded word, in part because people have strong feelings about it. In the language of economics, we use the term in a precise way because that helps us to communicate. Specifically, in economics we use the term capitalism to refer to a particular set of conditions and institutions. For our purpose, institutions are the laws and social customs governing the production and distribution of goods and services. The particular institutions that define capitalism as an economic system are private property, markets, and firms (firms being institutions that bring together capital and labor).
1.2. THE CAPITALIST REVOLUTION

The capitalist system combines the institutions of private property, markets and firms (which in turn combine capital and labor).

An additional note on terminology: Sometimes, economists refer to capitalist economies as market-based economies, or simply market economies. The term was used in particular to contrast capitalist systems from Soviet-style economies, where most economic decisions were made by central bureaucracies (central planning). Throughout the book, we will use both capitalism and market economies to designate the same system (with the understanding that, as we will see later in this section, there are many variants of capitalist, or market-based, economies).

A (VERY) BRIEF HISTORY OF THE ECONOMY

At the risk of oversimplifying, we may divide the history of the economy into three different phases, each with its own institutions. For centuries, the various economies around the world were characterized by self-sufficient, family-based production. Of the various institutions previously mentioned, the one that we most clearly observe is private property (arguably the oldest economics-relevant institution).
As these families produced a surplus of food and other products, they were able to exchange their surpluses in markets, typically local markets. This corresponds to a second phase in the history of the economy, the era of the market economy with family-based production. With respect to the previous era, we add a new economic institution: the market.

Finally, we come to the third phase in the history of the economy, the beginning of which is marked by Great Britain’s industrial revolution. The industrial revolution took place in the late 18th century and early 19th century. It was a process of rapid transition to mechanized production, a shift from agriculture to manufacturing, and rapid population growth. (Why did the industrial revolution first take place in Great Britain? This is one of the most fascinating questions in economic history, alas one we won’t have time to dive into.)

In terms of economic institutions, the industrial revolution led to the emergence of physical capital and of factories. This in turn led to the creation of firms and the distinction between capitalists (or capital owners) and workers. In sum, to private property (first era) and markets (second era), the third era in the history of the economy (the capitalist economic system) adds the institution of the firm (firms being institutions that bring together capital and labor).

The emergence of capitalism was associated with very rapid improvements in production technologies (i.e., technology progress) and gains in production efficiency. These trends led to unprecedented productivity improvements. For example, the economic value created by one hour of work increased very rapidly. Our next goal is to understand how and why this happened.

WHY CAPITALISM WORKS

The emergence of economics as a modern discipline can be traced to the emergence of capitalism. In Figure 1.3 we marked two events which took place in 1776: James Watt patents his steam engine; and Adam Smith publishes The Wealth of Nations. (US readers will note that 1776 was also the year the US declared independence from England.) Watt’s steam machine was not the first steam machine, but it was considerably more efficient than previous ones. It is frequently singled out as a watershed moment in Britain’s industrial revolution.
The Wealth of Nations, in turn, is normally considered as the birth of economics as a systematic discipline.

Adam Smith based his classic treatise on multiple visits to factories operating in England and Scotland. In other words, he observed the industrial revolution and the emergence of capitalism first hand. His theory of the benefits of capitalism rests on two pillars: the productivity gains from division of labour and the efficiency gains from free market exchange. We will return to these several times during the course. The Wealth of Nations had a huge influence in the discipline of economics. In a way, we can say that it created a new discipline.

Based on Smith’s work and on that of many economists since then, we have an understanding of how capitalism promotes economic growth. One first important point is that competition between firms incentivizes them to innovate. During the 18th and 19th centuries, it was a quest for a more efficient steam machine or a faster weaving machine. In the 21st century, we observe how Apple and Samsung, for example, vie for leadership in the smartphone market by constantly introducing new features.

A second factor is given by the division of labor and the increased efficiency it allows (i.e., gains from specialization). In Adam Smith’s own words (in The Wealth of Nations),

The greatest improvement in the productive powers of labour, and the greater part of the skill, dexterity, and judgment with which it is anywhere directed, or applied, seem to have been the effects of the division of labour.

Adam Smith proposed the example of a pin factory (an object which, as a result of the industrial revolution, went from a luxury item to one accessible to most people). If a pin factory employs one person only, then that person must execute all the tasks required to manufacture each pin. However, if the factory’s output is sufficiently high to justify hiring multiple workers, then each of them can specialize in one specific task and become good at it (i.e., efficient).

One reason why specialization allows for greater efficiency is scale economies. Suppose, for example, that each step in the pin production process requires a set up cost. If a worker works on one pin at a time, then the worker must incur this set up cost multiple
1.2. THE CAPITALIST REVOLUTION

A Ford Motor Company assembly line in 1913. Division of labor allows for worker specialization, which in turn leads to significant efficiency improvements.

...times. By contrast, if the worker specializes in one task only then the worker only has to incur the set up cost once.

A second reason why specialization allows for greater efficiency is **learning by doing**. If our worker sticks to doing one thing only, then the worker is better able to experiment with different ways of going about it, eventually choosing the best.

Last but not least, specialization allows for significant efficiency gains when there are differences in ability across workers. Suppose one of the workers is very good at cutting wire, whereas another one is very good at sharpening the ends. If they specialize in one of the tasks, then each of them will focus on what they are better able at, leading to higher overall productivity.

A related idea, one that is central in the work of Smith, is the idea that **market exchange** creates **value**. Consider the case of international trade. If Scotland and France are isolated from each other, then Scotland has to produce the wine it consumes and France the wool it uses. By contrast, if they trade with each other, then the Scots can specialize in what they’re best at (wool) and the French on what they are best at (wine), each exporting the surplus in exchange for the other country’s surplus. Note that, in the process trading wool for wine there is no change in the total quantity of wool and wine consumed in Scotland plus France. However, to the extend that Scotts value wine more than wool and the French value wool more than wine, we can say that trade creates value, that is, it increases the value attributed to it by the consumers who end up getting that wool and that wine. This is a very important idea, so important that we will return to it twice in the book.
International trade (one aspect of globalization) leads to the separation of location of production and location of consumption. This in turn allows for specialization at the world level, leading to significant efficiency improvements.

I previously mentioned the importance of the pin factory’s output being high so as to allow for the division of labor. If you live in a small town or a small country, then the demand for pins will be low. If that is the case, then the local pin factory will produce a low output level. If that is the case, then no division of labor will be possible, for the factory will employ one worker only, and that worker will have to take care of all the tasks required to make each pin. We conclude that the size of the market limits the extent of division of labor, one of Adam Smith’s important ideas.

**GLOBALIZATION**

Another important trend during the capitalist era is the emergence of globalization. When Scotland and France begin to trade, the market that French wine makers can sell to increases. Before, it was France; now, it’s France and Scotland. As we saw before, a bigger market brings greater opportunities for specialization and efficiency gains. In other words, globalization is a factor favorable to the efficiency gains promised by the capitalist system.

Take, for example, medical tourism is India. By 2015, the industry was estimated to be worth US$3 billion, and it is projected to grow to US$9 billion by 2020. According to CNN, each year about half a million patients visit India to seek medical care, coming from Bangladesh, Afghanistan, Iraq, Maldives, Oman, Yemen, Uzbekistan, Kenya, Nigeria and Tanzania, and many other countries. Dr Ashok Rajgopal, one of the country’s leading orthopedic surgeons, has performed more than 25,000 total knee surgeries, achieving un-
preceded efficiency levels (e.g., 28 replacements in 12 hours). We can see in this example the process described earlier: globalization leads to a greater market size, which in turn leads to specialization, which in turn leads to learning by doing, which finally leads to greater efficiency. Against these benefits, we must also admit that globalization implies a series of challenges. For example, a number of surgeons in Bangladesh, Afghanistan, etc, may have suffered a negative demand shock or even lost their jobs. Later in the course we will address some of these challenges.

**INNOVATION**

When we compare the 21st economy to the pre-industrial-revolution 18th century economy, we notice two important differences. First, we are more efficient at producing goods and services which already existed in the 18th century. Second, we created multiple new products and services. Both of these are instances of innovation, which we can divide into process innovation (producing the same product or service but in a more efficient manner) and product innovation (creating new products or services).

Consider first the case of process innovation. Steam engines have existed for a long time. Hero of Alexandria — mathematician and engineer in Roman Egypt — describes a simple turbine in the first century AD. In 1698, Thomas Savery developed the first commercial steam-powered device. Watson’s 1776 steam machine, which we alluded to earlier, was therefore not the first steam engine. Its importance lies in the fact that it was substantially better — substantially more efficient — than the previous ones. The onset of the industrial revolution is largely connected with this instance of process innovation.

Another particularly interesting example of process innovation is given by lighting. Box 1.1 shows how, over the centuries, we’ve become increasingly better at providing this basic service. However, despite these efficiency improvements in producing energy, electricity, wine, wool, pins, etc, the most salient difference between the 18th and 21st centuries is the incredible array of products and services to which we have access today but simply did not exist back then, no matter how much you would be willing to pay for them. And we do not need to go that far back to be shocked at the differences in
The figure below shows how much more efficient we have become, since 1800, at producing artificial light (source: *The Economy*). The horizontal axis measures time. The vertical axis measures, on a logarithmic scale, the amount of lumen produced by one average worker. Lumen is a unit of brightness. *Loosely speaking*, a 100W lightbulb would correspond to about 1600 lumens. One candle corresponds to about 12 lumens.

![Graph showing the increase in lumens per worker from 1800 to 2000](image)

Having started from close to $10^2 = 100$ lumens per hour of work in 1800, by the end of the 20th century we are closer to $10^7 = 10,000,000$ lumens per hour of work. In other words, we are $10^5$ more efficient. That’s 100,000 times more efficient!

So as to get a better idea of what this means, consider the following conceptual experiment. How many hours of work would it take to light up a ball room with the lumen-equivalent of 200 candles (think *Downton Abbey*) during one hour? If you were to do it in 1800 it would take about one person’s work (that is, one hour of work for one hour of light). In 1900, that number hand dropped to .002 workers, and in 2000 it was only .00002.

To look at it differently, consider the number of workers required to produce the lumen equivalent required to light up a typical sports stadium. If we were to use 1800 lighting technology, that would be 50,000 workers! By contrast, with 1900 technology we would only need 100 workers and with 2000 technology one worker suffices!
terms of product availability: we did not have personal computers before the 1970s, or (widespread) internet access before the 1990s, or the smartphone before the 2000s.

New products, more efficient production of existing products: In both cases, the capitalist system provides a very favorable system for the innovation that leads to higher living standards. What made inventors like Watson tick in the 18th century was largely the promise of economic profit from a better technology. Fast forward to the 21st century and we observe similar competitive races, for example Apple and Samsung constantly adding new products and features so as to increase their smartphone market share.

The capitalist system improves living standards in various ways: (a) market exchange creates value (both parties are better off); (b) markets allow for specialization, which in turn leads to greater efficiency; (c) the profit incentive leads individuals and firms to engage in product and process innovation.

VARIETIES OF CAPITALISM

Earlier we defined capitalism as the economic system that combines the institutions of private property, markets, and firms (capital and labor). Not all societies and economies have been organized according to this system. First, as an alternative to private property there have been a number of systems with common property. Examples include early Christian communities and Israel’s modern kibbutzim (plural of kibbutz). Similarly, not all societies and economies are
1.2. THE CAPITALIST REVOLUTION

based on markets. The Soviet Union (1917–1991) and China (until the late 1970s) are examples of societies where most economic activity was dictated by a centralized bureaucracy.

Moreover, capitalism is not a precisely-identified concept. Looking around today’s world we observe quite a number of varieties of capitalism. To a great extent, we may classify the US, France, and present-day China as capitalist systems. However, clearly they correspond to different varieties of capitalism. Take the issue of private property. In China, for example, state-owned enterprises play an important role, considerably more than in the US. In France, eminent-domain laws are more favorable to public projects than in the US. For example, if a rail line is planned between city A and city B and this requires crossing over privately-owned land, it will typically be easier for the government to obtain access in France than in the US.

Given this wide variety of systems, a more appropriate description of today’s societies is that they are mixed economies, a system combining free markets with state intervention (as well as private firms with public enterprises). One particular instance of the hybrid nature of today’s economies is given by the innovation system. As mentioned earlier, the capitalist system provides a very favorable setting for innovation and the resulting appearance of new products and services. However, from Finland to the US, from China to Switzerland, we observe that government investment and initiative has played an important role in innovation as well. Take for example the US innovation system. According to economist Mariana Mazzucato

The parts of the smart phone that make it smart — GPS, touch screens, the Internet — were advanced by the Defense Department. Tesla’s battery technologies and solar panels came out of a grant from the U.S. Department of Energy. Google’s search engine algorithm was boosted by a National Science Foundation innovation. Many innovative new drugs have come out of NIH research.

This should not diminish the genius of inventors and innovators like Steve Jobs. Knowing how to put things together, knowing how to combine previously developed innovations, is an art that few can accomplish. But one must also recognize the important role of
Scientific research and innovation provide a good instance of the value of a mixed economy, one which combines free markets and state intervention, private firms and public enterprises.

government-based financing in keeping the innovation system working. Perhaps the optimal system is neither the market economy nor the centrally planned one but rather the system that combines the best of free enterprise and government backing.

**CASE STUDY: CHINA’S GREEN REVOLUTION**

Private property is one of the institutions — one of the pillars — of the market economy. If we have not focused much on it it’s partly because private property is not exclusive to the capitalist system. As mentioned earlier, private property is arguably one of the oldest institutions in the history of the economy. But just like a healthy person does not value health until they lack it, so we tend to underestimate the role played by private property in making the market economy work.

One good example of this is China’s green revolution of the 1970s and 1980s. Decades of communist rule under Mao Zedong led the country to disastrous economic outcomes. For example, from 1958 to 1962 tens of millions of Chinese died of starvation. While there were various causes for the crisis, the famine is regarded as one of the greatest man-made disasters in human history, in particular the result of common property of land and other production means.

In 1976, Mao Zedong died and was succeed by Deng Xiaoping, a pragmatist (selected quote: “it doesn’t matter if a cat is black or white: as long as it catches mice, it’s a good cat”). Then two things happened, one in the halls of Beijing, one in the fields of a small rural village.
In December 1978, eighteen local farmers in Xiaogang, led by Yen Jingchang, agreed to break the law by signing a secret agreement to divide the land, a local People’s Commune, into family plots. Each plot was to be worked by an individual family. Each family committed to turn over a part of the crop to the government and the collective, and was allowed to keep the surplus. The results were nothing short of extraordinary: harvest levels during the first year were larger than the previous five years combined.

Inevitably, Beijing found out about the experiment. Deng Xiaoping, who could have punished the lawbreakers, decided to extend the so-called household responsibility system across China. He did so because he was a great pragmatist, but also because he listened to Xue Muqiao (1904–2005), the thought leader underlying the reform. In a 1977 letter, Xue prophetically states that

It is hard to motivate farmers if growth in agricultural production cannot bring corresponding growth in income. Any interest in working suffers if extra work is not rewarded. . . . Boosting farmers’ enthusiasm for agricultural production therefore outweighs improving the conditions for agricultural production.

In present-day economics jargon, we would say the system did not provide enough incentives (“enthusiasm”) for farmers to work hard.

In retrospect, the household responsibility system increased the farmers’ willingness to work hard and produce a significantly greater yield. Figure 1.5 shows the evolution of agricultural output following the rollout of the new system. The blue bars depict the average crop
index, where the 1978 value is normalized at 100 (left scale). The red line, in turn, represents the fraction of China’s land that adopted the household responsibility system (right scale). In 1978, no land had adopted the new system and the index was at 100. By 1984, all of China had adopted the new system and the output index had increased to more than 140.

The lesson of China’s rural reform is that private, well-defined property rights matter a great deal for the functioning of an economy. As a result of the reform, each family effectively had property rights over the additional output resulting from their extra effort. Or, as we would now say in economics jargon, the incentives were well aligned: I’m the one who is working harder, I’m also the one who reaps the benefit from that extra effort.

There is another lesson, one which economists are keen to repeat. Deng Xiaoping and the Xiaogang families are considered the heroes of the green revolution. However, as John Maynard Keynes famously noted,

Practical men who believe themselves to be quite exempt from any intellectual influence, are usually the slaves of some defunct economist. Madmen in authority, who hear voices in the air, are distilling their frenzy from some academic scribbler of a few years back.
1.3. THE LIMITS OF THE MARKET ECONOMY

In his 2009 documentary, *Capitalism: A Love Story*, director Michael Moore highlighted many of the shortcomings of American capitalism. (The film was described as “an examination of the social costs of corporate interests pursuing profits at the expense of the public good.”) In this section we focus on two: inequality (and discrimination) and the environment.

Among the many limitations of the capitalist system, inequality and the environment stand out as the most significant ones.

INEQUALITY

As mentioned earlier in the chapter, history’s “hockey stick” — the take-off of capitalism — affected different countries at different times, thus generating significant inequality across countries. In 1700 there was relatively little difference across countries in terms of per-capita income (a maximum factor of about 2, corresponding to UK vs China). By 2000 the difference is considerably higher (a maximum factor of 20, corresponding to the US vs India). It’s not that India became poorer in 2000 than it was in 1700. In fact, the opposite is true. What happened was that, while most countries in the world became richer in the past few centuries, some did so at a much, much faster rate — namely the ones that experienced the “hockey stick take off” at an earlier moment in history.

Notwithstanding the previous paragraph, we also observe a reverse trend in world inequality when considering specific pairs of countries. For example, for the past few decades we have observed that China, having taken off in the late 20th century, has narrowed the per-capita GDP cap with respect to the US (and other rich economies), so we may say that, to some extent, the degree of world inequality has decreased.

There are many dimensions of the inequality phenomenon. In addition to inequality across countries, we also have within-country...
inequality. Figure 1.6 partly documents the evolution of inequality in the US. In the 1960s, the top 1 percent of the population earned about 10% of total income. The bottom 50 percent of the population, in turn, earned a little more than 20%. This is already a rather unequal distribution, but it’s nothing compared to what came later. In 1995, the top 1 and bottom 50 curves crossed, and by 2010, their positions are reversed with respect to the 1960s: Now the top 1 percent earns close to 20% of total income, whereas the bottom 50 percent gets about 12.5% of the pie.

A big question in economics is the source of this increase in inequality. Is it about capital versus labor, as economist Thomas Piketty would have us believe? Many economists, including this one, would rather point to the effects of skill-biased technical change (e.g., artificial intelligence), globalization, and several other factors (which we will carefully analyze in Chapter 11).

Although much attention has been given to the US economy, this is hardly the only case of high levels of inequality. Figure 1.7 shows income concentration among the 10 percent of various countries (or country blocks). Europe is the relatively more egalitarian part of the world, where the top 10 percent “only” get 34% of total income. At the opposite end of the spectrum, among Middle Eastern countries the top 10 percent amass about two thirds of total income. The US lies somewhere in the middle of the sample. However, were we to
restrict to high-income countries, the US would “lead” the world in terms of income inequality. China is somewhere between Europe and the US.

A related problem with the capitalist system — though not an exclusive of the capitalist system — is that of discrimination. There are many forms of discrimination in our societies, but two in particular are frequently related to economic conditions: gender discrimination and racial discrimination. Figure 1.8 may be helpful for the purpose of analyzing the economic dimension of discrimination. It shows the mean weakly wage in the US for different groups of people. The first fact to notice is that, while the mean wage rate is $886, the top 10 percentile corresponds to $2,129 (that is, 10 per cent of the US worker population is paid more than $2,129 per week), whereas the bottom 10 percentile corresponds to $430 (that is, 10 per cent of the US worker population is paid less than $430 per week).

Focusing more specifically on the issue of discrimination, we notice that men earn on average $973, whereas women earn on average $789. This gap of approximately 20% has declined somewhat over the past few decades but remains high. We have to be careful before stating that women are discriminated in the labor market to the tune of 20%. One reason for this difference in averages is that women choose jobs that pay less on average than the jobs chosen by men. However, even correcting for the effect of selection into dif-

---

### FIGURE 1.7

different types of jobs we still observe that women are paid less than men. Moreover, the very choice of jobs is likely influenced by society’s prejudices regarding gender roles. The bottom line is that, while difficult to quantify exactly, the evidence for gender discrimination is significant. We will return to this issue in Chapter 11.

A similar qualification must be made regarding the economic dimension of racial inequality. African Americans earn an average of $694, whereas white Americans earn $916, about 30% more. Is this gap due to discrimination in the labor market? One reason to be cautious is that wages are highly correlated with education levels, and education levels vary considerably across races. Specifically, Figure 1.8 shows that Americans with less than a high-school education earn an average wage of $553, whereas, at the opposite end of the distribution, Americans with an advanced degree earn an average of $1,521. And access to advanced education is highly biased with respect to race. This suggests that, in addition to discrimination in the labor market, one must also consider differences in access to education.

Economists have studied extensively the factors underlying the clearly lower living standards of black and African-Americans. Education level, household structure, neighborhood of residence, etc — all of these economic-related factors play an important role. However, the danger of a purely economic analysis is to think that race inequalities can be treated as a “technical” problem, one for which politicians can find a “technical” solution. In fact, the problem runs
deeper and has its source in implicit (or explicit) prejudice which re-
mains in place even decades after the law has recognized the equal
status of all citizens. In sum, economics can help explain the prob-
lem, but is not itself sufficient to solve it. We will return to the issues
of inequality and discrimination in Chapter 11, within the broader
context of social justice.

THE ENVIRONMENT

A second important challenge to the current capitalist system is the
effect that it’s having on the environment. Unfortunately, the de-
bate on climate change has become too polarized and politicized. As
much as possible, economists try to bring the order of logical thought
into the matter. This we will attempt in the next few paragraphs.

There is a general consensus regarding two important correla-
tions: The correlation between CO$_2$ emissions and atmospheric CO$_2$
concentration; and the correlation between atmospheric CO$_2$ con-
centration and temperature anomalies. We next consider these.

Figure 1.9 plots the three time series in question. The blue line
shows the atmospheric concentration of CO$_2$ since 1700, with units
measured on the left scale (parts per million). As can be seen, atmo-
spheric concentration increased slowly until about 1950, from about
275 parts peer million in 1700 to about 310 part per million in 1950.
After that, we observe an increase at a faster pace. By 2010, the level
of atmospheric CO$_2$ concentration has already passed the 400 mark.

The second time series of interest is the level of CO$_2$ emissions
due to economic activity (transportation, travel, manufacturing, etc).
This corresponds to red line in Figure 1.9 and is measured on the
right scale (billions of metric tonnes). By the beginning of the 20th
century (the earliest date for which we have reliable worldwide data,
total emissions were about 10 billion metric tonnes. By 2010, the
value has already passed 35 billion metric tonnes, an almost four-
fold increase.

The third time series of interest is the difference in temperatures
with respect to the “normal” level, that is, the level one would esti-
mate to result from the “natural” evolution of the planet. This is a
controversial calculation — different scientists will give you differ-
ent numbers — but there is broad consensus with regard to the val-
ues in Figure 1.9 (measured in degrees centigrade, marked on the far
right scale). And the evidence suggests that, in the 18th century, the planet’s temperature was about a tenth of a degree colder than normal — if that — whereas, by the beginning of the 21st century, the temperature anomaly is rapidly approaching one full degree centigrade (about 2.5 degrees Farenheit).

I purposely chose the three scales in Figure 1.9 so as to make the two correlations easy to follow. First, the increase in CO₂ emissions during the 20th century is very closely followed by the increase in CO₂ atmospheric concentration. Second, the evolution of the temperature over a period of three centuries tracks very closely the evolution of atmospheric concentration.

While there is broad agreement regarding the correlation between these time series, there is some disagreement regarding the nature of causality. There is a general consensus that human activity — in particular economic activity in the more developed countries — has contributed significantly to the increase in atmospheric CO₂ concentration. There is some debate as to the precise level of human contribution to atmospheric CO₂ concentration, but it is generally agreed to be high.

The issue of the relation between atmospheric concentration and temperature is more controversial. It is generally agreed that it’s positive, but the agreement is less than universal when it comes to the
1.4. A SUSTAINABLE ECONOMY

As cargo ships steam across the oceans, the tiny aerosol particles in their exhaust act as seeds around which moisture in the atmosphere can condense. Occasionally this results in ship tracks becoming visible in cloud imagery.

NOAA National Environmental Satellite, Data, and Information Service (NESDIS)

precise relation between atmospheric CO\textsubscript{2} concentration and temperature. In other words, Figure 1.9 clearly establishes a close correlation between atmospheric concentration and temperature anomaly, that is, deviations from the “normal” temperature. However, as we will see Chapter 2, correlation does not imply causality. Climate scientists make a strong argument that this correlation corresponds to a causal relation, but there is considerable debate about the precise relation, in particular the rate at which changes in atmospheric CO\textsubscript{2} concentration translate into temperature increases, and ultimately on what the planet’s temperature will be in 10 or 100 years’ time.

Finally, there is also broad consensus that significant increases in temperature may have significant effects on the planet’s eco-system. Depending on whom you talk to — even among scientists — these effects may vary from very significant to catastrophic to apocalyptic. As the IPPC reports admit, it’s very hard to predict climate change, so the best we can do is to assign probabilities to the various possible outcomes. This includes some very drastic outcomes that occur with strictly positive probability.

In sum, notwithstanding disagreement regarding precise values, there is broad consensus that climate change is a very serious problem and that economic activity has played, and continues to play, a central role in it.

1.4. A SUSTAINABLE ECONOMY

In an article prepared for the 2017 World Economic Forum, we read that
Based on GDP and other measures of well-being, humanity has never been better off. ... Yet these historic advantages are being matched by a range of challenges felt keenly by many citizens.

This is an appropriate summary of the previous two sections. At some level, capitalism has improved living standards remarkably. However, in the process it has also created a number of imbalances, including inequality within and across countries, racial and gender discrimination, climate change and resource depletion. The precise extent to which the capitalist system is the cause of these imbalances is open to debate, but one must recognize that the imbalances have taken place within the context of the capitalist system.

These considerations form the basis for the concept of sustainable development or, alternatively, a sustainable economy. Broadly speaking, economists agree that a sustainable economy requires

1. The right incentives for individuals and firms to engage in cost-reducing and quality improving innovation, as well as efficient production.
2. A stable society, in particular one that is just.
3. A stable biophysical environment and resource base.

It’s fair to say that the average capitalist economy has done a decent job at addressing Point 1: Private property and market competition provide strong incentives for individuals and firms to come up with new and improved products, as well as becoming more efficient and producing existing ones. The greatest challenges are to address Points 2 and 3, social justice and the environment.

The United Nations’ 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a useful itemization of the generic goals of social justice and biophysical sustainability. At its heart are the 17 Sustainable Development Goals (SDG) to be achieved by 2030, including:

1. No Poverty
2. Zero Hunger
3. Good Health and Well-being
4. Quality Education
1.4. A SUSTAINABLE ECONOMY

The UN’s Sustainable Development Goals, adopted on September 25, 2015, as a part of the 2030 Agenda.

5. Gender Equality
6. Clean Water and Sanitation
7. Affordable and Clean Energy
8. Decent Work and Economic Growth
9. Industry, Innovation, and Infrastructure
10. Reducing Inequality
11. Sustainable Cities and Communities
12. Responsible Consumption and Production
13. Climate Action
14. Life Below Water
15. Life On Land
16. Peace, Justice, and Strong Institutions
17. Partnerships for the Goals

Some of these goals are more closely related to the economy, some are more a function of institutions such as democracy and the rule of law. All are at least indirectly related to the structure and performance of the economy. In what follows, we consider four particular aspects of the idea and ideal of a sustainable economy.

SOCIAL SUSTAINABILITY

It has become common to attribute the rise in inequality to the influence of particular parties within particular countries. Undoubtedly
politics has played an important role. However, one might say that the 21st century economy is characterized by "structures of inequality": There are features of the current economy that naturally lead to greater inequality. (By "natural" I mean "unless specific intervention to the contrary is taken."). The popular literature offers multiple accounts of this phenomenon in books with titles such as The Winner-Take-All Society, Average is Over, or The Vanishing Middle Class. The arguments vary, but the common theme is that various features of our economy and society are responsible for a rapid increase in levels of inequality. We will go through these in detail in Chapter 11.

Social sustainability requires keeping these inequality trends in check. There is no magic number telling us how much inequality a society can tolerate, but the level of social instability observed, for example, in the 2011 Occupy Wall Street protests or the 2018 Yellow Vests movement suggests that we are not far from that threatening magic number.

Social sustainability is not just about income inequality. The UN’s very first SDG (cf p. pagerefsgd) calls for an end to poverty by 2030. The number of people living in absolute poverty has fallen four-fold since 1980, a remarkable achievement for which the capitalist system deserves some credit. However, many countries still lag behind the economically developed world, and significant pockets of poverty persist in otherwise wealthy nations (including the US). Moreover, even when material poverty is alleviated or even eliminated we observe continuing — if not increasing — cases of social exclusion.

The distinction between poverty and social exclusion is impor-
tant as it helps explain the disconnect between the economic-growth and the social-justice diagnoses of what the capitalist system has achieved. If we examine income levels, it is undeniable that most of the world is better off now than it was half a century ago. Growth is good, growth has been good. But take two households, one living in 1970 and one living in 2020, with the same real income level (i.e., correcting for inflation), or perhaps with the 2020 household earning a slightly higher income. Suppose that neither household has Internet access. In 1970, lack of Internet access is a foregone conclusion, as there was no Internet back then. For a household in 2020, not having Internet access implies an enormous barrier to social integration, for example, difficulty in access to education and multiple other services. Therefore, even if both households’ earnings (in 1970 and in 2020) place them above the poverty line — no one is dying of hunger — there is a clear sense of social injustice about not having Internet access in 2020.

Social sustainability requires an increasing engagement by individuals and by various institutions — both governmental and non-governmental — in remedying — both structurally and on a case-by-case basis — the excesses implied by a system that disproportionally benefits some in favor of others. In Chapters 11, 12 and 13 we will address these issues in greater detail, including controversial issues such as wealth taxation.

Social sustainability also requires sustainable social relations. Gender inequality, for example, is an important hindrance to social sustainability. Box 1.2 reports on a specific aspect of gender inequality, the missing women problem. And, as recent events in the US show, racial conflict is also a major challenge to social sustainability.

SUSTAINABLE RESOURCE USE

In Section 1.3 we outlined the limitations of the capitalist system, or at least the versions of the capitalist system we’ve lived with for the past two centuries. One of these limitations is the negative impact that economic activity has had on the environment, both on the physical environment and on the biosphere.

If the market works so well in allocating resources — as Adam Smith argued — how come natural resources have been so poorly managed by the capitalist system? As we will discuss in greater de-
Box 1.2: 100 Million Missing Women.

Frequently we assume that there are slightly more women in the world than men. In the US and in 2013, for example, there were 161 million females and only 156 million males. However, worldwide there are fewer women than men. What explains the female/male ratio in the population? In an influential article, economist Amartya Sen stated that

At birth, boys outnumber girls everywhere in the world, by much the same proportion—there are around 105 or 106 male children for every 100 female children. Just why the biology of reproduction leads to this result remains a subject of debate.

If men and women receive similar nutritional and medical attention and general health care, then women tend to live noticeably longer than men. In Europe, the US, and Japan, despite the persistence of various types of bias against women, women outnumber men substantially. However, in South Asia, West Asia, and China, the ratio of women to men can be as low as 0.94. Specifically, Sen calculates that

A great many more than 100 million women are “missing.” These numbers tell us, quietly, a terrible story of inequality and neglect leading to the excess mortality of women.

Several scholars have attempted to explain the sources of this large gap, with explanations ranging from selective abortion and female infanticide to discrimination in access to employment and health services. The jury is still out as to the relative importance of each of these factors, but they all seem to reflect some form of gender discrimination.

tail in Chapter 9, the problem is essentially one of poorly enforced property rights. In some sense, everyone owns planet Earth; in some other sense, nobody does. Many actions that I perform harm the planet, which implies that they harm me as well. However, by harming the planet I also cause harm to billions of other people who have little or no control over my actions. In this sense, property rights over
1.4. A SUSTAINABLE ECONOMY

The planet are not well defined, or not well enforced, or both. It’s a bit like farms in China before the responsibility system was put in place (see p. 23).

**Sustainable resource use** requires, first of all, the acknowledgement that economic activity does not take place in a vacuum, rather in the context of a specific planet (or, looking at future prospects, in the context of a specific solar system). Figure 1.10 illustrates this point. Particularly important are the arrows in green, representing the use of natural resources related to land, raw materials, energy, water; and the red and blue arrows connecting the economic agents (firms, households) to the surrounding environment (arrows which represent, inter alia, pollution and waste). It is fair to say that, in the past, economic analysis has placed greater emphasis on the remaining arrows (market based relations, firm investment, etc) than on the environment-related arrows, that is, the ones that link the economy to its environment. It’s time to change that, for the sake of an economy with sustainable resource use.

From an economics point of view, it helps to distinguish three types of issues related to resource sustainability: (a) exhaustible resources, (b) environment, and (c) climate change. Let us next discuss each of these in turn. Since at least Thomas Malthus’ classic, *An Essay on the Principle of Population*, economists have been aware that the planet has limited amounts of various resources, and that the passage...
of time — as well as population growth — renders those resources increasingly scarce. Malthus predicted that

The power of population is so superior to the power of the earth to produce subsistence for man, that premature death must in some shape or other visit the human race.

Since then, and with remarkable regularity, multiple experts have predicted calamities of one sort or another on account of the limited supply of resources. In the 1968 classic, *The Population Bomb*, Paul Ehrlich warned of a worldwide famine in the 1970s and 1980s due to overpopulation. Four years later, *The Limits of Growth*, a report by an MIT team requested by the famous Club of Rome, predicted that the planet’s oil reserves would be depleted before the end of the century.

Limited resources remains a problem for humankind. In fact, the limited availability of resources is, in a certain way, the central problem in economics. However, of the three problems listed above — limited resources, environment, climate change — the threat posed by the scarcity of natural resources such as oil and copper and gold is arguably the least dramatic one. The reason is that market prices — Adam Smith’s “invisible hand” — tend to do a good job in turning economic agents away from resources that become increasingly rare. By contrast, most environmental problems result from market failure. In particular, climate change is a truly global problem (unlike most other environmental problems), that is, one that cannot be solved by the market or by local or even national authorities in isolation. In Chapter 9, we will discuss government policies, as well the role of individuals and non-governmental institutions, in achieving an environmentally sustainable economy.

**SUSTAINABLE BUSINESS**

In a famous — or infamous, you call it — *New York Times Magazine* article, economist Milton Friedman stated that

In a free-enterprise, private-property system, a corporate executive is an employee of the owners of the business. He has direct responsibility to his employers. That responsibility is to conduct the business in accordance with their desires, which generally will be to make as much
money as possible while conforming to the basic rules of the society, both those embodied in law and those embodied in ethical custom. Of course, in some cases his employers may have a different objective.

Friedman is frequently quoted with the more catchy and provocative statement, “The social responsibility of business is to increase its profits.” However, as the above paragraph shows, he makes two important qualifications. First, a manager’s responsibility toward the firm’s owners is moderated by “the basic rules of the society, both those embodied in law and those embodied in ethical custom.” Second, Friedman admits that “in some cases [firm owners] may have a different objective” than profit maximization.

These two qualifications are quite important. First, the norms of “ethical custom” related to business have evolved considerably in half a century (that is, since Friedman wrote his piece). Second, many corporations (most corporations?) espouse mission statements that go well beyond profit maximization. In 2019, the Business Roundtable (BRT), a non-profit association whose members are CEOs of major US companies, released an updated Statement on the Purpose of a Corporation by which the CEOs commit to lead their companies “for the benefit of all stakeholders — customers, employees, suppliers, communities and shareholders.”

Many analysts saw BRT’s Statement of Purpose as the ultimate vindication of the supremacy of stakeholders over shareholders, a condition *sine qua non* for the sustainability of business and society. To put it differently: A society must pursue the common good. Corporations are part of society, hence they too should pursue the common good. By contrast, other experts — mostly skeptical economists — interpreted BRT’s Statement of Purpose as little more than a marketing coup.

The concept of **corporate social responsibility** (CSR) largely corresponds to the evolution of beliefs reflected in BRT’s Statement of Purpose. (Other terms related to CSR include corporate sustainability, sustainable business, corporate conscience, corporate citizenship, conscious capitalism, and responsible business. These terms don’t necessarily mean the same but clearly overlap, and clearly contrast with Friedman’s view of the firm as pure profit maximizer.) While it is difficult to find a precise definition of CSR, most would agree
that it corresponds to a company’s sense of responsibility towards the community and the environment — both ecological and social — in which it operates.

One perspective on CSR is that it is good business. For example, the NYU Stern Center for Sustainable Business’s mission statement states that “sustainable business is good business: delivering better financial results while protecting the planet and its people.” The Center’s goal is to help business leaders “embrace proactive and innovative mainstreaming of sustainability, resulting in competitive advantage and resiliency for their companies as well as a positive impact for society.” This view is frequently encapsulated in the catchy phrase, “doing well by doing good” (where “doing well” is understood as shareholder performance and “doing good” is understood as being socially responsible).

A more “radical” perspective acknowledges that frequently the benefit of relevant stakeholders (e.g., local communities or the environment) comes at the expense of shareholders’ financial value. However, even then a socially responsible CEO does not necessarily violate a broader interpretation of the Friedman doctrine, whereby the CEO is responsible for pursuing the company’s mission. Fortunately, an increasing number of companies adopt mission statements which explicitly include elements of CSR; and those who acquire shares in those firms effectively buy into those mission statements. As a result, a CEO who conducts “business in accordance with [the shareholder’s] desires” (Friedman’s words) is effectively pursuing CSR even if at the expense of shareholder financial performance.

**HUMAN-CENTERED ECONOMY**

The *Universal Declaration of Human Rights*, the milestone document proclaimed by the United Nations General Assembly on December 10, 1948, states in its very first article that “All human beings are born free and equal in dignity and rights.” As part of his 2020 presidential campaign, then candidate Andrew Yang proposed that “The unit of a Human Capitalism economy is each person, not each dollar.” Finally, at a May 6, 2020 briefing to discuss the New York post-COVID-19 re-opening plan, Governor Cuomo asked and answered a rhetorical question: “The question comes back to how much is a human
life worth? To me, I say cost of a human a human life is priceless, period.”

An important condition for a sustainable economy is that it be based on people, not on money. In this regard and at a general level, the above statements seem bland, obvious, almost tautological. However, as one goes deeper into the issues one realizes there is considerable controversy. First, do all beings have the same dignity and rights? Highly influential philosopher Peter Singer defines a person as “a being who is capable of anticipating the future, of having wants and desires for the future,” thus excluding, for example, many handicapped infants and patients suffering from Alzheimer’s disease, a distinction some find offensive and excessively utilitarian.

Second, while we may consider human beings to be priceless, the truth is that, in practice, we are forced to make comparative judgments where the economic value of a life comes into play. For example, a new regulation that costs X and promises to save Y lives has to be judged based on some standard. In most countries, policy makers and courts employ the concept of the value of life (VL), an economic value used to quantify the benefit of avoiding a fatality. Different institutions in different countries in different years have proposed different VL. For example, in the US and in 2010, the Environmental Protection Agency set VL at $9.1 million, whereas the Food and Drug Administration set it at $7.9 million. The Department of Transportation, in turn, set VL at $9.2 million in 2014 and $9.6 million 2016. In other developed countries we find somewhat lower values, including $4.2 million in Australia (2014), $4.14 million in New Zealand (2016), $4.15 million in Sweden. Finally, for the purpose of estimating the value of life-extending measures, private and government-run health insurance plans apply VL on a per-year-of-quality-life basis, with values ranging from $50,000 to $129,000 per year.

In sum, while most agree with the primacy of the human person in a sustainable economy (and society), it is inevitable that some tension will arise between a purely utilitarian approach to personhood, at one end, and a working definition of value where economic calculus has no role to play, at the opposite end. Some common sense is required to balance these extremes. We will continue this important discussion (trade-offs) in Section 2.3. For now, let us return to Yang’s notion of “human capitalism,” the central tenets of which are:
Upon surveying a group of Yale students, economist Joel Waldfogel concluded that many of them valued holiday gifts received considerably less than their market price. Overall, Waldfogel estimated that ill-chosen gifts caused between $4 billion and $13 billion a year in economic waste. There is, however, considerable variation in the effect of gift-giving: When the two parties know each other well (e.g., father and daughter), then the likelihood of an ill-chosen gifts is smaller.

More important, even economists understand that an important component of the value of receiving a gift is not the gift itself but what it signifies (“it’s the thought that matters”).

Since the study on “the cost of Christmas” was made, one important development has taken place: the emergence of the gift card. Almost non-existent in 1995, by 2019 gifts cards already exceeded $163 billion in the US only. This is positive in terms of the cost of ill-chosen gifts, but it also implies a decline in the value of effort to find a good gift, that is, the value of the “thought” that matters.

**(Box 1.3: The economic cost of Christmas)**

(a) humans are more important than money; (b) the unit of a Human Capitalism economy is each person, not each dollar; and (c) markets exist to serve our common goals and values. While I don’t disagree with these principles, I prefer to rephrase them as follows:

1. Labor is more than just a production factor, that is, work has an intrinsic, subjective value beyond the economic value it creates.
2. Market exchange too generates value that surpasses the transaction value itself.
3. The value of economic freedom transcends the economic efficiency it engenders.

Let us next consider each of these in turn, beginning with the subjective value of work. A commonly proposed policy to address increasing inequality is to enact a Universal Basic Income: every citizen receives X regardless of whether they are employed or not. However, people derive value from being employed that goes well beyond the compensation provided by wage payments. Readers of a certain age
may remember one of the most popular TV shows of the 1970s, the *Mary Tyler Moore Show*. In its very first episode, Mary Richards interviews for a job with WJM News. Lou Grant, the station’s manager, offers Mary the job of Associate Producer. This comes as a pleasant surprise to Mary, who had actually applied for a secretarial job. Then this dialogue follows regarding job titles and compensation:

- The job pays $10 less a week than a secretarial job.
- That will be fine.
- If you can get by on $15 less a week, we’ll make you producer.
- No, no, I think all I can afford is associate producer.

On a more serious note, in a 2019 interview economist Robert Shiller stated that

> Jobs are more than a source of income. One has a story about one’s life that involves one’s job, and a job defines how I am important and why people should love me.

This matters, for example, when discussing the impact of AI on human welfare: There is some fear that it may lead to the “loss of even more of one’s identity.” More generally, Shiller argues that

> Economists should focus more on the ‘meaning of life’ and look beyond the figures on the page.

The second point raised is that market exchange creates value beyond the transaction value itself. Economic exchange involve human beings, individuals whose welfare depends not only on the result of the transaction but also on the relationship that is established in the context of such transaction. This becomes an issue, for example, when debating the impact of online sellers such as Amazon on local commerce (the idea that local stores create more value than the value of the goods they sell). In Chapter 9 we return to this important debate. Box 1.3 presents another example.

Finally, the third point is that economic freedom is a good in of itself. In Chapter 7 we show how, under certain conditions, a market economy leads to an efficient allocation of resources (a result economists refer to as the First Welfare Theorem). For many economists, this is a strong argument in favor of free markets. However, in Part IV of the book we present extensive evidence that the
conditions for the First Welfare Theorem to hold largely fail to hold. For many, this amounts to a justification for tight regulation of markets or, even more than that, for replacing markets with public provision of a variety of goods and services (housing, education, health, energy, etc). One argument against this view is that the defense of economic freedom is not solely based on its efficiency properties, rather economic freedom is an important component of human dignity. This is the view that “man’s eminence is to be seen in the fact that he chooses between alternatives,” and that the virtues of economic freedom — entrepreneurship, personal responsibility, equality of opportunity, meritocracy — are the best guarantee of upward mobility and ultimately social justice. One problem with this view is that it easily morphs into a radical form of libertarianism, where the individual should “exist for his own sake, neither sacrificing himself to others nor sacrificing others to himself.” It’s complicated.
KEY CONCEPTS

Gross Domestic Product (GDP)  per-capita GDP

growth rates  purchasing power parity

history’s hockey stick  take-off  capitalist revolution

capitalism  private property  market  firms

market economies  industrial revolution  capital

labor  technology progress  productive efficiency

productivity  competition  division of labor

specialization  scale economies  learning by doing

market exchange  value  globalization  innovation

process innovation  product innovation

varieties of capitalism  mixed economies

household responsibility system  discrimination

selection  sustainable development

sustainable economy  Sustainable Development Goals

sustainable resource use  corporate social responsibility

value of life
This book is about economics and the economy. If the previous chapter dealt primarily with the economy (capitalism: its accomplishments and its limitations), this chapter deals with the discipline of economics: what it’s all about, how it relates to other areas of social research, and what its main themes are.

2.1. SCOPE AND METHOD

What is economics? Instead of offering a formal definition, let us consider three important aspects of microeconomics (which correspond to different parts of this course): choice, markets, and public policy. By choice we mean how consumers, workers, firms, etc, make choices: primarily choices regarding what you would think of as economic choices (e.g., whether to buy X or to buy Y), but also choices you might not have thought as economic choices (e.g., how many hours to study for the Microeconomics final). We will be particularly interested in situations when agents (consumers, workers, firms, etc) must choose between alternative uses of scarce resources (such as time and money). Part II of the book is devoted to the study of choice, first in generic terms, then with specific applications to household choices and to choices by firms.

By markets we mean the interaction between buyers and sellers who transact products or services: a cup of coffee, help designing a
website, shares in Apple computer are three of a host of possible examples. Part III of the book deals with markets: supply, demand, and the interaction between supply and demand. We will be particularly interested in understanding how prices are formed and whether — or in what conditions — markets work well (for a precise definition of “well”).

Finally, by public policy we mean how the government (or government-like institutions) impose constraints on consumers, firms and markets to address failures in the system. Examples of economics-related public policies include merger policy (i.e., deciding whether firms A and B should be allowed to merge), carbon taxation (that is, taxing activities that generate greenhouse gases), or welfare transfers to the poor (e.g., green stamps). Broadly speaking, Parts IV and V of the book focus on public policy. Part IV is devoted to market failure (i.e., cases when a market left to itself fails to work well). Part V, in turn, is devoted to issues of social justice (e.g., cases when markets work well but not for all).

The reader may have noticed that I have used the terms “economics” and “microeconomics” interchangeably. The time has come to be more precise. The economics discipline is normally divided into two main branches: microeconomics and macroeconomics. Macroeconomics is the branch of economics that deals with aggregate economic variables, such as GDP growth, interest rates, unemployment, inflation. By contrast, Microeconomics is the branch of economics that deals with behavior of individual economic units — consumers, firms, workers — and the markets that these units comprise. The previous chapter started off with a macro perspective on the economy, namely GDP as a measure of economic activity. However, the rest of the chapter — and the rest of the book — will focus on the micro approach to economics.

MODELS AS MAPS

Methodologically, economic analysis is based on a series of theories, and theories are primarily expressed in the form of economic models. An economic model is a description of an economic situation by means of words, diagrams and mathematical expressions.

It helps to think of models as maps. Consider the following example, illustrated by Figure 2.1. You are currently at the Village Van-
2.1. SCOPE AND METHOD

Finding your way from the Village Vanguard to the Blue Note

guard and want to find your way to the Blue Note (both are Jazz clubs in Greenwich Village). Figure 2.1 displays two different images: a satellite photo (left) and a stylized map (right). The photo is clearly a more realistic depiction of reality, whereas the map is a rather stylized one. In other words, you might say that the satellite photo has more information, whereas the map is missing a lot of the details of the reality in question.

In spite of having less information, the map is a more useful tool for finding your way. In fact, it is more useful precisely because it has less information: it allows you to focus on what is essential (in particular, the streets, which are represented as “unrealistic” lines). The same is true of a good economic model. It abstracts from a number of details and zooms in on the essential aspects of the economic setting in question. This does not mean that all economic models are good, just as there are some really bad and unhelpful maps.

Economic models, like maps, are simplification of reality, focusing on the most relevant elements and how they relate to each other.

There are many other dimensions of simplification we will be engaging in throughout the course. Particularly important is the choice of models with two firms, two countries, two consumption goods, etc. Obviously there are more than two firms, countries, consumers, in the world. Our focus on only two is done for the purpose of under-
standing the main effects (and, in many cases, being able to represent them on a graph).

The level of abstraction in economics varies. Alternatively to maps, we can think of economic models as construction blueprints. Sometimes what we need is a rough hand-drawn draft (for example, to provide a broad idea of a building design concept). By contrast, sometimes we need a very detailed blueprint, to make sure the constructor follows the architect’s design as closely as possible. The same is true in economic modeling. Sometimes a brief description in English suffices to present an economic idea. By contrast, sometimes we need a detailed mathematical description of a model’s “moving parts”.

Modern economics uses a lot of math. The advantage of mathematics is that it allows for a more precise definition of concepts and relations. Natural language is full of nuances which are great for poets and punsters but can be rather annoying for economists. The term “capitalism”, introduced in the previous chapter, is a case in point. Words can be very loaded, which sometimes is a good thing but at other times leads to more confusion than clarity.

Unfortunately, as is the case in many other contexts, a good tool can be poorly used. Not infrequently, poorly written mathematical models create more confusion than clarity. Moreover, the mathematics barrier can be an exclusion factor for would-be economists with important ideas to contribute.

Reacting to criticisms of excessive “mathiness” in economics analysis, a recent trend has placed greater weight on data analysis: let the data speak. This has been aided by two important developments: a significant increase in the amount of available data (partly a result of the growth of online commerce), and an increased reliance on data from experiments, both laboratory experiments and field experiments (i.e., experiments in a real-world context). However, reliance on data without theory can be very limiting. How can we distinguish between correlation and causality unless we have a good model of how the world works? How can we generalize data from a special case (e.g., ten small factories in India) to the rest of the world unless we have a good model?

A related problem is that of out-of-sample prediction. How can we predict the demand for a new Hyundai model given that we only have data from old Hyundai models? How can we predict the effect
2.1. SCOPE AND METHOD

In sum, data needs the support of theoretical frameworks if it is to be useful for decision-making. But economic modeling that is not based on evidence — or at least that is not checked by evidence — is a dangerous endeavor, and most likely a failed one.

CORRELATION AND CAUSALITY

One of the most common mistakes in analyzing and presenting data is to confuse correlation with causality. Consider the two time series plotted in Figure 2.2, sociology doctorates awarded in the US and worldwide non-commercial space launches. As can be seen, they are highly correlated (during the 1997–2009 period). Does that mean that one causes the other? Clearly not. If you have the time and want to have some fun, I recommend you visit the website Spurious Correlations. What this site does is to troll the web for time series and then attempt all possible combinations of pairs of time series. Eventually, you are bound to find some that produces a high correlation. And there are lots of them, some ridiculously funny.

Here’s a more serious example. A few years ago, the New England Journal of Medicine, one of medicine’s leading journals, published an article on the correlation between per-capita number of Nobel laureates and per-capita chocolate consumption. It was not intended as a
2.1. SCOPE AND METHOD

Chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations. The principal finding of this study is a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates in various countries. Of course, a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism.

What’s wrong with above paragraph? What’s wrong is that it misses the fourth and most likely possibility, namely that the sample considered in the study is much too small. Perhaps the authors tried a bunch of variables (Brussel sprouts, pineapple, soy sauce, etc) and chocolate came out as the best correlation.

It is good to hold a healthy skepticism with regard to correlations. Don’t jump into the conclusion of causality too easily. This includes some of the correlations we’ve considered so far in the book, for example capitalism and growth or human CO$_2$ emissions and climate change.

Another instance of the correlation / causality confusion is what economists refer to as the post hoc fallacy. There’s a Latin phrase, “Post hoc ergo propter hoc,” which in English means, “after this, therefore because of this.” Politicians are particularly prone to this fallacy: “I became president and GDP grew by 3%, that’s how good a president I am.” Which of course ignores that economic growth results from hundreds of factors, very few of which are under the president’s control. “After this but not necessarily because of this,” that should be our motto. This provides a natural segue into the next topic, where we provide general guidance for uncovering causality relations.

COUNTERFACTUALS

One way to avoid falling into the post hoc fallacy is to perform what economists and historians refer to as a counterfactual. Suppose that, at the beginning of December, Luigi’s lowered the price of its famous
jeans. Revenues in November were $80 million, whereas revenues in December were $113 million. It seems that lowering the price was a good idea, right? Not necessarily: After this but not necessarily because of this! In fact, if you are in the business you should now that, with the holiday season, jeans sales tend to increase in December even if you don’t change prices.

As it happens, Luigi’s product manager has access to data on the sales of Kabral’s jeans, a brand for which price remained the same in November and December. Their sales were 44 in November and 67 in December. With all of this data at hand, what can we say about the price-cutting strategy? One was to answer this question is to construct a counterfactual of December sales with no price change. In other words, given the data we have for Kabral’s sales, what do we expect Luigi’s December sales would have been had Luigi’s not changed its price? Kabral’s December sales were greater than their November sales by a factor of $\frac{67}{44} \approx 1.52$ (that is, a 52 percent increase). Assuming that the seasonal variation in jeans sales is the same for Luigi’s and Kabral’s, our counterfactual estimate of Luigi’s December sales would be 80 (November sales) times 1.52 (November-December change factor given no price changes). This implies a counterfactual of 121.6.

We conclude that, while Luigi’s sales increased from 80 to 113, sales would have increased to 121.6 had price not decreased. We thus estimate that the causal effect of decreasing price was to decrease revenues by $121.6 - 113 = 8.6$. The problem is not that revenues did not increase, rather that they did not increase by as much as we would have expected them to increase had we not played with price. Note that the identifying assumption — that is, the assumption that allows us to identify the causal effect of the price decrease — is that the seasonal variation of sales at Luigi’s is the same as at Kabral’s. Not a perfect assumption but probably better than just comparing November and December sales.

For a different application of the counterfactual strategy, I suggest you look at problem “COVID-19: Sweden vs Norway” in the companion book, *Exercises in Microeconomics*. The issue of identification of causality is so important that we will next look at it from a slightly different perspective.
ALL OTHER THINGS EQUAL

Unlike chemistry and other so-called hard sciences, economics does not have the luxury of controlled laboratory experiments. (You can perform behavioral experiments in the laboratory, but not with the level of precision and generality that chemistry and other hard sciences can attain.) The beauty of a laboratory experiment is that one can control all of the relevant variables and then, by changing the value of X only, study the effect of X on Y (for example, X could be temperature and Y pressure). Economics is a behavioral and social science. The phenomena it studies involve multiple economic agents interacting in the most complex ways. For this reason, our best hope is to find real-world data that allows us to effectively keep all variables constant except for X.

As a specific example, consider an empirical study of the relation between hours of study and course grade (sneak preview: it’s a positive relation, so keep studying). The research analyzes the habits and performance of 84 Florida State University students. For each student, researchers determined how many hours the students studied, as well as other variables. They also recorded the students’ course grade. One question we may ask is whether hours of study are related to grade. A first possible answer is given by the following table:

<table>
<thead>
<tr>
<th></th>
<th>high study time (42 students)</th>
<th>low study time (42 students)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average GPA</td>
<td>3.43</td>
<td>3.36</td>
</tr>
</tbody>
</table>

In other words, if we divide the students into two groups (high study time and low study time) then we observe that students who study for more hours receive an average grade of 3.43, higher than the average of the group of students who study fewer hours (3.36). The difference seems rather small and may suggest that there isn’t much of a causal relation between study and grade.

However, as we saw in the previous subsection, one must be careful not to confound correlation with causality. Sometimes we find a correlation which does not correspond to any particular causal relation. The opposite is also possible: although there is no simple correlation between two variables, they may actually be related by a causal link. To see how this is possible, consider the following break-
down of the data:

<table>
<thead>
<tr>
<th>Study Environment</th>
<th>High Study Time</th>
<th>Low Study Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>3.63 (11 students)</td>
<td>3.43 (31 students)</td>
</tr>
<tr>
<td>Poor</td>
<td>3.36 (31 students)</td>
<td>3.17 (11 students)</td>
</tr>
</tbody>
</table>

We now observe that most of the students who spent many hours studying did so in a poor study environment, whereas most students who spent few hours studying did so in a good study environment. This implies that in the first table we were essentially comparing apples with oranges, that is, not correcting for sample selection effects. Students who spend more hours studying do not get a much higher grade because, in addition to spending more hours studying they are also more likely to do so in a poor study environment.

This is one of the greatest challenges of social science: When we compare two situations, there are many dimensions on which they differ, and as a result it is very difficult to assign the variation in $y$ (e.g., grade) to a particular change in $x$ (e.g., hours of study). In this regard, one of the methodological goals of economics is to proceed according to the ceteris paribus principle. Ceteris is Latin for things (as in, for example, “et cetera” or simply “etc.”); and paribus is Latin for similar (as in, for example, “parity” or “comparable”). So, in plain English we would say “all other things equal.” To return to our example, we would like to know the effect of increasing hours of study keeping all other things equal (including, in our particular example, the study environment).

In practice and in research studies like the present one, economists typically utilize multi-variate regression analysis to take into account the multiple factors that vary along with the number of hours of study. Students who are interested in this type of data analysis are encouraged to take the course ECON-UB.251: Econometrics or a similar one. For now, suffice it to say that the research study in question concluded that, all other things equal, one extra hour of study is associated with a grade increase of 0.24. This may not seem a lot, but is certainly a lot more than the 0.07 suggested by the first table above.
POSITIVE AND NORMATIVE ANALYSIS

Continuing our overview of economics methodology, we now come to the important distinction between positive and normative analysis. **Positive analysis** corresponds to statements that describe a cause-effect relationship, that is, statements that predict future behavior or explain past behavior. In other words, statements about what is. By contrast, **normative analysis** corresponds to statements about what ought to be. In other words, normative analysis consists of, and is often supplemented by, value judgments.

---

Positive analysis corresponds to statements that describe what is.
Normative analysis corresponds to statements about what ought to be.

---

Take for example the relation between CEOs and shareholders. A positive statement would be that CEOs make choices so as to maximize firm value, specifically the total value of present and future revenue streams. (Economists disagree on this.) A normative statement would be that it is the CEO’s duty to maximize firm value. (Economists also disagree on this.)

The distinction between positive and normative analysis is important in a number of ways. For example, economist Milton Friedman remarked that

Two individuals may agree on the consequences of a particular piece of legislation. One may regard them as desirable on balance and so favor the legislation; the other, as undesirable and so oppose the legislation.

In other words, keeping the positive/normative dichotomy in mind can be helpful to knowing when to agree and when to agree to disagree.

At a more fundamental level and as we will see in the next section, the central model of economic behavior assumes that agents — firms, consumers, workers, etc — are self-interested optimizers, that is, they try to do the best for themselves. This is a positive statement, that is, a set of predictions about the actual behavior of people and organizations. In subsequent chapters in this book we will examine the extent to which this paradigm of economic modeling is correct.
and useful. For now it suffices to remark that, with the exception of a small fraction of the economics profession — those whom we might describe as Ayn Rand followers — we take the self-interested model as a positive statement, not a normative one (the world would be a better place if we cared more about others than we do — and now you know I am not an Ayn Rand follower).

One final note regarding positive economics: Many economists have come to regard the positive approach as a goal — almost an ideal — that the economics discipline should aim for. As Milton Friedman put it

> The ultimate goal of a positive science is the development of a “theory” or “hypothesis” that yields valid and meaningful ... predictions about phenomena not yet observed.

This perspective on economics has largely been influenced by what some refer to as physics envy: the aspiration of turning economics into a hard science independent of the value judgements associated with normative economics. Many other economists — including this one — think this is not only impossible but also undesirable. Can a conservative economist be neutral when writing about economics and gun control? Can a liberal economist be neutral when writing about economics and abortion? Even if research is based on “objective” statistical analysis, it’s almost inevitable that researchers are more likely to keep results that better confirm their prediction or expectation. This bias can be partly resolved by replication studies (extremely rare in economics) or by asking researchers to pre-announce their research design (also extremely rare). More important, even if we are able to correct for these potential biases there would remain a fundamental one, namely the choice of a research topic.

In sum, the distinction between positive and normative economics is helpful. The ideal of a neutral, value-free economist is probably not. There is only one thing worse than a biased economist, and that is a biased economist who is convinced he or she is neutral.

### 2.2. BEHAVIORAL AND SOCIAL SCIENCE

Economics is both a behavioral and a social science. In other words, it deals with human behavior — how humans make certain types
of decisions — and with social interaction among individuals, especially in the context of firms and markets. As such, economics relates to other behavioral sciences (e.g., psychology) as well as to other social sciences (e.g., sociology or history).

As a behavioral science, economics is based on a model known as homo economicus (Latin for “economic man”). This model portrays humans as agents who are consistently rational and self-interested, specifically agents who act with the purpose of optimally pursuing their subjectively-defined ends. If you have any knowledge of human behavior — including your own — you know this is not entirely realistic, to put it mildly. We are, at best, boundedly rational. In fact, research by psychologist Daniel Kahneman suggests that a large fraction of our decisions are based on mental shortcuts or rules of thumb. This is not necessarily inconsistent with the homo economicus model. In fact, to the extent that decision making implies costs — gathering information, processing data, etc — one would expect a rational decision-maker to develop simplified decision rules. That said, there are a number of instances where the homo economicus model seems to fail, and psychology provides a helpful framework to understand these.

**PSYCHOLOGY**

As a behavioral science, economics benefits from the influence of related behavioral sciences. The field of behavioral economics studies the effects of psychological factors on the behavior of individuals and institutions, in particular to the extent that it differs from the behavior predicted by the rational economic behavior model. One classical example of such deviation is given by decisions regarding inter-temporal trade-offs. Suppose I give you the option of one bag of M&M now or two bags of M&M tomorrow. Most people will go for one bag now. Nothing wrong with that, an economist might say: it’s simply that people have a strong preference for consumption in the present. Now suppose that you’re given the choice between one bag of M&M one year from now or two bags one year and one day from now. Faced with this trade-off, most people will go for two bags in one year and one day. But if you put yourself in the shoes of you one year from now you will effectively be picking two bags of M&M “tomorrow” over one bag “today”.
A second setting where behavior tends to deviate from the homo economicus prediction relates to risk-return trade-offs when there are small probabilities of high-payoff events. For example, wearing seat belts imposes a cost on a lot of people, who find it rather uncomfortable. This cost, which is visible and easy to predict, must be traded-off against the benefit of wearing seat belts in case an accident takes place, a rather low-probability event. Objectively speaking, that is, considering the enormous benefit of wearing seat belts when an accident takes place, buckling up is the rational thing to do. However, left to their own will many drivers and passengers prefer — or at least choose — not to wear seat belts.

There is a certain pattern to these failures to behave according to economic rationality. These are situations involving long time gaps or large payoffs with low probability. This suggests that while the economic model is reasonably appropriate for common and repeated agent decisions, it may falter when applied “on the boundaries” of agents’ economic decisions. The message then is that economists should beware of the boundaries of applicability of their model of rational decision making.

An analogy can be made with respect to Newtonian mechanics. For centuries, this branch of physics has been of extraordinary use in a variety of applications. The deterministic predictions of Newtonian mechanics are fairly accurate when it comes to objects at a “human” scale: cars and bikes, projectiles, chairs and desks, buildings and bridges, etc. However, when we move to the subatomic level, Newtonian mechanics fail miserably. By no means does this imply that Isaac Newton’s work was in vain, only that his laws must be interpreted and applied within a limited scope.

**SOCIOLGY**

Psychology contributes to economics by enriching our view of the cognitive limitations and biases of decision makers. By contrast, sociology deals with the patterns of social relationships, social interaction, and culture that surrounds everyday life, including economic life. Consider the example of work and employment. From a mainstream economics point of view, labor is a production factor — i.e., part of the process of transforming inputs into outputs — as well as an income source. The labor market brings together the firm’s de-
mand for labor and the individuals’ labor supply, where the latter results from an optimization process involving income, leisure and other factors.

Although theoretical and empirical economic analysis allows for “other factors”, in practice these “other factors” are not given as much importance as they should. For example, Section 1.4 includes the quote that “one has a story about one’s life that involves one’s job, and a job defines how I am important and why people should love me.” The issue of social status plays an important role in sociology, less so in economics.

One’s job is by no means the sole source of social status. In his classic, *The Theory of the Leisure Class*, economist and sociologist Thorstein Veblen — an early critic of the capitalist system — argues that social status is largely acquired by means of conspicuous consumption, that is, consumer spending on luxury goods and services to publicly display economic power rather than — or in addition to — enjoying them for their intrinsic value. Positional goods — a refinement on the concept of conspicuous consumption — are valued in terms of relative consumption. Here’s an interesting thought experiment:

You must choose between ... world A, in which you will live in a 4,000-square-foot house and others will live in 6,000-square-foot houses; and world B, in which you will live in a 3,000-square-foot house, others in 2,000-square-foot houses. ... If only absolute consumption mattered, A would be clearly better. Yet most people say they would pick B.

Again, we find that

Social interaction plays an important role in an individual’s actions and value judgements.

In the above example social interactions takes the form of relative consumption comparisons. More generally, from an economics point of view, understanding these sociological patterns is important in order to explain economic behavior. It may also have specific policy implications. For example, several authors argue that luxury taxes
— and more generally taxes on positional goods — may be a way of addressing the “arms race” of conspicuous consumption.

Gender and racial discrimination is another instance where economics can benefit from neighboring social sciences. Traditionally, economists have classified discrimination in two different “bins”: statistical discrimination and taste discrimination. An example of statistical discrimination is provided by car insurance rates. A 16-year-old woman pays an average 6-month premium of $3,378, whereas a 16-year-old man pays a higher $3,897. The fact that women are charged less does not reflect any animus against men, simply the fact that, statistically speaking, 16-year-old men are more prone to accidents than 16-year-old women. Similarly, the fact that a 24-year-old man pays on average $1,381 — substantially less than his younger counterpart — does not necessarily reflect age discrimination in the sense that the expression normally has. By contrast, the differential way that racial minorities are treated in the labor market frequently goes beyond statistical outcome expectations, rather reflecting a distaste for hiring minorities.

The distinction between taste and statistical discrimination is helpful and has guided much of the economics research. However, many criticize the approach as incomplete. Sociologists, in particular, place a greater weight on institutional discrimination, the idea that differential treatment by race is also — perhaps mainly — perpetrated by organizations if not codified into law. We will return to these issues in Section 11.2, where we will show how pervasive all of the above types of discrimination are.

POLITICAL ECONOMY

At some level, France, China and the US may all be characterized as societies based on the capitalist system. However, one may argue that the differences are greater than the similarities. More broadly, we observe considerable differences across countries, both in terms of their economies and in terms of their political systems. One of the most fascinating research questions — for economists and scholars in general — is why such disparities exist. In 1960, Argentina’s per capita GDP was 7.3 bigger than South Korea’s. By 2018, South Korea’s GDP was 2.6 times bigger than Argentina’s. How does one explain such wide variation in outcomes? In a famous 1992 memo,
one of Bill Clinton’s strategists reminded campaign workers to stay on message: “It’s the economy, stupid.” Clearly, economic conditions play an important role in addressing questions such as the Argentina-South Korea gap. For example, in Section 1.2 we saw how history’s “hockey stick” explains a considerable portion of the cross-country variation in economic outcomes. However, economists have come to believe that

Institutions — in particular political institutions — play a central role in a country’s development.

Some argue that economic and social development requires social orders “to control violence, provide order, and allow greater production through specialization and exchange.” These orders may be supplied by a political system that limits its abuse or, in a small number of more recent cases, by means of political competition (that is, free entry into economic and political organizations). In this context, failed nations result from extractive economic institutions, “which destroy incentives, discourage innovation, and sap the talent of their citizens.” Examples include lack of property rights, absence of rule of law, and many other.

It’s not only political institutions that matter. In old and modern societies alike, non-governmental institutions — churches, charities, political action groups — have played an important role. Witness for example the impact that Greta Thunberg has had on the climate change debate, or the Gates Foundation on public health.

ECONOMIC HISTORY

Spanish philosopher George Santayana pithily observed that “those who cannot learn from history are doomed to repeat it.” (The actual quote is slightly different but the idea is the same.) The study of history plays an important role in general, and in particular in the study of economics. Unlike hard sciences such as chemistry or biology, social sciences such as economics lack the ability to test theories in the lab. (This is not entirely true — see p. 50 — but is largely true.) In this context, the study of history can be of great help. In fact, many questions that were relevant in the past remain relevant in our day, and the lessons from the past can be lessons for today.
2.2. BEHAVIORAL AND SOCIAL SCIENCE

The introduction of copyright protection in 19th century Italy provided a big boost to the creation of new operas. Can this tell us something about the current debate on extending the period of copyright protection?

Take for example the issue of copyright protection, which gives its owner the exclusive right to make copies of a creative work (e.g., a book, a song, a movie, etc). There is considerable cross-country variation in copyright duration. For example, in the US and Europe copyright lasts until 70 years after the creator’s death; in Canada, 50 years after the creator’s death; in Mexico, 100 years after the creator’s death. Which one is right? Does it make a difference? Specifically, do the incentives for artistic creation depend on the existence and the duration of copyright protection? Lest the reader may think this is an irrelevant issue, suffice it to say that the Copyright Term Extension Act (CTEA) of 1998, which extended copyright terms in the United States, was a highly controversial and hotly debated law.

Can history help address the above questions? Economic historians Michela Giorcelli and Petra Moser describe how Napoleon’s military victories in Italy in the late 1700s effectively introduced copyright law into various Italian states. This led to a significant increase in the number and the quality of operas written by Rossini and the other great composers of 19th century Italy. However, when the duration of copyright protection was extended beyond the creator’s life, there were no additional effects to speak of. In sum, history suggests that

Some copyright protection goes a long way into incentivizing an author to create, but extending copyright protection beyond the author’s death does not seem to provide any additional gain.

Had copyright protection not been extended in the US (in 1998) from 50 to 70 years after the creator’s death, Walt Disney’s Mickey Mouse
would by now be in the public domain. Would the benefits from open access outweigh the costs in terms of creativity incentives? History suggests the answer is yes.

Consider now a current and controversial topic in the news: immigration. The main economic rationale for immigration quotas is that it protects domestic employment. Against this alleged positive effect we must take into account the human talent that is lost by preventing would-be immigrants from entering the country. Can history tell us anything about it? Research by economic historians Petra Moser and Shmuel San suggests the negative effects of immigration quotas can be significant. Their view is based on an important historical precedent: Between 1921 and 1924 the US first adopted immigration quotas for “undesirable” nationalities, so as to stem the inflow of Eastern and Southern Europeans (ESE). It is estimated that, due to these quotas, 1,170 ESE-born scientists were missing from US science by the 1950s. This in turn led to a 68 percent decline in patenting in the fields where ESE immigrants researched. Moreover, these effects were still felt well into the 1960s. This historical evidence confirms the relative consensus among economists that

*Immigration quotas, especially when applied to skilled immigrants, produce significant harm to the domestic economy.*

**GAME THEORY**

Economic agents — buyers and sellers, employers and employees, households, firms and countries — frequently act in relation to other agents. Economics is a social science. Economic behavior is not only a matter of optimizing with respect to a series of constraints; it is also a matter of forming beliefs regarding other agents’ behavior, predicting how our behavior will influence theirs and how best to react to their behavior, etc.

The field of game theory — long associated with economics — provides useful tools for the analysis of such situations of interdependent behavior. As a motivating example, consider one of the biggest problems of our time: climate change. Although this is not strictly an economics problem, the game-theory approach can be useful in understanding the fundamental issues. In essence, climate change is a social dilemma: Whenever there is a common resource, individual
choices, good as they may be from each individual’s point of view, may result in an inferior collective outcome.

Consider the following simplified depiction — a model — of the climate change social dilemma. Suppose there are only two countries in the world, the US and China. Each country must determine how much effort it will put to reduce greenhouse gas emissions. For simplicity, assume the US and China can choose a high effort level or a low effort level. The outcome (i.e., the payoffs received by each country) depends on the choices by both players. In fact, this is the essence of interdependency: my fate does not depend on my actions only.

Figure 2.3 illustrates the climate change dilemma, where the US’s choice is listed in rows and China’s choice is listed in columns. If both the US and China opt for a high effort, then the dangers of global warming and climate change are avoided. By contrast, if one of the countries chooses to go for a low effort, then such a country gets a relatively good outcome: it benefits from the other country’s effort to reduce greenhouse emissions without needing to pay the cost. Finally, if both countries go for low effort, then both are plunged into the undesirable outcome of extremely high risk of global warming and climate change.

The next step is to measure these outcomes quantitatively. Economists and climate scientists have build complex dynamic models to estimate the effects of different policies over the next decades. Fig-
FIGURE 2.4
The climate change game

Figure 2.4 summarizes the dilemma faced by the US and China in terms of quantitative payoffs. For each combination of a choice by the US and a choice by China, the matrix in Figure 2.4 shows the payoff for the US and the payoff for China. We follow the convention of marking the row player’s payoff in the lower left corner and the column player’s payoff in the upper right corner. Moreover, we will assume that each player’s goal is to maximize their individual payoff.

Notice that the US’s ultimate payoff depends on its choices as well as on China’s choice. Specifically, the model in Figure 2.4 corresponds to a game. Game theory is a branch of economics dealing with behavior when outcomes depend on the interplay of various agents, who we refer to as players. One goal when using games, similarly to other models, is to understand and predict behavior. What outcome would we expect to unfold in a situation like the one depicted in Figure 2.4?

A first observation is that the US has a dominant strategy: If China chooses high effort, the US is better off by choosing low effort. If China chooses low effort, the US is again better off by choosing low effort. So, regardless of what China does, the US is better off with low effort. More generally,

A strategy $s$ for player $a$ is a dominant strategy if it yields a higher payoff than any other strategy independently of player $b$’s strategy.

Moreover, if $a$ has a dominant strategy, then all other strategies are said to be dominated strategies. What’s the point of mentioning dominant strategies? Simple: One expects rational players to choose dominant strategies (if they exist). So, in the present case we would expect the US to choose low effort. Looking now at China’s payoffs,
we conclude that, like the US, it too has low effort as a dominant strategy. Putting it all together, we would expect both the US and China to choose low effort, which is bad for both the US and China.

The above simplified analysis provides an example of a game. A *game* is a model which combines four elements:

- Players (in the previous example, the US and China)
- Rules (in the previous example, simultaneously choose effort level)
- Strategies (in the previous example, high or low)
- Payoffs (in the previous example, an index of national welfare that includes both economic and climate-related outcomes)

Figure 2.4 illustrates a fundamental and recurrent social dilemma: both countries are better off (collectively) if both choose high effort, but each country is better off (unilaterally) by choosing low effort and free riding on the other country’s effort. Although this is a specific game and a specific situation, it is sufficiently common that we give it a special name: the *prisoner’s dilemma*.

> *In the prisoner’s dilemma game, each player chooses their dominant strategy but the resulting outcome is the worse for both players.*

At first blush, this seems a contradiction: if each player chooses what’s best for them, how can the outcome be the worst for them? That is the reason why models such as games can help clarify and understand otherwise confusing and complex social interaction events.

To conclude this subsection, note that not all instances of social interaction have the nature of a prisoner’s dilemma. In fact, Chapter 7 presents one of the most important results of microeconomics, stating that, under certain conditions, the result of each agent’s choosing their optimal strategy is an outcome which is globally optimal (in a specific sense). Essentially, this result — known as the First Welfare theorem — corresponds to Adam Smith’s “invisible hand” metaphor (cf Section 1.2). One of the biggest debates in economics and politics is precisely the relative relevance of this theorem vis-a-vis the prisoner’s dilemma. Both have a point to make. Ultimately, it’s a case of glass half full versus glass half empty.
2.3. CENTRAL THEMES

Although there are still many chapters to come where to talk about economics and economics principles, in this section we will cover some of the more central themes in microeconomics. Think of it as a sort of trailer for Chapters 3–7, where we cover the core of microeconomic theory. All of this implies some repetition, namely economics concepts that appear more than once throughout the book. As they say in computer lingo, “it’s not a bug, it’s a feature,” (meaning, it’s on purpose, not by mistake). Some of these economics concepts are not obvious or immediately intuitive, and repetition can be very helpful.

MARGINAL THINKING

As often happens with other disciplines, economics is more than a series of ideas: it’s a way of thinking about things. Non-economists may have noticed it — and be annoyed by it — when talking to economists. First, economists have a tendency to think about problems strictly in terms of costs and benefits (monetary or otherwise), i.e., they tend to follow a cost-benefit approach. For example, on May 2019 Madonna’s Madame X Tour was officially announced, with concerts starting in September 2019. In this context, one important decision for Madonna is how many tour performances to include. Any time there is a “how many” question, economists think about it in terms of marginal variations. Suppose the current plans call for 84 dates (it did as of January 2020). Is this an optimal number of concerts? To answer the question, Madonna should evaluate the benefits (monetary and otherwise) from an additional concert, both in terms of ticket sales and other related revenue stream. This should then be compared against the cost (monetary and otherwise) of an additional concert. If cost is greater than benefit, go for it; if not, don’t.

But there is more: Suppose that Madonna really thinks like an economist. Then it must be the case that 84 concerts is the number such that the benefit of an additional concert is about the same as the cost of an additional concert. Why? Well, if the benefit of the 85th concert is greater than the cost, then Madonna should extend the tour to 85 dates (at least). By contrast, if the benefit is lower than the cost, then it’s also likely the case that the benefit of the 84th concert itself was lower than the cost, in which case Madonna would be better off
2.3. CENTRAL THEMES

How many shows should Madonna include in her Madame X tour? She should think at the margin, an economist would say. By shorting the tour to 83 dates (or less). We will return to this in Chapters 3, 4, 5, and 6 (at least). It’s that important!

The level of a given economic activity should be increased if and only if the additional benefit exceeds the additional cost. The optimal level of economic activity is such that the additional benefit is very close to the additional cost.

SUNK COST AND OPPORTUNITY COST

The Airbus A300-600, better known as the Beluga, is a modified Airbus aircraft used to transport large payloads. Its creation follows the needs of a company with geographically dispersed production facilities: Although the Airbus assembly plants are located in Toulouse and Hamburg, major parts (including wings and landing gear) are manufactured in Airbus consortium plants located in Spain, Britain, and Germany. In addition to airplane parts, the Beluga can also be deployed to carry cargos owned by third parties. Examples of past loads include United Nations helicopters, sections of the International Space Station, and a collection of 17th-century paintings in an environmentally controlled container the size of a small house.

The Beluga is an input (one of many inputs) used in the construction of Airbus passenger planes. (Note a possible source of confusion: one plane — the Beluga — is an input in the production of another plane, an Airbus commercial plane.) Suppose the Beluga costs $D$ in development costs (specific to the Beluga). Moreover, leasing of a Beluga yields an average revenue of $L$ per year. Based on this infor-
2.3. CENTRAL THEMES

The Airbus A300, also known as Beluga (in the picture, the A330-743L version, a.k.a. Beluga XL).

When it comes to cost accounting, it is common to distinguish the accounting perspective from the economics perspective. The typical accounting approach would be to amortize the cost of developing the Beluga and include that value in the production cost of an Airbus passenger plane. By contrast, the economics approach is to make the distinction between sunk cost and opportunity cost. A **sunk cost** is a cost that will be paid independently of the current course of action. An **opportunity cost**, by contrast, refers to something that needs to be given up in order to obtain a certain good $x$.

A sunk cost in an accounting cost but not an economic cost. An opportunity cost is an economic cost but not an accounting cost.

(The above is a bit of a caricature of the accounting approach, though it’s fair to say the accounting approach tends to be more history focused, whereas the economics approach tends to be forward looking.) Going back to the Beluga example: By the time Airbus makes decisions regarding its passenger planes (e.g., what price to sell them), the cost of developing the Beluga, $D$, is a sunk cost: no matter what price decisions are made regarding the passenger plane, Airbus will have to pay the Beluga development cost. By contrast, there is a cost which an accountant would ignore but an economist would not: the opportunity cost $L$ of employing the Beluga to transport airplane parts as opposed to providing services to third parties. So, in
Economic theory states that sunk costs should not matter in decision making (a normative statement). Do they matter in actual decision making (a positive statement)? Baseball provides an example of when they do, in which case we say the decision maker suffers from a sunk cost “fallacy”. Here’s the story: Chris Davis, a 33-year-old slugger for the Baltimore Orioles, went through a long stretch of hitless at-bats in 2018, the worst batting average in major league history (.168). Still, the Orioles manager kept fielding him. According to an article,

In truth, the decision to keep playing Davis almost certainly has more to do with his $17 million salary this year and the $93 million the Orioles owe him beyond 2019 in salary and deferred payments, which will have the team sending him paychecks through the 2037 season.

The point is that the money the Orioles owe Davis is a sunk cost. It will be paid regardless of whether he plays or not. As such, it should be irrelevant to the decision of whether to field Davis. The fact he was fielded in spite of his poor performance suggests that not all behavior is consistent with basic economic thinking. This may not surprise you, but it continues to puzzle economists, especially in situations when the stakes are high.

terms of the above notation, the economic costing of Airbus passenger planes should include $L$ but not $D$; whereas the accounting-based costing would include $D$ but not $L$.

FEASIBLE SET

There are many ways of defining economics. Some state that “economics is about making choices,” which feels like a very broad definition, perhaps a little too broad. Others put a stress on the “relationship between ends and scarce means that have alternative uses.” The concept of feasible set, which provides an apt segue to the concept of opportunity cost, is the natural reflection of the scarcity-focused def-
inition of economics. The **feasible set** is simply the set of outcomes that are attainable. This sounds like a tautology, but you would be surprised by how often people — in particular politicians — ignore this basic fact of life: you cannot have everything.

For the sake of concreteness, let us consider a current and highly controversial issue: the extent to which — and speed at which — we should “re-open” the economy as we recover from a pandemic such as COVID-19. **Wealth vs health**, **lives vs economics**, **health crisis vs economic crash**, etc. — with these or similar terms, the media raises the question of what the optimal policy is. **One expert** states that

> If you keep the shutdown going for 2 months more than we need to, that’s just an unbelievably costly mistake. …
> If we lift the shutdown 2 months too soon, that would be an unbelievably costly mistake.

This is neither reassuring nor informative. A more useful tool might be to estimate the relevant feasible set. Suppose that we have two main goals: economic growth and saving lives. Figure 2.5 measures the economy’s growth rate on the horizontal axis and the number of lives saved on the vertical axis, where \( h \) represents the entire population. The region shaded blue represents the feasible set: the set of combinations of economic growth and lives saved that is feasible. Notice in particular that \( h = \bar{h} \) is never part of the feasible set: it is
impossible to save all lives, regardless of how much we focus on saving lives. The best we can do is to attain point $E_1$, where $h_1$ lives are saved and economic growth is at its lowest level (probably a highly negative value). At the opposite extreme, we could push economic growth to the max but then very few lives would be saved (point $E_2$).

It seems clear that $E_2$ is not an acceptable outcome. It is not as clear, but equally true, that $E_1$ is likely an undesirable outcome. Once we get close to extreme lockdown levels and duration, the opportunity cost of saving lives by means of an even stricter lockdown becomes very high, possibly prohibitively high. As was recently noted, economic decline itself has an adverse effect on health, for example by increasing suicide rates.

What can we then say about the health vs wealth tradeoff? First, as the feasible set shows, many times there is no tradeoff at all! Going from $E_3$ to $E_4$, for example, is a no-brainer: it improves health and it improves the economy. Specifically, a well-designed system of testing and tracing and protective measures might do the trick. But once all of the “low hanging fruit” has been caught, we still have to make a choice, for example between $E_5$ and $E_6$. This is not something that technical analysis per se can do, it’s a decision we need to make as a society. What economists can do is provide an estimate of what the trade-off is: what the opportunity cost of increasing lives saved from $h_6$ to $h_5$ is in terms of economic growth. Or, conversely, what the opportunity cost of increasing economic growth from $w_5$ to $w_6$ is in terms of lives lost.

In other words, the typical economist will tell you that, on the one hand, we could try to open the economy slowly, which would probably save more lives (point $E_5$), but, on the other hand, we could open the economy quicker ($E_6$), though that would cost us a few more lives. Oral tradition has it that President Harry Truman, when briefed by an economist, pleaded “please give me a one-handed economist!” Alas, once you are on the frontier of the feasible set, all choices are going to be of the “on the one hand—on the other hand” type. In this sense, there are no one-handed economists.

At a May 6, 2020 briefing to discuss the New York post-COVID-19 re-opening plan, Governor Cuomo stated that his “reopening plan doesn’t have a tradeoff.” With due respect, Governor, you just got an F on the first quiz.

Exercise “COVID-19: Sweden vs Norway” (in the companion ex-
2.3. CENTRAL THEMES

exercise book) explores COVID-19-related trade-offs by comparing Sweden and Norway’s radically different approaches. We will return to the concept of feasible set, in an explicit manner, in Chapter 3 and then in Chapter 12, and implicitly at many points in the text. It’s that important a concept!

DECREASING MARGINAL BENEFIT

Some people cannot start the day without a good cup of coffee. In economics terms, we would say that the benefit (also known as utility) you get from a first cup of coffee is very high. If you feel particularly sleepy, or if you like coffee a lot, then a second cup of coffee will still be greatly appreciated. By the time you are drinking a third cup of coffee (and it’s still 8am), you realize you’ve had enough: the benefit you would get from that additional cup of coffee would be very small or even negative. This is a typical pattern in economics, so much so that we have a name for it: the law of decreasing marginal benefit. When it applies to consumers like you and me, we refer to it as the law of decreasing marginal utility. It’s important to distinguish total utility from marginal utility. The utility I get from drinking two cups of coffee is greater than the utility I get from one cup of coffee. However, the utility I get from the second cup of coffee is lower than the utility I get from the first cup of coffee.

When we are talking about firms or individuals producing goods or services, instead of benefits we normally refer to the law of decreasing marginal returns. For example, suppose that I own a restaurant with a kitchen of a certain size. The more workers I hire to the kitchen, the more dinners I am able to serve. However, you can see how the contribution of an additional worker would decline as more and more are added to the kitchen. In fact, at some point the marginal contribution of a worker might actually be negative! Too many cooks spoil the broth, so the saying goes. The concept of decreasing marginal returns will reappear multiple times in this book, in particular in Chapters 3 and 5.

The benefit provided by an economic good in increasing in the quantity of such good, but the benefit provided by each additional unit of the good tends to decline as total quantity increases.
Closely associated with the concept of decreasing marginal returns is the concept of increasing marginal cost. Continuing with the restaurant kitchen example, suppose initially there is one cook and output is 10 plates per hour. If I add a second cook, it’s unlikely they will equal the rate of 10 plates per hour (for the reasons discussed before). It follows that, if I want to increase the kitchen production rate to 20 plates per hour, I may need to go from one cook to three cooks. This means that the cost of getting the second 10 meals is greater than the cost of getting the first set of 10 meals. The concept of increasing marginal cost will prove very important when discussing import tariffs (Section 7.3) and climate change policy (Section 9.2).

**GAINS FROM TRADE**

As mentioned in Chapter 1, Adam Smith observed a fairly obvious fact: Scotland produces excellent wool but horrible wine, and France, by contrast, produces lousy wool and excellent wine. This clearly calls for an exchange of Scottish wool for French wine, thus making both countries happier with what they wear and drink.

Normally an economic transaction creates value for both parties.

All too often we think of transactions as a zero-sum game ("my gain is your loss"). This may be true in chess and other such games, but not in most business situations. In fact, in today’s economy the number of individuals and corporations involved in the creation of "stuff" is relatively small. Most of the modern economy produces services, and within services a big chunk corresponds to bringing together supply and demand. This is true for Walmart, Amazon, and for a multitude of online platforms. Although these individuals and corporations and platforms do not create any new "stuff", they do create considerable value simply by inducing trade. When you buy those jeans or earbuds or whatever, you typically get more than you would be willing to pay; the seller gets more than it cost them to produce; and the platform itself gets a cut in the process. In sum, there is a lot of value that’s just been created, even if no new product was created in the process. We will return to this issue in Chapter 7.
2.3. CENTRAL THEMES

The market value of diamonds is higher than the market value of water, but the value in use of water is higher than the value in use of diamonds.

Paul Sapiano

VALUE IN USE AND MARKET VALUE

There aren’t many economist jokes (certainly not compared to lawyer jokes), but one that comes up frequently is this: An economist is someone who knows the price of everything, but the value of nothing. This is a funny joke, but as an economist I must say it is not entirely fair. First, I should clarify that it all comes from the famous playwright Oscar Wilde. In Lady Windemere’s Fan, he had Lord Darlington quip that a cynic was “a man who knows the price of everything and the value of nothing” (my emphasis). A cynic, not an economist.

Second, and most important, economists have a pretty good theory of value. Consider the so-called water-and-diamonds’ paradox (also known as the paradox of value). Which of the two has greater value: water or diamonds? It really depends on what notion of value you consider. To quote Adam Smith:

The word value, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods which the possession of that object conveys. The one may be called “value in use”; the other, “value in exchange”. The things which have the greatest value in use have frequently little or no value in exchange; on the contrary, those which have the greatest value in exchange have frequently little or no value in use. Nothing is more useful than water: but it will purchase scarcely anything; scarcely anything can be had in exchange for it. A dia-
mondd, on the contrary, has scarcely any use-value; but a very great quantity of other goods may frequently be had in exchange for it.

In other words, the **market value** of diamonds is higher than the market value of water, but the **value in use** of water is higher than the value in use of diamonds. To put it in one sentence, I’d rather live without diamonds than without water, but I’d rather own DeBeers than ConEdison. In Chapter 7 we will return to this concept when discussing whether and how markets are efficient.

---

**Market value refers to the transaction monetary amount. Value in use refers to the benefit received from using a good. The two values are not necessarily related.**

---

**COMPARATIVE ADVANTAGE AND SPECIALIZATION**

David Ricardo, a 19th century economist, read Adam Smith’s argument that trade creates value when different countries have advantage in supplying different products. For example, Scotland is better at producing wool and France is better at producing wine. But when Ricardo applied Smith’s thinking to England and Portugal he was faced with the observation that Portugal produces better wine than England (easy), but is also more efficient at producing textiles (i.e., does so at a lower cost).

Put this way, it does not seem there is much room for value-creating exchange, but there is. The genius of Ricardo is to develop the concept of **comparative advantage**: What matters is not whether Portugal is better than England at producing wine or textiles. What really matters is what Portugal is relatively better at. As it happens — not very surprising — Portugal was better than England at producing textiles but much better than England at producing wine. It follows that the two countries can jointly create value by exchanging English textiles for Portuguese wine.

---

**An economic agent should specialize on the activity for which they have a comparative advantage.**
Car traffic in Auckland, New Zealand. Car fuel efficiency depends greatly on the incentives provided by gasoline prices, which in turn depend on gasoline taxes.

REVEALED PREFERENCE

You should put your money where you mouth is, so goes the popular expression. In other words, your actions should reflect your convictions. One can also think about it backwards: show me your actions and I will tell you what your convictions are. Economists have a version of this: the concept of revealed preference. The idea is simple: If economic agents — consumers in particular — are rational decision makers, then we can infer their preferences from their actions. Suppose, for example, that Lena buys a YouTube TV subscription when the monthly rate is $40 but does not buy a YouTube TV subscription when the price is $60. Then I can infer that Lena’s maximum willingness to pay for YouTube TV lies somewhere between $40 and $60.

This is not rocket science. The economic theory of revealed preference is more complicated than this, partly because it’s not just the subscription price of YouTube TV that changes, lots of things change at the same time in the real world. However, the idea is fairly straightforward:

\[
\text{If we have enough observations of an economic agent’s decisions, then we should be able to infer their preferences with reasonable precision.}
\]

We will return to these issues at various points in the book, in particular in Chapter 6.

INCENTIVES

Many different fields of inquiry are involved in the business of understanding and predicting human behavior. Psychology, for ex-
2.3. CENTRAL THEMES

Founded in 2008, Stack Overflow (SO) is the largest online Q&A platform where programmers ask and answer programming-related questions. As of June 2020, 13 million users had posted nearly 20 million questions. More than 70% of the questions were answered, in many cases by more than one user. Considering the quality and length of many of these answers, it is clear that many hours were spent helping other users, with no direct financial compensation to speak of. Is it a case of altruism, or are there ulterior motives behind private contributions to such a public good?

Affiliated with SO, the Stack Overflow Careers (SOC) site hosts job listings and contributors’ CVs so as to match employers and employees. The information regarding each job candidate includes their employment history as well as various summary statistics regarding their contribution to SO: how many questions and answers were posted, as well as how many positive votes such contributions got from peers.

The weeks leading up to a programmer’s job change show an increase in SO activity, with a sudden drop right after the job change takes place. While there are various interpretations for these upward and downward shifts, statistical research suggests that users increase their SO activity as a means to improve their reputation and thus receive better job offers.

Box 2.2: Voluntary contributions

ample, explores behavior and mental processes, including perception, cognition, attention, emotion, intelligence, and so forth. Sociology, in turn, focuses on patterns of social relationships and social interaction that surround everyday life. The economics approach, as we saw in Section 2.1, is based on the homo economicus framework. Economists think of consumers, workers, firms, etc, as rational agents who choose what’s best for them, agents for whom economic incentives provide an important instrumental cause. To put it in a somewhat crass but nevertheless apt way, we assume that, faced with an economic situation, each agent asks him or herself, “what’s in it for me.”

Consider the example of car fuel efficiency. Motivated by a desire to reduce dependency on oil (1970s and 1980s), or to reduce greenhouse gas emissions (more recent decades), governments have
been keen on increasing car fuel efficiency. In the US fuel efficiency has been pursued primarily by means of regulation (e.g., fuel efficiency standards). By contrast, European countries levy a substantial tax on gasoline consumption. Regulations notwithstanding, US consumers (resp. manufacturers) have little economic incentive to buy (resp. sell) fuel efficient cars. In Europe, the incentive is as high as gasoline taxes are. The results are clear: In 2013, average fuel efficiency in the US was 32 miles per gallon, whereas the EU showed a whopping 45. (For European readers, this corresponds to 5.2 liters per 100 kilometers in the EU against 7.6 liters per 100 kilometers in the US.)

The car fuel efficiency examples illustrates the importance of economic incentives. In particular, it illustrates one of the dearest economics principles:

*Market prices provide crucial information and incentives to decision makers.*

Box 9.2 provides a more in depth look at the gasoline case. Box 2.2 discusses an example where one might have thought economic incentives do no play a role: voluntary contribution to an online site. As it happens, economic incentives do play an important role. The issue of incentives will be pervasive throughout the entire book, beginning with the next section, where we take a critical view at the role played by economics in our current society. A more in-depth discussion of the role played by incentives in economic relations is provided in Chapter 10.

**2.4. A FORCE FOR THE GOOD**

“Blame Economists for the Mess We’re In” — such was the title of a recent controversial and much talked about OpEd. It makes two points about economists and economic policy in the US. First, that economists, who until the 1970s were largely ignored (“they don’t know their own limitations”), gradually took over the reins of public policy. Second, that this resulted in a complete disaster: “The rise of economics is a primary reason for the rise of inequality,” a reference in particular to the free-market economics that inspired the Reagan and Thatcher deregulation waves.
This view misses two points. First, the influence exerted by economists has been quite significant since at least Adam Smith. Chapter 1 includes a famous quote by John Maynard Keynes, namely that “practical men who believe themselves to be quite exempt from any intellectual influence, are usually the slaves of some defunct economist.” And living economists as well, beginning with Keynes himself (“nobody could have been more influential,” John K Galbraith once said).

Second, it seems a bit unfair to place the rise of inequality at the feet of economists. For example, as we will see in Chapter 11 a combination of the digital revolution and globalization have also played an important effect. Societies are naturally attracted to scapegoating, and the role played by economists and economics makes for an easy target.

That said, the past few years have justifiably been a time of reckoning for economics as a field of research and public policy. From a methodological point of view, the preference given to formal — often mathematical — tools has unwittingly led economists to focus more on efficiency (which is more easily measurable) than on social justice (which is considerably more difficult to measure and act upon).

What is, then, the way forward? There is an old German expression — used by Luther, Kepler and Goethe, among others — which, translated and adapted into English, became “don’t throw the baby out with the bathwater.” We should acknowledge that some aspects of economics thinking have done more harm than good to the economy and society. But ridding public policy from its economics foundation would likely lead to an even worst outcome. Economics is part of the problem, but it is also part of the solution. It is impossible to think about the effects of public policy without having a good idea of how economic agents will act and react. And it’s impossible to predict behavior without a good model of what makes economic agents tick. This applies to all sorts of policies: which taxes to levy and at what level, how to set import tariffs and what immigration policy to implement, when to allow firms to merge and when not to, how to reopen the economy in a post-pandemic context, and so on.

Equally important, it’s impossible to move forward without a clear knowledge of what the options are. We can and should think outside of the box, but we cannot live outside of the feasible set (cf p. 72). In one of the many recent critical appraisals of microeconomic
US President Ronald Reagan and British Prime Minister Margaret Thatcher pushed privatization and deregulation during the 1980.

analysis, an article states quite categorically that “our priority should be to build resilient systems explicitly designed to withstand worst-case scenarios,” whereas “mainstream economics has a single overarching objective,... efficiency.” If we think of economic resiliency and economic efficiency as two different goals, then this statement makes little sense: it’s like insisting on one of the extreme points in Figure 3.3. Why on earth should the extreme of maximum resiliency be optimal? If you want to minimize the probability of dying today (worst case scenario), then you should simply sit down at home and avoid any of a series of possible activities that carry some risk of death, small as it might be (e.g., crossing the street). Like it or not, life is made of trade-offs. Granted, economists have insisted too much on economic efficiency as opposed to other important goals (resilience, equity, sustainability), but again we should not throw the baby out with the bathwater.

In sum, the tool of microeconomics plays a crucial role in policy design. But as we use it, we must take into account that behind each policy there are individual people with individual aspirations and goals, with individual problems and limitations. Individuals who are as human as any other human being. Only then can economics become a force for the good.
PART II
SCARCITY AND CHOICE
In Chapter 2 we referred to economics as the study of optimal allocation of scarce resources. Two important components in this definition are (a) allocation, which we may also refer to as choice; and (b) scarce resources. This chapter focuses precisely on the choices made by economic agents so as to optimally allocate scarce resources.

INCOME AND LEISURE OVER TIME

One of our our scarcest resource is time: 24 hours a day for a limited number of years. We can use our time resource to engage in a variety of activities. A particularly important distinction is between time spent working and leisure time. We define leisure as time not working, that is 24 minus time working, measured in hours per day; or, equivalently, $24 \times 365$ minus the yearly total number of hours of work.

Figure 3.1 shows the number of hours of work as well as per-capita income in the US from 1870–2016 (both on a yearly basis). In 1870, Americans worked about 3,100 hours a year, which corresponded to about 75 hours a week. By 2016, the average American only worked for about 1,800 hours a year, that is, a little more that one half of the 1870 average. (The number of hours per week did not drop as much as the total number of hours because the number of work days also declined considerably during the 1870–2016 period.) In the meantime, per-capita income increased by a factor of about 16.
Excluding the 1938 observation, which corresponds to a very special period in American economic history (the Great Depression), we observe that, over time, Americans have a higher income and work fewer hours. To put it differently, today’s average American has a higher income and enjoys more hours of free time than an average American in 1870.

From an economics point of view, the positive relation between income and free time is not obvious. The increase in income is closely associated with an increase in the wage rate. So, over time we observe an increase in the wage rate and a decrease in the number of hours worked. One might have expected that, as the wage rate increases, Americans might be interested in working more hours, not fewer hours. As we will see, from a theoretical point of view it could go either way, as there are two effects of opposite sign: On the one hand, a higher wage makes working longer hours more attractive, but on the other hand, as income and consumption increase, people attach greater value to leisure, and more leisure time implies less work time. Which effect dominates? This chapter lays out the basic framework to address questions like this, and the next chapter addresses a variety of applications including the wage-leisure problem.
INCOME AND LEISURE ACROSS COUNTRIES

Consider now the data shown in Figure 3.2. We fix a particular date (2016) and compare income and free time combinations across countries. One first stylized fact we get from eyeballing the scatter plot is that higher income is associated with more free time. This is consistent with what we observed for the US over time (cf Figure 3.1).

A second observation regarding Figure 3.2 is that there are many “outliers” from the purported positive correlation. Consider, for example, France and the US. The average French person has lower income than the average American, but also more free time than the average American. Why is this so? Which is better in terms of living standards?

ROADMAP

Our goal in this and the next two chapters is to analyze the optimal allocation of scarce resources (time, income, natural resources, etc) by economic agents (consumers, firms, workers, students, etc). Throughout the present chapter, we will consider a specific example: Alexei must choose how to use time (scarce resource) to study or simply enjoy leisure.

In Chapter 2 we read about research on the habits and performance of 84 Florida State University students. It was estimated that
3.1. FEASIBLE SET

one extra hour of study is associated with a grade increase of 0.24. In other words, there is a trade-off between enjoying leisure time and getting a better course grade (surprise!).

In the next sections, we address this trade-off in a formal way. We first derive Alexei’s feasible set, that is, the combinations of leisure and course grade that are attainable. We then model Alexei’s preferences regarding grade and leisure time by means of indifference curves. Finally, given Alexei’s preferences and feasible set, we derive Alexei’s optimal choice.

3.1. FEASIBLE SET

Consider the fictional case of Alexei, a student who must choose how many hours to study. Suppose that we have determined that the relation between hours of study and Alexei’s grade is given by the following table

<table>
<thead>
<tr>
<th>Study hours</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>0</td>
<td>20</td>
<td>33</td>
<td>42</td>
<td>50</td>
<td>57</td>
<td>63</td>
<td>69</td>
</tr>
<tr>
<td>Study hours</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15+</td>
</tr>
<tr>
<td>Grade</td>
<td>74</td>
<td>78</td>
<td>81</td>
<td>84</td>
<td>86</td>
<td>88</td>
<td>89</td>
<td>90</td>
</tr>
</tbody>
</table>

Both leisure time and grade are goods: the more Alexei has of them, the better off he is. Ideally, Alexei would like to have 24 hours of leisure and a grade of 100. Unfortunately, such combination is not feasible. Alexei’s feasible set is the set of combinations of leisure and grade that Alexei can attain.

Alexei’s feasible set is illustrated in Figure 3.3. Alexei owns the resource time, specifically 24 hours per day. Alexi can turn time into grade according to the above table. If Alexei does not study at all, he gets all 24 hours of leisure, but a grade of zero. It follows that (24,0) is a point in Alexei’s feasible set. Naturally, Alexei can enjoy up to 24 hours of leisure, but he can also enjoy fewer than 24 hours, so all points \((x,0)\), where \(x \leq 24\), also belong to Alexei’s feasible set.

Suppose now that Alexei works for 6 hours. This leaves up to \(18 = 24 - 6\) hours of leisure to be enjoyed. Moreover, the 6 hours of study yield a grade of 63. This generates another point of Alexei’s
feasible set. More generally, we can choose a value of hours of leisure, determine the corresponding number of hours of study (24 – leisure), and, based on Alexei’s studying abilities, determine the expected grade. We thus obtain a point on the (leisure, grade) map corresponding to the boundary of Alexei’s feasible set.

MARGINAL RATE OF TRANSFORMATION

The boundary of Alexei’s feasible set illustrates the trade-off that Alexei faces in the choice of grade vs leisure. Specifically, we measure this trade-off by computing the marginal rate of transformation (MRT). The MRT is given by the absolute value of the slope of the outer edge of the feasible set. The MRT indicates how much Alexi needs to give up of one good in order to obtain more of the other. In other words, the MRT measures the opportunity cost of getting more leisure (the good on the horizontal axis) in terms of forgone good grades (the good on the vertical axis).

The marginal rate of transformation (MRT) measures the opportunity cost of getting one additional unit of \( x \) in terms of foregone units of \( y \).

In Table 3.1, we compute the MRT at each point on the edge of Alexei’s feasible set. For example, if Alexei studies for 4 hours then...
TABLE 3.1
Marginal Rate of Transformation

<table>
<thead>
<tr>
<th>Study (hours)</th>
<th>Grade (points)</th>
<th>Leisure (hours)</th>
<th>MRT (point/hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>24</td>
<td>N/A</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>57</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>74</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>84</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>88</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>89</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>90</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>16+</td>
<td>90</td>
<td>9—</td>
<td>0</td>
</tr>
</tbody>
</table>

he enjoys 20 hours of leisure time. Moreover, 4 hours of study produce a grade of 50 (not so good). What is the MRT at this point? If Alexi wants to enjoy an additional hour of leisure — the 21st unit — then he must reduce study time to 3 hours. This implies a drop in grade of \(8 = 50 - 42\). It follows that the MRT is given by \((50 - 42)/(21 - 20) = 8\).

Notice that, as Alexei spends more and more hours of time on leisure, the academic cost of an additional hour of leisure time increases. We can see this in Table 3.1 as we move up the third column (more hours of leisure) and check the corresponding value in the fourth column (MRT). Conversely, as Alexei spends fewer and fewer hours on leisure, and more and more time studying, the benefit (in terms of grade) of an additional hour of study time decreases. We can see this in Table 3.1 as we move down the first column (more hours
TABLE 3.2
Alexi’s indifferent combinations

<table>
<thead>
<tr>
<th>leisure</th>
<th>grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>84</td>
</tr>
<tr>
<td>16</td>
<td>75</td>
</tr>
<tr>
<td>17</td>
<td>67</td>
</tr>
<tr>
<td>18</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>

of study) and check the corresponding value in the fourth column (MRT). We will return to this in Chapter 5 and show that this corresponds to the property of decreasing marginal product: the more you study, the less an additional hour of study helps improving your grade. In terms of Figure 3.3, this corresponds to a feasible set boundary that is concave with respect to the origin.

3.2. PREFERENCES

In the previous section we dealt with Alexei’s feasible set: what Alexei can do in terms of grade and leisure. In this section we deal with Alexei’s preferences: what he likes in terms of grade and leisure. Economic choice is precisely the combination of preferences/goals (what we want) and constraints (what we can do).

Unlike the feasible set, which is given by objective data, preferences are a highly subjective business: there’s no arguing about tastes, so the saying goes. However, economists believe that preferences, subjective as they may be, can be measured. With that in mind, suppose that Alexei is known to be indifferent between the combinations of leisure time (in hours) and final grade (in points) listed on Table 3.2. We say all of these combinations give Alexei the same level of utility, a concept that describes a person’s preferences.

Consider the following alternative combination of free time and grade: 13 hours of leisure and a grade of 84. Would this give Alexei a lower or a higher utility than any of the combinations above? The answer is: lower. The reason is that one of the combinations in Table 3.2 is (15, 84). The proposed alternative, (13, 84), has the same of good
And more of a good is better (so we assume). And since all of the combinations in Table 3.2 provide the same utility level, it follows that (13, 84) is worse than any of the combinations in the table above (by transitivity, which we also assume).

Consider the following alternative combination of free time and grade: 18 hours of leisure and a grade of 70. Would this give Alexei a lower or a higher utility than any of the combinations above? The answer is: higher. The reason is that one of the combinations in the table above is (18, 60). The proposed alternative, (18, 70), has the same of good 1 and more of good 2. And more of a good is better (so we assume). And since all of the combinations in the table above provide the same utility level, it follows that (18, 70) is better than any of the combinations in the table above (by transitivity, which we also assume; more below).

Generally speaking, given two choices \( c_1 \) and \( c_2 \), either Alexei

- prefers \( c_1 \) to \( c_2 \) (which we denote by \( c_1 \succ c_2 \))
- prefers \( c_2 \) to \( c_1 \) (which we denote by \( c_2 \succ c_1 \))
- is indifferent between \( c_1 \) and \( c_2 \) (which we denote by \( c_1 \sim c_2 \))

If \( c_2 \) has more of both good things (e.g., leisure and final grade) than \( c_1 \), then \( c_2 \succ c_1 \). Otherwise, it depends on each person’s trade-off between the two goods.

More generally, we assume that economic agents (including Alexei) have preferences characterized by a relation \( \succ \). Specifically, we say that \( c_1 \succ c_2 \) if the agent prefers \( c_1 \) to \( c_2 \); and we say that \( c_1 \sim c_2 \) if the agent is indifferent between \( c_1 \) and \( c_2 \). Moreover, we assume the preference relation satisfies the following behavioral postulates or behavior axioms:

1. **Completeness**: given \( c_1 \) and \( c_2 \), either \( c_1 \succ c_2 \) or \( c_2 \succ c_1 \) or \( c_1 \sim c_2 \)
2. **Transitivity**: if \( c_1 \succ c_2 \) and \( c_2 \succ c_3 \) then \( c_1 \succ c_3 \)
3. **Monotonicity**: if \( c_2 \) has more of every good than \( c_1 \) then \( c_2 \succ c_1 \)

In economics we refer to agents who choose their best option based on preferences consistent with these postulates as rational agents (*homo economicus*). Let us now look at each of these in detail. Completeness means that “I have no idea” does not apply when it comes
to preferences: each agent can always tell, given any two options, which one she prefers (or whether she is indifferent). Note that this does not imply that the agent is fully cognizant of the pluses and minuses of each option. In most real-world situations, there is a lot of uncertainty regarding outcomes. That said, we assume that, factoring in all of that uncertainty, economic agents are able to — and actually do — compare alternative options and evaluate their relative merits.

Transitivity is a fundamental postulate of rational behavior. If Alexei prefers \( a \) to \( b \), prefers \( b \) to \( c \), and finally prefers \( c \) to \( a \) then there is something wrong with Alexei, or so an economist would say. Finally, Monotonicity is simply the formal counterpart of the idea that we are dealing with goods: more of a good thing is better. What about that time when you ate way too many potato chips and felt sick for two days? The way economists deal with that is by remarking that you can freely dispose of extra amounts of goods. In other words, if you are given 100 tubes of Pringles you don’t need to eat them all, so having more tubes of Pringles cannot make you worse off. You can see how the assumption is controversial, and we will return to this in Section 3.3.

How do we deal with “bads” such as pollution? Unfortunately, the free disposal assumption clearly does not apply here. Let \( x \) be the amount of pollution that an agent suffers from. The trick is then to say that the agent benefits from a “good” (call it “lack of pollution”, for example) in the amount \(-x\). A greater value of \(-x\) then increases the agent’s utility, which is the same as saying the agent benefits from less pollution.

INDIFFERENCE CURVES

Let us return to the values in Table 3.2. We can plot these combinations of free time and grade on a graph. We can then connect the various combinations that Alexei is indifferent about. We refer to this line as one of Alexei’s indifference curves. Since Alexei is indifferent among all of the points in this indifference curve, and since Alexei prefers combinations that induce higher utility levels, it follows that all points along the indifference curve correspond to the same utility level, which we denote by \( U' \).
A brief digression which may help understand the concept of indifference curves. There is an analogy between indifference curves and iso-altitude curves in topographic maps. Iso-altitude curves connect points on the terrain that have the same altitude. If the lines are very close together, we know the terrain is very steep at that point. More generally, iso-altitude lines allow us to represent a three-dimensional object in a two-dimensional graph. Figure 3.5 shows a topographic map (top) as well as a section of a two-peaked mountain. The line labeled “20” on the top map, for example, connects all points with an altitude of 20. Similarly, the middle indifference curve in Figure 3.4 connects all points with utility level $U_0$.

Similarly to a topographic map, which has many iso-altitude lines (one per altitude level), we also have a multitude (in fact, a continuum) of indifference curves. In Figure 3.4 we see the indifference curves corresponding to utility levels $U'$, $U''$ and $U'''$. Given our monotonicity postulate (more of a good thing implies higher utility) we can conclude that $U' > U''$ and $U''' > U'$. In fact, when it comes to indifference curves we know that the farther away they are from the origin, the higher the utility level they correspond to. The same is not true for maps, that is, we don’t know which way altitude increases. For this reason, we need to label each iso-altitude line to know which way the mountain slopes.

Just like the shape of iso-altitude lines show us the shape of a
3.2. PREFERENCES

Digression: topographic maps

Given area, so the shape of the indifference curves gives us an idea of a person preferences. For example, if Alexei did not care about grade at all, then his indifference curves would be vertical. The idea is that, for a given number of hours of leisure, giving Alexei a higher grade would not change his utility.

If you look at a topographical map, you will note that iso-altitude lines do not cross. Why not? Two different iso-altitude lines correspond to two different altitude levels. Were two lines to cross we would have a point with two different altitudes simultaneously, which is clearly impossible. Similarly, no two different indifference curves can cross. This is illustrated in the top panel of Figure 3.6. According to the indifference curves, \( A \sim B \) and \( A \sim C \), hence \( B \sim C \). But \( B \) has more of both goods than \( C \), so \( B \succ C \), which in turn implies a contradiction!

Another property of indifference curves is that they are convex (generally speaking). This is illustrated in the bottom panel of Figure 3.6. Combinations \( A \) and \( B \) lead to the same utility level (they lie on the same indifference curve). Consider now combination \( C \), which consists of one half of \( A \) and one half of \( B \). Since the indifference curve containing \( A \) and \( B \) is convex, combination \( C \) lies to the north east of the indifference curve, that is, \( C \) corresponds to a higher utility level than \( A \) or \( B \). Intuitively, convexity of the indifference curve corresponds to the property that people prefer a moderate amount of
3.2. PREFERENCES

A ~ B and A ~ C but B > C, a contradiction

balanced bundles (e.g., C) are better than “extreme” bundles (e.g., A, B)

FIGURE 3.6
Properties of indifference curves

each good to a very high amount of one and a very low amount of the other.

Going back to the analogy between indifference curves and topographical maps. The latter correspond to altitude, which is a well-defined quantity (e.g., 1,000 feet). The former correspond to an individual’s utility level, which is a not-so-well defined quantity. In fact, as long as the indifference curves remain the same, it really does not matter what the specific values of $U'$, $U''$ and $U'''$ in Figure 3.4 are: Indifference curves help describe an individual’s preferences, not an absolute value of “happiness.”

Let us summarize what we’ve seen so far regarding utility and indifference curves.

• Better combinations (higher utility) correspond to indifference
TABLE 3.3
Alexei’s MRS

<table>
<thead>
<tr>
<th>leisure</th>
<th>grade</th>
<th>MRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>84</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>75</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>67</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>4</td>
</tr>
</tbody>
</table>

curves farther from the origin (monotonicity, also know as non-satiation assumption)

- Indifference curves slope downward (more of one good compensates for less of the other good)
- Indifference curves are typically convex (a mix is better than extremes)
- Indifference curves never cross (if that were the case, we would have a violation of transitivity, thus a violation of rationality)
- Particular shapes of indifference curves reflect an agent’s preferences (there’s no arguing about tastes)
- It doesn’t matter what units we measure utility with, so long as utility levels are consistent with an agent’s indifference-curve mapping

MARGINAL RATE OF SUBSTITUTION

The shape of each person’s indifference curves reflects that person’s preferences. We are particularly interested in the slope of the indifference curves. The **marginal rate of substitution (MRS)** corresponds precisely to the absolute value of the slope of a person’s indifference curve. This slope can be approximated by the the ratio of point-to-point variations between two nearby points of an indifference curve. Since in general indifference curves are not linear, the slope depends of the particular combination we consider. Accordingly, the MRS varies from point to point.

The MRS measures how a person trades off one good for another. Specifically, it quantifies how much an agent is willing to give up of
3.2. PREFERENCES

one good in order to obtain more of another good. Consider Table 3.3, which includes various combinations that Alexei is indifferent about. In order to increase leisure from 15 to 16 (i.e., study one hour less), Alexei is willing to receive a 9 point lower grade. This implies that the MRS (at 15 hours of leisure and a grade of 84) is (approximately) given by \((84 - 75)/(16 - 15) = 9\), that is, loss in grade divided by gain in leisure.

As mentioned earlier, the shape of the indifference curves reflects an individual’s tastes. And the MRS is equal to the slope (in absolute value) of an individual’s indifference curves. Given that, we conclude that an individual with high MRS cares a lot for the good measured on the horizontal axis. For example, a student who cares a lot for free time has a very high MRS of free time for grade.

The definition of MRS has some similarities with respect to MRT. Recall that MRT indicates how much an agent has to give up of \(y\) in order to obtain one additional unit of \(x\). By contrast, MRS indicates how much an agent is willing to give up of \(y\) in order to obtain one additional unit of \(x\). In words, the two definitions seem very similar, but there is a big difference between “willing to give up” and “having to give up”, as many of us have found out in our own lives. In other words, the concept of MRS relates to an individual’s preferences, whereas the concept of MRT relates to an individual’s feasible set. The essence of this chapter is precisely to bring together the two concepts — what an individual wants and what an individual can do — so as to determine an individual’s optimal choice.

As mentioned earlier, indifference curves are normally convex. Figure 3.7 shows Alexei’s indifference curve corresponding to the combinations on Table 3.3. At 15 hours of leisure and a grade of 84, the slope of Alexei’s indifference curve is approximately equal to 9. However, at 18 hours of leisure and a grade of 60 (another point in the same indifference curve), the slope of the indifference curve is approximately equal to 5, the ratio of the change in final grade, 60 - 50, divided by the change in leisure time, 20 - 18. In other words, considering two points on the same indifference curve, the one with a higher value of \(x\) (leisure in this case) has a lower slope (in absolute value). Since MRS is equal to the absolute value of the slope of the indifference curve, the above property is equivalent to the law of decreasing marginal rate of substitution. Specifically, MRS is decreasing with respect to the variable measured on the horizontal axis.
The law of decreasing marginal rate of substitution has a simple economic intuition: the more you have of one given thing, the less you care about having more of it in relative terms, that is, in relation to other goods. For the sake of illustration, suppose you have pizza and soda for lunch. If you have 10 slices of pizza and no soda in front of you would easily give up one slice of pizza for a can of soda. If instead you have one slice of pizza and three cans of soda, then you don’t particularly care for an additional can of soda.

The marginal rate of substitution of good $x$ is decreasing in $x$: the more you get of $x$, the less you are willing to give up of $y$ in order to get more of $x$.

In terms of indifference curves, a decreasing marginal rate of substitution implies that indifference curves are convex, as shown in Figure 3.6. In this regard, there are two extreme cases of indifference curves. If two goods are perfect substitutes, then the corresponding indifference curves are straight lines (no strict convexity). At the opposite extreme, if two goods are perfect complements, indifference curves are L shaped (extreme convexity).

A note on notation: The way we refer to the MRS can sometimes be confusing. Therefore, it helps to set on a series of conventions which we will try to follow throughout the book. Suppose that the
good on the vertical axis is called $y$, whereas the good on the horizontal axis is called $x$.

- When we say MRS of $x$ for $y$, we mean replacing $y$ with $x$. One way to think about it is that “for” means “in place of”.
- We represent MRS of $x$ for $y$ as $\text{MRS}_{xy}$.
- When we simply refer to MRS we mean $\text{MRS}_{xy}$.
- Given the above convention, $\text{MRS}_{xy}$ is the absolute value of the slope of the indifference curve.
- While I will not insist a lot on this, $\text{MRS}_{xy}$ corresponds to the ratio of incremental (or marginal) utility of $x$ divided by incremental (or marginal) utility of $y$.
- A similar convention applies to $\text{MRT}_{xy}$ (cf Section 3.1).

To conclude this section, let us summarize the main points regarding agent preferences:

- If individuals obey certain behavioral postulates, then they are able to rank all available options.
- The ranking can be illustrated with an indifference curve map (i.e., a series of iso-utility curves).
- The negative of the slope of the indifference curve measures the marginal rate of substitution (MRS).
- MRS is the rate at which an individual would forego a certain good $y$ in order to get one more unit of good ($x$).
- MRS decreases as $x$ is substituted for $y$.

3.3. THE MARGINAL RULE

Having derived Alexei’s preferences (what he likes) and Alexei’s feasible set (what he can do), we can now turn to Alexei’s optimal decision making.

Alexei’s optimal choice corresponds to a combination of leisure and grade. Specifically, Alexei prefers the combination which (a) belongs to the feasible set and (b) is associated with an indifference curve located the farthest possible from the origin. Figure 3.8 illustrates this. Combination $c^*$ is Alexei’s optimal choice. There is no
indifference curve associated with utility level greater than $U_2$ which contains a point contained in the feasible set (area in blue).

To put it differently, all points in the feasible set other than $c^*$ correspond to a lower utility level than $c^*$. Take for example combination $c'$. It is associated with an indifference curve $U_1$, which corresponds to lower utility than $U_2$. To see this, note that there are points in the $U_2$ curve with strictly less of final grade and leisure than $c^*$. This implies that $U_1 < U_2$. Since $c'$ belongs to the $U_1$ curve and $c^*$ belongs to the $U_2$ curve, it follows that Alexei prefers $c^*$ to $c'$. The argument would be even stronger if we considered a combination $c'''$ located strictly within the feasible set. (Why?)

Note also that combinations such as $c''$ are strictly better (from Alexei’s point of view) than $c^*$. Does that imply that $c^*$ is not optimal? Not really: Since $c''$ is not feasible (that is, does not belong to the feasible set), $c''$ cannot be optimal. Strictly speaking, the concept of optimality we are defining here is one of constrained optimization. After all, that’s what economics is all about: finding the optimal resource allocation in a world where resources are scarce (thus the idea of constrained optimization).

We now come to a very important rule regarding constrained optimization. In Figure 3.9, we see that at the optimal point $c^*$ the slope of the boundary of the feasible set is equal to the slope of the corresponding indifference curve (that is, the indifference curve that goes
through point $c^*$). Since the slope of the frontier of the feasible set is equal (in absolute value) to the MRT, and the slope of the indifference curve is equal (in absolute value) to the MRS, we conclude that, at the optimum point

$$MRS = MRT$$

I should add that I frequently choose font size to reflect the importance of a result. The above equality is typeset in a very, very large font size.

To understand why the equality $MRS = MRT$ must hold, it may help to understand what happens when it does not hold. Take for example point $c'$ in Figure 3.8. At this point, the feasible set boundary is flatter than the indifference curve going through the same point. This means that $MRT < MRS$. This in turn implies that Alexei is willing to give up more grade for an extra hour of leisure than is required by his feasible set. But then there must be a combination which is both (a) feasible and (b) better than $c'$. Specifically, point $c'''$ in Figure 3.8 corresponds to this move.

Note that the slope of the segment connecting $c'$ and $c'''$ is higher (in absolute value) than the MRT at $c'$. This implies that $c'''$ is feasible. In other words, by moving from $c'$ to $c'''$ we’re giving up more grade than required by the feasible set. Moreover, the slope of the segment connecting $c'$ and $c'''$ is lower (in absolute value) than the MRS at $c'$. This implies that $c'''$ is preferred to $c'$. In other words, by moving from $c'$ to $c'''$ we’re giving up less grade than required to keep the same utility level.

Combination $c^*$ is the only combination where such argument does not hold, precisely because $MRS = MRT$. Figure 3.9 presents the above arguments in a more systematic way. Whenever $MRS \neq MRT$, there is room for improvement. Combination $c^*$ is the only combination where such argument does not hold, precisely because $MRS = MRT$ to begin with.

RATIONAL BEHAVIOR?

Given the emphasis that economics and economists place on the “marginal approach” (MRT, MRS, etc.) to decision making, it’s im-
important to pause and think a bit about the validity of the economics model of human behavior. First, the economics model assumes that each individual is doing what’s best for him or herself. One common objection to this assumption is that the economics model ignores the fact that we care for others. Here it’s important to make the distinction between positive and normative analysis (cf Section 2.1). It is a wonderful thing to aim for a society where we all care for each other (normative statement), but we must take into account that we have a natural tendency to think about ourselves (positive statement). We should therefore design social institutions that foster social behavior taking as given the constraint of individualistic behavior. Moreover, taking into account that we do care for others, we could, for example, include Ana’s grade as part of Alexei’s preferences.

A second criticism is that the models — at least the ones we’ve considered so far — are far too simplistic. Alexei’s problems, for example, go well beyond determining how many hours to study. He needs to find a good tutor and a quiet place to study, he has a bunch of errands to run and no time to do it, etc, etc. The answer here is the idea of models as maps introduced in Section 2.1. Deriving the optimal number of hours of study is not the only decision that Alexei needs to make, but it is certainly one of the more important ones. And simplifying the analysis by focusing on that one decision helps understanding the main trade-offs and how to handle them.

The idea of models as analogical narratives of human behavior is
important when we consider the various concepts introduced in this and the next chapter. With very few exceptions, economic agents have no idea of what an indifference curve is or what is the definition of marginal rate of substitution. So why all this talk about the “marginal approach” and formulas such as MRS = MRT? The answer is that, from an economic analysis point of view, what is relevant is whether economic agents behave as if they were finding the point where MRS = MRT. Economists such as Milton Friedman emphasized this point with examples such a baseball playing. Suppose a baseball batter pops up a ball. If you want to predict where a good left fielder will move to, all you need to do is determine the ball’s trajectory, for good left fielders find their way to the ball and catch it. This you can do by solving a differential equation (taking as inputs the strength and direction of initial impact, wind conditions, and perhaps a few other parameters). Most baseball players are not familiar with differential equations. All they do is look at the ball and gradually adjust their position as a function of the ball’s position and movement. It’s as if the baseball player were constantly re-calculating the solution to a differential equation, Friedman would say. So, for the purpose of predicting the baseball player’s behavior we can assume that he solves a mathematical equation in his head and then acts based on the equation’s solution. This is not literally true, but it does the job of describing and predicting the player’s actions.

Similarly, by repeatedly facing the choice between two consumption goods, for example, a consumer might try small adjustments and determine how good each particular bundle is. By dint of repeated experimentation the consumer will eventually converge to something that is likely close to the point where MRS = MRT. Were that not the case, there would be small local changes that would make the consumer better off, in which case I would expect the consumer to continue experimenting. So, even though no consumer explicitly solves the equation MRS = MRT, in practice they end up with a choice that corresponds to solving MRS = MRT.
KEY CONCEPTS

- feasible set
- marginal rate of transformation (MRT)
- utility
- behavior axioms
- rational agents
- completeness
- transitivity
- monotonicity
- indifference curves
- marginal rate of substitution (MRS)
- decreasing marginal rate of substitution
- perfect substitutes
- perfect complements
- constrained optimization
In the previous chapter we presented a basic framework to study an agent’s optimal behavior. We did so in a very particular context: a student who must decide the optimal trade-off between leisure and course grade. In this chapter we consider more common applications of the optimal-choice framework, in particular decisions in a market context, that is, in a context where agents face prices. Examples include choice of consumption bundles, labor supply, and fertility decisions. We also look at the effect of changes in prices on an agents’ optimal choices.

4.1. CONSUMPTION

Consider a particular, but very important, choice problem: the choice of an optimal consumption bundle given market prices. One example is given by Exercise 2.5 in the companion exercise book. In it we learn that Maria has an income of 56 which she spends entirely on clothing and food. The price of food is \( p_f = 1.75 \), whereas the price of clothing is \( p_c = 1.12 \). Figure 4.1 plots Maria’s feasible set. Suppose that Maria were to spend all of her income on clothing. Then she could afford \( 56 / 1.12 = 50 \) units of clothing. Alternatively, she could spend all of her income on food. Then she could afford \( 56 / 1.75 = 32 \) units of food. Maria can also attain any combination of these extremes, that is, positive amounts of clothing and food such that total
expenditure falls below or equals her available income. Specifically, if Maria is to spend all of her income on clothing and food then the following equality holds:

\[ p_x x + p_y y = I \] (4.1)

In other words, the feasible set is defined by a straight line in the \((x, y)\) space, where the coefficients on \(x\) and \(y\) are given by the prices of each good.

The term feasible set, first introduced in Paragraph 3.1, is a fairly broad term to describe the constraint that an economic decision maker is subject to. In the particular but important case of a consumer with a given income level the feasible set is referred to as the consumer’s **budget set**. Moreover, the boundary of the consumer’s budget set is referred to as the consumer’s **budget line**.

Referring back to Figure 4.1, we see that Maria’s budget line is defined by points \(A\) and \(B\). The vertical coordinate of point \(A\) is given by \(I/p_f\), whereas the horizontal coordinate of point \(B\) is given by \(I/p_c\). It follows that the slope of the budget line, in absolute value, is given by \((I/p_f)/(I/p_c) = p_c/p_f\). (Note that the price in the numerator corresponds to the good on the horizontal axis.) We can also find this by solving (4.1) in order to \(y\):

\[ y = I/p_y - (p_x/p_y) x \] (4.2)
which shows that the line’s slope (in absolute value) is given by $p_x / p_y$. It follows that, for Maria, the MRS = MRT formula, introduced in Section 3.3, implies $MRS = p_c / p_f$. In words, Maria should choose consumption levels of food and clothing such that the marginal rate of substitution of clothing for food (how much food she’s willing to give up for an extra unit of clothing) is exactly equal to the ratio (price of clothing divided by price of food). (Why is this the optimal rule?) In terms of Figure 4.1, the point on the budget line where this optimality condition is satisfied corresponds to point $C$.

The optimal consumption mix corresponds to the equality of the MRS and the price ratio.

Economists engage frequently in the exercise of predicting how a change in $x$ will affect $y$, where $x$ could be an event and $y$ an economic variable of interest. For example, $x$ might be “an oil rig in Saudi Arabia is destroyed” and $y$ the price of natural gas. This prediction exercise is generally referred to as **comparative statics**. This is not a very helpful expression, but we will stick to it (and discuss at length in Section 7.1). In what follows we will examine the impact of changes in two particularly important variables: income level and the price of one of the goods.

**CHANGES IN INCOME**

Consider the generic case when a consumer has income $I$ which he splits between goods $x$ and $y$ (as in, for example, Figure 4.2). From Equation (4.2), we see that an increase in $I$ implies a parallel shift in the budget line (that is, the intercept changes but the slope remains the same). What effect does an increase in $I$ have on the optimal consumption bundle? As frequently is the case, the answer is — it depends. Specifically, it depends on the consumer’s preferences. The two panels in Figure 4.2 illustrate two possibilities regarding the impact that an income increase has on the consumption level of $x$. In the top panel, the increase in income leads to an increase in the consumption of $x$. In the bottom panel, the increase in income leads to a decrease in the consumption of $x$.

Economists classify goods according to the relation between income and consumption. Specifically, we say a good is a **normal good**
if its consumption increases when income increases. As the name suggests, most goods are normal goods. However, the opposite is also possible: we say a good is an **inferior good** if its consumption decreases when income increases. Although inferior goods aren’t all that common, it’s fun to think of examples. Spam comes to mind, on the assumption that anyone with enough money would eat something else. This example also suggests that what’s a normal good for one person may be an inferior good for another one; and that what’s a normal good for a given person at time 1 may become an inferior good at time 2. Can you think of examples? In terms of Figure 4.2, we would say that the top panel illustrates a case when $x$ is a normal good, whereas the bottom panel illustrates a case when $x$ is an
4.1. CONSUMPTION

Consider a consumer with income $I$. Initially prices are given by $p_x$ for good $x$ (quantity $x$ measured on the horizontal axis) and $p_y$ for good $y$ (quantity $y$ measured on the vertical axis). The consumer’s budget constraint is given by $b_A$, a line with an intercept of $I/p_y$ on the $y$ axis and $I/p_x$ on the $x$ axis. Intuitively, if the consumer spends all of her income on $y$ then all she can afford is $I/p_y$ units of $y$, ditto for $x$. Given the consumer’s preferences and her budget constraint, the optimal choice is given by point $A$, corresponding to $x_A$ of $x$ and $y_A$ of $y$.

Now suppose that the price of $x$ changes from $p_x$ to $p'_x$ (a price increase), so that the budget constraint pivots from $b_A$ to $b_C$. (We know $p'_x > p_x$ because the intercept $I/p'_x$ is lower than the intercept $I/p_x$.) This implies (for this particular consumer) a shift in optimal choice from A to C. In particular, the consumption of $x$ drops from $x_A$ to $x_C$. This is not surprising: when the price of $x$ increases, normally its consumption decreases.

The shift from A to C may be decomposed into two parts, which we refer to as the income effect and the substitution effect of the price of $x$ on the consumption of $x$. Before getting into formal definitions,
4.1. CONSUMPTION

a bit of economic intuition. If the price of $x$ increases, then the consumer’s real income decreases. In other words, a value $I$ of income does not go as far as it did when $p_x$ was lower. Moreover, if the price of $x$ increases then the relative price of $x$ (relative to $y$) increases. In other words, getting one additional unit of $x$ requires sacrificing more units of $y$. We would like to disentangle the effect of a change in $p_x$ (on the consumption of $x$) into these two effects: how much of the change in $x$ is due to the fact that $x$ is relatively more expensive than $y$, and how much of the change in $x$ is due to the fact that the consumer’s real income is now lower.

The substitution effect corresponds to the change in $x$ caused by the change in relative prices keeping real income constant. The income effect corresponds to the change in $x$ caused by the change in real income.

In what situation could we say that the consumer’s real income remains constant in spite of an increase in $p_x$? A natural definition of constant real income is that, given the new, higher price of $x$, the consumer also has higher nominal income so that her utility level remains constant. In terms of Figure 4.3, budget line $b_B$ corresponds to the new price ratio but a higher income, so much so that the consumer gets the same utility as before the increase in $p_x$. Notice that this is a hypothetical budget line. We are simply asking what income would the consumer need to get in order to keep the same utility level as initially.

Faced with the new, hypothetical budget line (the dashed line in Figure 4.3), the consumer optimally chooses $B$, which gives her the same utility level as point $A$. We thus say that the shift from $A$ to $B$ corresponds to the substitution effect. In other words, the substitution effect corresponds to the change in $x$ caused by the change in relative prices only, that is, keeping real income constant, i.e., staying on the same indifference curve.

Once we derived the substitution effect, the income effect is simply what’s left to account for the total effect of a change in $p_x$. Specifically, an increase in $p_x$ implies a shift from $A$ to $C$. In terms of consumption of $x$, we are talking about a decrease in $x$ from $x_A$ to $x_C$. Note that $x_A$ and $x_C$ are data, that is, are observable. Given our hypothetical dashed budget line, we determined the shift from $x_A$ to $x_B$.
as the substitution effect of the increase in $p_x$. Finally, we identify the shift from B to C, that is the drop in $x$ from $x_B$ to $x_C$, as the income effect of the increase in $p_x$.

To better understand the nature of the income effect, notice that the (vertical) intercept of $b_B$ is given by $(I + DL)/p_y$. This means the difference between $b_C$ and $b_B$ is that the latter corresponds to an extra $DL$ income. This means that, given the new price level (higher $p_x$) we would have to give the consumer an extra $DL$ in order for her utility to remain the same. In reality, she did not get that extra income $DL$, which implies she is at C, not at the hypothetical B. The shift from the hypothetical $b_B$ to the actual $b_C$ corresponds to the change in real income, and the shift from the hypothetical $x_B$ to the actual $x_C$ corresponds to the income effect.

Formally, the **income effect** on $x$ of an increase in $p_x$ corresponds to the effect of a reduction in income, from the level required to keep the initial utility level $(I + DL)$ to the actual (nominal) income level $(I$, which never changed). In other words, the income effect corresponds to the effect of the change in real income (which decreases when $p_x$ increases) which results from the change in price. To rephrase the above point: An increase in $p_x$ to $p'_x$ makes the consumer poorer (prices are higher and income remains at $I$). However, if we give the consumer an extra income of $DL$, then her welfare under the new prices remains the same as with the initial prices. It follows that $DL$ measures the (negative) shock to real income resulting form an increase in $p_x$.

In terms of Figure 4.3, the income effect is given by $x_C - x_B$. The substitution effect, in turn, is given by $x_B - x_A$. Notice that, in the present case, both the income and the substitution effects are negative (that is, $x_C < x_B$ and $x_B < x_A$).

The income effect may have the same sign or the opposite sign of the price change. If $x$ is a normal good (as is the case in the above example), then the income effect on the consumption of $x$ has the opposite sign than the change in price of $x$. In the previous example, the price of $x$ increased and the income effect is negative ($x_C < x_B$). However, if $x$ is an inferior good then the income effect on the consumption of $x$ has the same sign than the change in the price of $x$.

The substitution effect on the consumption of $x$ always has the opposite sign than the change in price of $x$. In the previous example, the price of $x$ increased and the substitution effect is negative ($x_B <
This property is known as the law of demand. In other words, if the price of $x$ increases then the substitution effect on $x$ is negative; if price of $x$ decreases, then the substitution effect on $x$ is positive.

Frequently, we refer to the law of demand as implying that when the price of $x$ increases then the demand for $x$ decreases (and vice versa). This is not quite right, that is, this is not true in general. It is possible that the demand for good $x$ increase when its price goes up (and vice versa). This happens when the income and substitution effects have opposite signs and the income effect is bigger (in absolute value). (For aficionados: when this happens, we say $x$ is a Giffen good.)

The law of demand states that the substitution effect is negative. In most cases, this means that an increase in price implies a decrease in quantity demanded.

WHO CARES?

At this point (and often throughout the course) you may ask yourself: Who cares? Policymakers do, for a variety of reasons. First, in a world with fluctuating prices and income levels, economists are interested in estimating how such changes affect consumer welfare. One way to do so is to compute consumer price indexes and use these to calculate the value of real income, that is, income level corrected for price changes. The income effect / substitution effect decomposition suggests a specific path for this correction.

Another practical application of the income effect concept is given by environmental policy. For decades, economists have advocated the creation of a carbon tax to induce lower consumption of CO$_2$-intensive goods. William Nordhaus, co-winner of the 2018 Nobel Prize, already recognized the threat of CO$_2$ emissions in the early 1970s. In 1992, he proposed a modest carbon tax as a means to rein in increasing emissions. In Section 9.2 we will return to this important issue. For now, we recognize that implementing a carbon tax is a difficult political problem: nobody likes taxes, least of all politicians who fear they might not get re-elected.

One possible political strategy to implement a carbon tax would be to accompany it by a compensating subsidy. Let us return to
4.2. LABOR SUPPLY

Figure 4.3. Suppose that \( x \) corresponds to carbon-intensive goods whereas \( y \) corresponds to “green goods”. A carbon tax implies an increase in \( p_x \), that is, consumers need to pay more for gasoline, airfares, etc. This in turn implies that, for a given dollar income, consumers will be poorer: their dollar will not reach as far as it did before. How much do we need to compensate the consumer for the drop in real income stemming from the carbon tax? The answer is, exactly \( \Delta I \). If we give consumers \( \Delta I \) when the carbon tax is implemented, then their consumption bundle switches from A to B, that is, moving along the initial indifference curve, consumers substitute green goods for carbon-intensive goods while their welfare standard is maintained.

4.2. LABOR SUPPLY

In Chapter 3, we considered Alexei’s problem of balancing leisure time and course grade. A more common problem is to balance leisure time and earned income. Specifically, for the sake of concreteness consider Ana’s choice of how many hours to work. Anna likes income (from work) as well as leisure. An employer offers to pay \( w \) per hour that Anna works. What should Ana do?

Before continuing, a brief note on the motivation for studying Ana’s problem. In some cases, Ana’s decision problem is very much a problem that specific individuals need to make. For example, if I live in New Jersey and drive an Uber/Lyft car on weekends I need to decide how much income I want versus leisure time.

Generally speaking, we do not see employees deciding how many hours to work. Most people hold 9-to-5 type jobs, that is, jobs with well-determined hours. How realistic is then the idea of Ana choosing how many hours to work? The question becomes relevant if Ana must choose between various possible jobs, each of which has different hours. For example, suppose Ana has just graduated from law school and received several job offers. One is a government job in DC. The salary is low but the hours are very good: eight hours a day with very little overtime. Another possibility is to work for a top New York law firm. The pay is considerably higher but so are the hours. In between we find two or three more offers which correspond to intermediate levels of income and leisure. So, effectively Ana’s
decision problem is similar to the Uber/Lyft driver: even though within each job Ana does not have much leeway in terms of how many hours to work, at the moment of choosing which job to take she is effectively making a choice analogous to that of the weekend Uber/Lyft driver.

Just as Alexei had preferences for leisure and grade, so Ana has preferences for leisure and income. Just as Alexei was constrained by a feasible set, so Ana too is constrained by a feasible set. However, whereas Alexei’s feasible set is determined by his study abilities, Ana’s feasible set is determined by the wage rate she gets paid. This is illustrated in Figure 4.4, where Ana’s feasible set is given by the triangle shaded blue. To see why this is indeed Ana’s feasible set, consider one first possibility, namely Ana does not work at all. This gives her 24 hours of leisure but an income of zero. At the opposite extreme, Ana could work for 24 hours a day (though I would not recommend it). This leads to a combination of zero hours of leisure and an income of $24w$, where $w$ is the hourly wage rate. In between, Ana can attain any combination of these extremes, that is, all points in the segment with extremes $(20, 0)$ and $(0, 24w)$.

Why is the boundary of Alexei’s feasible set curved whereas Ana’s is a straight line? Because in Alexei’s case the feasible set is determined by Alexei’s studying abilities (i.e., the study → grade mapping), whereas Ana’s feasible set is determined by a market,
Microeconomic theory models consumer choice as a comparison between marginal rates of substitution and price ratios.

namely the labor market; and in this market Ana is a **price taker**, that is, she is paid the going hourly wage rate \( w \) regardless of how many hours she works. (This is a reasonable assumption for most workers. In other parts of the book we will consider several cases when economic agents, in particular firms, are *not* price takers.) In other words, the non-linear relation we find in Alexei’s pursuit of grade is absent from Ana’s pursuit of income: each additional hour of work produces an extra \( w \) of income. (That said, both the Uber/Lyft driving interpretation and the multiple job offers interpretation of Ana’s problem may involve some curvature in the feasible set: It’s not the same thing driving an Uber at 3pm or 3am; and the different jobs offered to Anna likely correspond to different hours and different wage rates.)

Notice that the slope of the boundary of Ana’s feasible set is given by \(-w\), or simply \( w \) in absolute value. To see this, note that the absolute value of the slope is given by the vertical axis intercept, \( 24w \), divided by the horizontal intercept, 24. We thus have \((24w)/24\), or simply \( w \). The economic intuition is as follows: MRT (the slope of the boundary of the feasible set) measures the opportunity cost of increasing \( x \) (the good on the horizontal axis) in terms of units of \( y \) (the good on the vertical axis). The opportunity cost of getting one extra hour of leisure (\( x \) in the present case) is the foregone income (\( y \) in the present case). And the foregone income of enjoying an extra hour of leisure is the income foregone from working one hour less, that is, \( w \).

We can also appreciate the relation between Ana’s income-leisure problem and Maria’s food-clothing problem. In both cases we have a budget line with slope given by the price ratio. This may seem a
littlly tricky in Ana’s case: What is the price of income? It’s simple 1: it costs $1 to have an extra $1 (we are not allowing for investment in the present analysis). And what is the price of leisure? As we saw above, it’s best to think about the opportunity cost of leisure, which is \( w \). So the absolute value of the slope of the budget line is given by \( w/1 = w \). It follows that the MRS = MRT formula implies MRS = \( w \).

In words, Ana should choose a level of leisure such that the marginal rate of substitution (how much income she’s willing to give up for an extra hour of leisure) is exactly equal to the wage rate. Why is this the optimal rule?

**CHANGES IN INCOME AND PRICES**

Similarly to the consumption problem (Section 4.1), we can now examine the effects of a change in income or a change in price. Specifically, we are interested in two comparative statics problems: (a) the impact of changes in (non-labor) income; and (b) the impact of changes in the wage rate. Answering these questions will take us some way in the direction of understanding the evidence with which we started Chapter 3, in particular the evolution of income and leisure over time (cf Figure 3.1).

One way in which the income-leisure problem is trickier than the consumption problem is that in the former we have two possible sources of income: labor income and non-labor income. Labor income is given by \( w \) times the number of hours worked, that is, 24 minus the number hours of leisure. In what follows, when we talk about the effect of a change in income we refer to a change in non-labor income.

Figure 4.5 illustrates the effect of an increase in (non-labor) income by \( \Delta I \). As can be seen, this corresponds to a parallel shift of the feasible set frontier. To understand this, consider first the case when Ana chooses 24 hours of leisure. Before, her income was zero (her income was exclusively labor income). Now, even not working any hours, Ana has an income of \( \Delta I \). Similarly, suppose that Ana works 24 hours a day, thus leaving no time for leisure. Before, her labor income was 24 \( w \) (as we saw on p. 115). Now, Ana enjoys a total income of 24 \( w + \Delta I \).

Continuing with Figure 4.5, we see that the effect of an increase in non-labor income by \( \Delta I \) is to change Ana’s optimal point from \( c_1 \).
Consider now the effects of a higher wage, specifically an increase from $w_1$ to $w_2$. This is illustrated in two panels of Figure 4.6. If Ana chooses 24 hours of leisure, and thus zero hours of work, then an increase in wage has no effect on her labor income. By contrast, if she works for 24 hours then her labor income increases from $24w_1$ to $24w_2$. Putting these two points together, we conclude that the wage increase implies a rotation of the feasible set boundary around the point (24,0), similarly to the effect of a change in $p_x$ in the consumption case considered in Section 4.1. What is the effect of a wage increase on Ana’s optimal choice? In the top panel of Figure 4.6, we observe a change from $c_1$ to $c_2$ which corresponds to a higher income level and a lower level of leisure (that is, more hours of work). In the bottom panel of Figure 4.6, we observe a change from $c_1$ to $c_2$ which corresponds to a higher income level and a higher level of leisure (that is, fewer hours of work).

In other words, following an increase in wage the level of leisure may increase or decrease; it depends on Ana’s preferences for in-
4.2. LABOR SUPPLY

come and leisure. In terms of our income-substitution effect decomposition, the case on the top panel of Figure 4.6 corresponds to the case when the income effect is so large that it outweighs the substitution effect: Although the price of leisure increases ($w$ is the price of leisure), the amount of leisure chosen by Ana increases!

To conclude, recall that daily leisure time equals 24 hours minus work hours. As such, what was said about the demand for leisure can be reinterpreted as labor supply. In particular, the above analysis shows that an increase in the wage rate may result in an increase or a decrease in the supply of labor. The latter is the case when the income effect of a wage increase is sufficiently large to compensate for the substitution effect of a wage increase.
The supply of labor may react positively or negatively to a change in the wage rate depending on the relative magnitudes of the substitution and income effects

LEISURE CHOICES IN THE US AND SWEDEN

Having gone through the framework of optimal economic decision making, let us return to the data in Figures 3.1 and 3.2. Since leisure equals total time minus time of work, we can re-plot the points in Figure 3.1 with leisure on the horizontal axis. Specifically, annual hours of leisure are equal to 365 times 24 minus the value on the vertical axis of Figure 3.1. Given per-capita GDP (plotted on the horizontal axis of Figure 3.1), we estimate average wage as per-capita GDP divided by hours of work. This is not really correct, for income (measured by GDP) is equal to labor income plus capital income. However, to the extent that the share of labor income has remained relatively constant over the period in question, the evolution of the ratio per-capita GDP divided by hours of work provides a reasonable approximation of the evolution of the wage rate over time.

We next make an additional approximation which is a bit rough but nevertheless may help understand the evolution of income and leisure in the US: We assume that the average values plotted in Figures 3.1 and 3.2 correspond to a representative American worker; that is, to a “typical” American. This is OK so long as we remain aware that underneath an average there is significant dispersion.

In 2016 average income in the US was $33,259 (in 1990 PPP dollars), whereas the average number of hours worked was 6,979 hours. This implies that the point (6.979,33.259) was attainable by the 2016 average American consumer, that is, the point was on the consumer’s budget set (in fact, on the consumer’s budget line). Since the actual number of leisure hours was 6979, we conclude the number of hours worked was 8760 – 6979 = 1781 hours. This in turn implies that the average wage rate was 33259/1781 = $18.67 per hour (in 1990 dollars).

We also know that the consumer could have opted for not working at all. This would have yielded 24 × 365 = 8,760 hours of leisure and zero income, the point on the lower right end of the graph. Assuming that the average consumer could have worked as many
hours as she wanted at the going wage rate, we conclude that the consumer’s budget line includes points (6.979, 33.259) and (8.760, 0). By a similar process, we also get the budget lines in 1938 and in 1960. As can be seen, the average American consumer’s budget set has expanded over the years.

Specifically, from 1938 to 1960 we observe an increase in wage, corresponding to a steeper feasible set frontier. Moreover, we observe that the average American chose to enjoy fewer hours of leisure, that is, to work for longer hours. In terms of the preceding analysis, we say that, from 1938 to 1960, the wage increase was associated with a substitution effect greater than the income effect: as the wage rate increased, Americans increased the number of hours they work.

By contrast, from 1960 to 2016, we observe an additional increase in the wage rate, but this time the average number of hours of leisure increases, that is, the average number of hours of work decreases. In terms of the preceding analysis, we say that, from 1960 to 2016, the wage increase was associated with a substitution effect smaller than the income effect: as the wage rate increased, Americans decreased the number of hours they work.

This is a good time to recall that an economic model (just like any other model) is a simplification of reality. There are many, many aspects of industrial and labor relations, politics, law, technology, and so on, which contributed to the evolution depicted in Figure 4.7. Can
you think of other factors may have explained the evolution from 1938 to 2016?

If Figure 4.7 is based on the data underlying Figure 3.1, Figure 4.8 is based on the data underlying Figure 3.2. For the year 2016, we derive the values of income and average wage for the US and Sweden. By our approximation, the average wage rates in the US and Sweden in 2016 were about the same. As a result, the feasible set for an average American was not very different from the feasible set for an average Swede (assuming that labor is the sole income source). However, the representative American selected considerably fewer hours of leisure than the representative Swede.

How can we explain such stark differences in choice between Americans and Swedes, knowing that, in terms of feasible set, they face a similar problem? The preceding analysis of optimal economic decision making (MRS = MRT) suggests that the preferences of Americans are different than the preferences of Swedes. Specifically, Figure 4.8 depicts indifference curves for a representative American and for a representative Swede which are consistent with the observed choice being optimal.

Notice that, for the values actually chosen by the representative agents, the MRS for the American is approximately the same as the MRS of the Swede. This is because, at the optimum, each equates MRS to MRT, and MRT is given by the wage rate, which in turn is approximately the same in the US and in Sweden.
However, the fact that MRS is the same in the US and in Sweden does not mean that the preferences of Americans are the same as the preferences of Swedes. To see this, consider point $c$, where the particular indifference curves in Figure 4.8 cross. At this point, the indifference curve of the American is flatter than the indifference curve of the Swede. At point $c$, the comparison of MRS is meaningful, for we are considering the same values of leisure and income.

In the present context, a flatter indifference curve means that you care relatively more for the good on the vertical axis. In words, Figure 4.8 means that Americans care relatively more for income (as opposed to leisure) when compared to Swedes. This is consistent with the observation that, for approximately the same wage rate, Americans work longer hours than the Swedes.

As in the case of Figure 4.7, we must acknowledge that there are many other factors influencing the observed choices in the US and Sweden other than differences in preferences. Can you think of any?

### 4.3. OTHER HOUSEHOLD DECISIONS

In this and the previous chapter we have considered three examples of decision problems: Alexei’s choice between leisure and grade; Maria’s choice between food and clothing; and Ana’s choice between leisure and income. Even though the feasible set changes from case to case, all of these examples have one thing in common: the optimal solution is given by the equality of marginal rate of substitution and marginal rate of transformation, where the latter may be given by a production function, a price ratio, or a wage rate.

The fact that economists make so much use of marginal rates, and more generally derive optimal solutions by looking at small potential variations in choice variables, is reflected in the general expression **marginal approach** (see also Section 2.3). Mathematically speaking, this corresponds to the use of calculus, but as the above analysis shows the same intuitive results can be obtained by a combination of verbal, graphical and algebraic statements.

There are many other examples of economic decision problems which follow a similar pattern. In this section, we look at three specific examples: savings; risk and insurance; and fertility choices.
SAVINGS

Why do households save? There are multiple reasons, even for a given household. One motivation is to build up a reserve against unforeseen contingencies (what economists refer to as the precautionary motive). A related one is to enjoy a sense of independence and the power to do things. For others, it’s a matter of accumulating resources with the view of starting a business, or simply putting the downpayment required for a house or a car. The microeconomics framework we’ve presented in this and the previous chapter suggests a different perspective, one that economists refer to as life-cycle optimization (or the intertemporal substitution motive). Specifically,

A household’s decision to save can be thought of as a choice between consumption today and consumption in the future.

In this context, one question of interest is how the trade-off between consumption today and future consumption depends on the interest rate, specifically the rate at which households can turn present savings into future income. Do savings increase when the interest rate increases? Figure 4.9 plots the evolution of the US aggregate personal savings rate (savings as a percentage of income), as well as the US bank prime loan rate. Can you find a relation between interest rating and savings? From 2000-2010 they seem inversely correlated,
Retirement planning can be thought of as a trade-off between consumption in the current period and consumption in a future period. Whereas from 2010-2020 somewhat positively correlated. Does microeconomic analysis have anything to say about it?

For simplicity, suppose that there are only two periods in a person’s lifetime: the period while she is working and the period while she’s retired. Suppose the income during the first period is given by $I$ and that the person in question expects to receive no income during her retirement period.

The top panel of Figure 4.10 illustrates this situation. On the horizontal axis we measure consumption today, whereas on the vertical axis we measure consumption tomorrow. If the consumer decides to spend all of her current income in consumption, then consumption today is equal to $I$, whereas consumption tomorrow is equal to zero. If, at the opposite extreme, the consumer decides not to spend anything on consumption today then her consumption during the retirement years is given by $I(1 + r_a)$, where $r_a$ is the return on invested savings. In practice things are a little more complicated. First, there isn’t one single rate of return to consider: I can invest in bonds, in the stock market, or in other financial assets. Second, there is usually considerable uncertainty regarding the actual value of $r_a$. But for simplicity let us assume our consumer has a good idea of how much a dollar saved today implies in terms of additional revenue during her retirement years. Finally, any intermediate solution between the extremes of consume everything and consume nothing give us the consumer’s budget line $b_a$ (very much like the ones we saw for the consumption and labor supply decisions). Given the consumer’s relative preferences for consumption today and future consumption, the optimal choice is given by point $A$, which corresponds to consuming $c_a$. 


4.3. OTHER HOUSEHOLD DECISIONS

If \( b I (1 + r_b) \) then consumption is smaller than substitution effect consumption today \( (c) \) consumption during retirement \( (f) \)

If \( a I (1 + r_a) \) then income effect is greater than substitution effect consumption today \( (c) \) consumption during retirement \( (f) \)

FIGURE 4.10
Effect of a decrease in the interest rate on consumption and savings

today and \( f_a \) during retirement.

Now suppose that the rate at which the consumer can invest her savings drops from \( r_a \) to \( r_b \). This implies a shift in the budget line from \( b_a \) to \( b_b \). Notice the budget line rotates around the horizontal axis intercept: if the consumer spends all of her current income, then it does not matter what the rate of return is. Given the consumer’s relative preferences for consumption today and future consumption, the new optimal point is given by point \( B \), which corresponds to consuming \( c_b \) today and \( f_b \) during retirement. Notice that, on the upper panel of Figure 4.10, \( c_b > c_a \). In other words, the drop in the rate of return on invested savings leads to an increase in today’s consumption. This makes sense: the opportunity cost of today’s consumption
is foregone consumption tomorrow. If the rate of return is lower, then the amount of consumption tomorrow I need to forego in order to consume one more unit today is lower, thus current consumption becomes more attractive.

By definition, savings are defined by investment minus consumption. Our consumer has an income of $I$ and consumes $c_a$ in the current period when the interest rate is $r_a$. It follows that the consumer saves $s_a = I - c_a$. When $r$ drops to $r_b$, savings drop to $s_b = I - c_b$. Since $c_b > c_a$, we observe a drop in savings given by $c_b - c_a$, a variation that corresponds to the arrow in the top panel in Figure 4.10.

However, as we’ve seen multiple times, changes in prices — and the interest rate is a price — lead to substitution as well as to income effects. The bottom panel in Figure 4.10 illustrates the case when the income effect is so significant that the overall effect of a drop in the rate of return on investment savings is to increase the savings rate, as illustrated by the arrow in the bottom panel of Figure 4.10.

What’s going on here? The bottom panel of Figure 4.10 obviously corresponds to a different consumer than the top panel. You can tell this by the different shape of the indifference curves. In the bottom panel the indifference curve going through point B is “flatter” than the indifference curves in the top panel. In the present context, flatter indifference curves correspond to the consumer placing relatively greater value on consumption during retirement. (In the limit, if the indifference curves were completely flat, then consumption during retirement would be all that the consumer cared about.) You can see how a flatter indifference curve implies an optimum farther to the left, that is, a solution with a lower level of consumption today. So, what’s happening is that, all things equal — in particular, for a given level of consumption today — the drop in the interest rate implies that the consumer can expect a drop in consumption during retirement. If consumption during retirement is very important for the consumer — as is the case for the consumer in the bottom panel — then the consumer optimally compensates for this drop in future consumption by saving more today. We thus end up with the somewhat surprising conclusion that, as the rate of return on investment savings decreases, the consumer reduces today’s consumption level and increases the level of savings.

So, does a decrease in the interest rate lead to a decrease or an increase in the savings rate? As often is the case in economics, the
answer is — it depends. In the present case, it depends — among other things — on each individual’s preferences for consumption today and future consumption. The situation on the bottom panel of Figure 4.10 is more likely to be the exception than the rule. That said, it’s definitely a possibility. Empirical studies suggest that the substitution effect is negative as theory would predict. However, based both on studies for develop and developing countries, the total effect tends to be rather small. All in all, this is consistent with the idea that the substitution effect and the income effect are present, have opposite signs, and are approximately of the same order of magnitude, possibly with the income effect being a bit smaller (in absolute value) than the substitution effect. So, perhaps it’s not that surprising that we do not find a clear relation between the two time series in Figure 4.9. Moreover, it’s clear that there are many other factors affecting the savings rate other than the interest rate. On this, the reader is referred to Exercise 4.16 in the companion exercise book.

An increase in the interest rate typically leads to an increase in savings. Empirically, this effect is small, possibly because of the income effect of the interest rate increase.

So far, we’ve focused on the effect of a change in the price (the interest rate). As in previous sections of this chapter, we are also interested in understanding the effect of an increase in income. Common sense suggests, and empirical evidence confirms, that consumption today and consumption tomorrow are both normal goods. Therefore, an increase in today’s income is reflected in both an increase in today’s consumption and an increase in tomorrow’s consumption. Since tomorrow’s consumption is directly related to today’s savings, we conclude that an increase in income leads to an increase in savings. To conclude, an extra dollar in today’s income implies both an increase in today’s consumption and an increase in today’s savings. Macroeconomists are particularly interested in this savings function, which essentially corresponds to how many cents of an extra dollar in income are turned into savings.
The way economists model uncertainty and risk is to assume the future is made up of multiple possible states, usually referred to as states of the world. For simplicity, let us consider two possible states of the world. In the $h$ state, the consumer is healthy, whereas in the $s$ state she is sick. Suppose moreover that if the consumer is sick then she must pay a large sum to recover her health, so much so that, absent any health insurance, her net income (net of health expenditures) drops from $h_1$ to $s_1$.

Figure 4.11 illustrates this situation. On the horizontal axis we plot net income in the $h$ state of the world, whereas on the vertical axis you plot net income in the $s$ state of the world. This may seem a bit confusing, for in the past we considered choices between actual goods (food and clothing, income and leisure time, consumption today and consumption tomorrow). In the present case we consider trade-offs between income in state $h$ and income in state $s$. The source of possible confusion is that, in reality, only one of these states will turn out to be true (hopefully state $h$). But a rational agent must be prepared for all eventualities and make choices that cover all possible states.

The line marked as $45^\circ$ corresponds to the main diagonal. This is the line such the value of $h$ is equal to the value of $s$. For example, $h_2 = s_2$. In words, points along the $45^\circ$ correspond to situations
where there is not income risk: my future income is the same regardless of the state of the world. Point $E_1$, the point that describes the consumer’s initial situation, is clearly below the 45° line, so it clearly involves risk: if the individual happens to be sick then her net income is considerably lower than if she is healthy.

As before, we assume that the consumer has preferences over the possible states, and that these preferences are described by indifference curves. In the present context, the shape of the indifference curves reflects the consumer’s degree of risk aversion. In the limit, a risk neutral consumer’s indifference curves are straight lines, where the slope is such that all points correspond to the same expected value. For example, if the consumer is healthy with probability $\rho$ and sick with probability $1 - \rho$, then expected income is given by $\rho h + (1 - \rho) s$. To be more specific, suppose that $\rho = 80\%$, so you are sick with probability 20%. Suppose moreover that $h = 200$ and $s = 120$. Then expected income would be $0.8 \times 200 + 0.2 \times 120 = 184$.

Suppose for simplicity that the probability of being sick is equal to the probability of being healthy. (Not a very cheerful example, I know, but one that hopefully simplifies our lives and helps understand the issues.) Then the risk-neutral consumer’s indifference curves are straight lines with slope -1. The green line in Figure 4.11 provides an example. If I go to a risk neutral consumer and propose to take away $1 in the $h$ state to compensate for an extra $1 in the $s$ state she will shrug her shoulders, as she will be indifferent between these two possible deals. In particular, she would be indifferent between point $E_1$ and point $E_2$. Note that point $E_2$ corresponds to a combination of income levels such that there is no risk (it falls on the 45° line) and expected income is the same as in $E_1$ (it falls on the same indifference curve of a risk-neutral consumer).

Most consumers that I know are risk-averse, especially when it comes to large sums of income. In terms of choices over states, risk aversion corresponds to convex indifference curves, like the blue curves in Figure 4.11. If a risk-averse consumer were given the option to pick a point from the green line he would definitely choose $E_2$, the point that has no risk. In other words, if I am to choose among different combinations of $h$ and $s$ that have the same expected value, then I’d rather go with the one that has no risk, which in this case is point $E_2$. In fact, at $E_2$ I reach indifference curve $U_3$, which corresponds to higher consumer utility than any other point along the
green curve.

Unfortunately, one cannot simply say I’d like to be in \( E_2 \) rather than \( E_1 \). One can, however, buy insurance and get to a combination of \( h \) and \( s \) that is better than the initial \((c_1, s_1)\) combination. In the context of our framework, insurance consists in renouncing to part of \( h \) in exchange for an increase in \( s \). Specifically, I pay a premium \( p \) but if I get sick — in other words, if state \( s \) happens — then I receive a payment \( q \) from the insurance company. This implies that, starting from \((h_1, s_1)\), my “bundle” is now given by \((h - p, s + q)\).

The rate at which the consumer trades off \( p \) for \( q \) depends on the conditions offered by the insurance company, specifically on the insurance premium charged by the insurance firm. An example of an insurance policy is given by the red line in Figure 4.11. Notice that the slope of the red line is lower (in absolute value) than the slope of the green line. The green line corresponds to constant expected value. If an insurance company were to offer a premium corresponding to the green line than on average the insurance company would make no profit. Since the insurance company must pay costs and make some profit, they offer less \( q \) for the same amount of \( p \) than the green line (or, they ask for more \( p \) for the same \( q \)).

Given the insurance policy corresponding to the red line, the consumer’s optimal choice corresponds to point \( E_3 \). Basically the consumer “buys” a compensation \( s_3 - s_1 \) in case she is sick. The cost of buying this contingent compensation is given by \( h_1 - h_3 \). Notice that point \( E_3 \) is associated with the indifference curve farthest to the NE (in this case curve \( U_2 \)) and still within the consumer budget set (which is determined by the red line). In sum, by buying insurance, the consumer moves from utility level \( U_1 \) to utility level \( U_3 \). It pays to buy insurance!

To conclude this section, we introduce two important concepts which you will likely come across in future economics and finance courses: certainty equivalent and risk premium. Consider again our consumer’s outlook in Figure 4.11 assuming that no insurance is bought, that is, assuming the consumer is in point \( E_1 \). One question one might ask is how averse to risk this particular consumer is. As we saw earlier, risk aversion corresponds to convex indifference curves. Broadly speaking, the more concave the indifference curves are, the more risk averse the consumer is. Is there a more quantitative way of measuring this? Yes. First, we define the concept of certainty
4.3. OTHER HOUSEHOLD DECISIONS

equivalent}, the payoff level for sure such that the consumer is indifferent with respect to the uncertain outcome \((h, s)\). Specifically, the certainty equivalent corresponding to the uncertain outcome \(E_1\) is given by \(E_4\), the point that lies on the 45° line and on the same indifference curve as \(E_1\). In other words, the consumer is indifferent between the uncertain outcomes \((h_1, s_1)\) and the certain outcome \((h_4, s_4)\). Then \(h_4 = c_4\) is the certainty equivalent to \((h_1, s_1)\).

Finally, the risk premium is defined as the amount an individual is willing to give up to replace an uncertain outcome with a sure one. In terms of Figure 4.11, the risk premium associated with \(E_1\) is given by \(h_1 - h_4\). The more risk averse a consumer is, the greater her risk premium.

FERTILITY

In 1968, a Stanford professor published a highly controversial book, The Population Bomb, where he prophesies that

The battle to feed all of humanity is over. In the 1970s hundreds of millions of people will starve to death.

This was not an isolated warning. Just a year before, the best-seller Famine 1975! had made similar predictions. Figure 4.12 marks the time when these books was published, as well as the evolution of
three key variables: world population, world GDP (constant prices), and the world average fertility rate. The fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year. At the time when the dire predictions were made, world population was at about 4 billion and increasing rapidly; fertility rates were higher than 5 children per woman; and per-capita GDP (in 2010 prices) was less than \((20 \times 10^{12}) / (4 \times 10^9)\), or simply $5,000. By 2020, per-capita GDP (in 2010 prices) is about \((85 \times 10^{12}) / (7.5 \times 10^9)\), a little over $11,000, that is, more than twice the value in the late 1960s. World population has almost doubled (we are now close to 8 billion), whereas the fertility rate dropped to 2.5, about one half of what it was in the late 1960s.

Clearly, the world has plunged into a series of crises (inequality, social unrest, climate change, to name a few), but mass famines were not the main problems of the past few decades. What the doomsayers of the late 1960s missed is that there are adjustment mechanisms which respond to pressures such as rapid population growth. The green revolution, which we dealt with in Section 1.2, was one of them.

Understanding these phenomena is important, lest we blindly project trends and follow the recommendations that come with those projections. *The Population Bomb*, for example, suggested taxes on children, luxury taxes on childcare goods, support for sex-selective abortion (to account for families that have a preference for boys), and even floated the idea of adding “temporary sterilants” to the water supply and staple foods.

What does economics have to say about this? One of the noticeable trends in Figure 4.12 is the sharp decline in the fertility rate in the past half century. In some cases, this resulted from coercion (e.g., China’s one-child policy) or other public policies. In most cases, however, it resulted from household decisions. The goal of this section is precisely to better understand fertility rates as an economic household decision. It may seem a bit crass to think of childbearing as a purely economic decision: defining children as an economic good, measuring the price of children, etc. Obviously it’s not just an economic decision. However, it is a decision with important economic components and certainly with important economic conse-
Women’s access to education and the job market is frequently associated with a drop in fertility rates.

Therefore, following the methodological approach of many particular sciences, we narrow our perspective to an economic model of choice with the goal of understanding how a variety of economic factors impact the evolution of the fertility rate.

The microeconomics approach is to **model** the tradeoffs between having children and consuming other goods. The idea is that having a child implies a series of costs. For starters, there’s the mother’s personal cost of childbirth. But there are many other economic costs: direct costs such as food, housing, clothing, schooling, etc; and, equally important, the opportunity cost implied by the time spent with children, namely the opportunity cost in terms of foregone labor income.

By now, we know how to represent the choice between two goods. In this regard, there isn’t much new about the present application. In Figure 4.13, we measure the number of children (**fertility rate**) on the horizontal axis and consumption on the vertical axis. Since consumption is measured in $, the “price” of consumption is 1 (it costs $1 to buy $1 of consumption). Suppose the household initially has an income of $I_1$. Then one option is for the household to have zero children and consume $I_1$. Let $p_1$ be the “price” of one child. As per the previous paragraph, we should think of $p_1$ as including both the direct costs as well as the opportunity costs of having a child. Making the simplifying assumption that this per-child cost is the same regardless of the number of children, we conclude that the household is faced with a budget constraint to its consumption and childbearing decision, namely the budget line $b_1$. In particular, if the household were to spend all of its income in children it would be able to “afford” $I_1/p_1$ children. Next we assume that the household’s pref-
4.3. OTHER HOUSEHOLD DECISIONS

References can be represented by an indifference curve mapping. And finally we find the optimal combination of consumption and number of children, which in this case is given by $E_1$, a point corresponding to $n_1$ children and a consumption level $c_1$.

Figure 4.12 shows that per-capita income doubled since the late 1960s, whereas the fertility rate dropped to a half of what it was half a century ago. A tantalizing conclusion is that children are an inferior good, that is, a good the consumption of which decreases as income increases. In addition to the time-series correlation, cross-family and cross-country evidence also suggests that family size and income are negatively correlated: high income families tend to have few children and, analogously, high income countries tend to have low fertility rates.

However, if you read Chapter 2 you will remember that one of the most important points of methodology to keep in mind is that correlation does not imply causality. In the present context, in addition to an increase in income we must also consider that the “price” of children may have changed in the past half century or so. It did, for two reasons. First, the direct costs with childbearing have increased. Second, and most important, the opportunity cost has increased considerably. The average educational attainment of women increased in most countries of the world, and so have their job market opportunities. This implies that, from a labor market point of view, women
have more to lose from childbearing activities. (There is a third factor underlying the trend in Figure 4.12, namely the changes in contraceptive technology, which allowed for a reduction in unplanned and/or unwanted pregnancies. Important as this was, it falls out of the focus on Figure 4.13.)

So, if we want to compare a typical household in 2020 to a typical household in 1970 we should consider not only an increase in income but also an increase in the price of children. Figure 4.13 depicts a plausible new equilibrium — point $E_2$ — which results in both an increase in $l$ and an increase in $p$ with respect to point $E_1$. Note that at $E_2$, which might represent the current situation, we observe a drop in $n$ (that is, $n_2 < n_1$) and an increase in $c$ (that is, $c_2 > c_1$), both of which are reflected in the data. Although it is not obvious from the aggregate data, more detailed studies suggest that children are a “normal” good, that is, for most households an increase in income leads to an increase in family size. The problem with Figure 4.12 is that the negative substitution effect of the increase in $p$ more than outweighs the positive effect of an increase in household income.

The drop in fertility rates is in part explained by the increase in the opportunity cost of childbearing (and not by the increase in income).

If the substitution effect of changes in the price of children is so im-
important, then it makes sense to think of public policy as it affects such price and the trade-offs that follow from it. In the 21st century several European countries suffer from the opposite problem than *The Population Bomb* warned about sixty years ago: Italy, Spain, Portugal, Greece and other countries suffer from very low fertility rates, so low that population size is declining. To be sure, some of the factors underlying the increase in the “price” of children are good news: women have better access to education and the labor market than before. However, there are many other components of this price that can be lowered. Government spending on family benefits, in particular, is likely to contribute to an increase in fertility rate by virtue of decreasing the “price” of children.

Figure 4.14 illustrates some of these points. For a series of OECD countries — mostly European countries but also including US, Canada, Japan and Australia — it plots the labor market participation — i.e., the percentage of women holding jobs — on the horizontal axis; and fertility — i.e., the average number of children per woman — on the vertical axis. Looking at the cloud of points as a whole, there does no seem to be any recognizable pattern. However, if we group the observations by the size of government benefits directed at the family (above median in blue, below median in red) then we do observe a bit of a pattern: Countries where the government provides higher levels of family benefits tend to have higher fertility rates. It is also noticeable that Scandinavian countries such as Sweden, Norway and Iceland, show both higher rates of female labor market participation and higher fertility rates. Of course, there are a lot of other factors. In this sense, it helps to compare countries that are closer to each other, for example France, Italy and Spain. Of these three, the only country with above-mean government support is also the one with (significantly) higher fertility rate. In sum, both the theoretical analysis and the empirical evidence suggest that policy measures which decrease the “price” of children do result in higher fertility rates.
### Key Concepts

- budget set
- budget line
- comparative statics
- normal good
- inferior good
- substitution effect
- income effect
- law of demand
- real income
- carbon tax
- price taker
- labor supply
- representative agent
- marginal
- life-cycle optimization
- savings function
- state of the world
- risk aversion
- risk neutral
- expected value
- insurance premium
- certainty equivalent
- risk premium
- fertility rate
In Chapter 3 we focused on economic decision making at a generic level. In Chapter 4, we covered household decision making (consumption, labor supply, etc). It is now the time to focus on economic decisions by firms.

Economists model firms as organizations that transform inputs into outputs. General Motors own a series of factories with a series of machines and employs a number of workers and managers to produce cars. Similarly, New York University may be thought of as an organization that owns buildings and employs faculty and staff in order to offer educational services.

Here’s the roadmap for this chapter. In Section 5.1, we introduce the concept of a firm’s production function, the mapping that describes the transformation of inputs into outputs. Next, in Section 5.2 we focus in one of the firm’s important economic choices, namely picking the optimal input mix (workers, machines, materials, energy, etc). Finally, in Section 5.3 we look at the choice of price and output level. In all cases, we apply the basic framework introduced in Chapter 3.

5.1. PRODUCTION FUNCTION

At the risk of oversimplifying, we can think of a firm as a process of transforming inputs into outputs. This is easier to see for a firm that
5.1. PRODUCTION FUNCTION

makes actual things. For example, a bagel bakery uses water, flour and other ingredients, together with machinery (an oven) and labor (someone has to put it all together), to produce tasty bagels. Firms that offer services also go through a similar process. For example, a consulting firm uses hours of labor — many, many hours, I’m told — together with some capital (mainly laptop computers) and materials (paper and paper clips), to produce solid advice to corporations that need it. The firm’s **production function** is the mapping that tells us, for a given set of inputs, how much output a firm is able to produce. Normally, this depends on the particular firm, as some firms are more efficient than others at transforming inputs into outputs. It also depends on the quality of inputs, for example skilled versus unskilled labor.

**PRODUCTION WITH ONE INPUT**

In Section 3.1, we derived Alexei’s feasible set when faced with the choice between leisure and grade. We took as given the following relation between hours of study and Alexei’s grade:

<table>
<thead>
<tr>
<th>Study hours</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>0</td>
<td>20</td>
<td>33</td>
<td>42</td>
<td>50</td>
<td>57</td>
<td>63</td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study hours</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>74</td>
<td>78</td>
<td>81</td>
<td>84</td>
<td>86</td>
<td>88</td>
<td>89</td>
<td>90</td>
</tr>
</tbody>
</table>

This table relates one input (hours of study) to an output (grade). Normally we associate production functions to firms, but by analogy we can also think of Alexei as having a course-grade production function. Figure 5.1 displays the various points of Alexei’s production function. On the horizontal axis we measure input level (in the present case, hours of study), whereas on the vertical axis we measure output (in the present case, course grade). The figure also includes a line that connects the points from the table above, reflecting the assumption that Alexei can also choose fractional values of the input variable “hours of study”.

We now define two important concepts based on the production function. First, **average product** (or simply AP) is defined as the ratio between output level and input level. In this case, this corresponds
to grade per hour of study. The second, related, concept is marginal product (or simply MP), defined as the additional output level resulting from a one-unit increase in input level. In this case, this corresponds to grade per additional hour. In general, both AP and MP depend on the input level at which they are computed. Therefore, we can compute the AP function and the MP function. This we do in Table 5.1. Suppose, for example, that Alexei is currently studying 6 hours a day. This implies that he is shooting for a grade of 63. His average product is then given by \( \frac{63}{6} = 10.5 \). Had Alexei studied for 5 hours instead of 6, he would have been aiming for a grade of 57. This implies that the 6th hour of study gives Alexei a 6 point kick in grade. We thus say that marginal product at 6 hours of study is 6 points.

Based on the values of this table, we can construct Figure 5.2, where again the input level is measured on the horizontal axis. One first observation from this figure is that marginal product is decreasing. This corresponds to a very important economics “law”, the law of decreasing marginal returns. The idea is that each additional unit of the input (hours of study) has a lower contribution than the previous ones. In Section 6.1 we will see that many production functions have increasing MP for low input levels and then decreasing MP for higher input levels. A more general “law” of decreasing marginal returns indicates that the MP of a given input is decreasing for a suf-
5.1. PRODUCTION FUNCTION

<table>
<thead>
<tr>
<th>Hours</th>
<th>Grade</th>
<th>AP</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>20.00</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>16.50</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>14.00</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>12.50</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>57</td>
<td>11.40</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>10.50</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>9.86</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>74</td>
<td>9.25</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>8.67</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>8.10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>84</td>
<td>7.64</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>7.17</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>88</td>
<td>6.77</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>89</td>
<td>6.36</td>
<td>1</td>
</tr>
<tr>
<td>15+</td>
<td>90</td>
<td>6.00</td>
<td>1</td>
</tr>
</tbody>
</table>

ficiently high level of that input.

There are a number of justifications for decreasing marginal returns. One of the most important ones is sequencing of actions. Suppose that I study better during the morning, say from 8am to 12pm, not so well after lunch, and very poorly at night. If I decide to study only for 2 hours a day then I would study in the morning, when those hours of study have a greater impact (and absent other considerations against studying in the morning). If instead of 2 hours I decide to study for 6 hours, then I will optimally study 4 hours in the morning and 2 in the afternoon, knowing that the afternoon hours will not be as productive as the morning ones. In this context, the marginal effect of the sixth hour of study (in the afternoon) is lower than the marginal effect of the first hour (in the morning).

Another explanation, one that we will explore later in the course, is related to production with multiple inputs. Consider, for example, the production function of restaurant meals. It requires two inputs,
kitchen space (in square feet) and workers. Fixing the input “kitchen space”, you can see how the contribution of an additional worker would decline as more workers are added. In fact, at some point, the marginal product of a worker might actually be negative! Too many cooks spoil the broth, so the saying goes.

There are multiple justifications for decreasing marginal product, including optimal sequencing and fixed production factors.

Figure 5.2 shows that, in additional to marginal product, average product is also decreasing. There is an economic intuition for this. At one hour of study, MP and AP are the same. Because of decreasing marginal returns, MP at two hours of study is less than MP at one hour of study, which in turn is equal to AP at one hour of study. In words, this means that the second hour contributes to grade less than the average hour up to then. This implies that that second hour of study effectively lowers the average. More generally, if MP is lower than AP, then an additional hour of study effectively lowers AP, which in turn implies that AP is decreasing.

Average product is greater than marginal product. Both are decreasing in output level.
To conclude this section, Figure 5.3 plots a generic production function. There is one input, $x$, and one output, $y$. As in the case of Alexei’s grade production function, we see that the relation between $x$ and $y$ is a concave function. A concave production function corresponds to a decreasing MP mapping. In fact, when we have a continuous production function such as $y = f(x)$ in Figure 5.3, MP is given by the slope of the tangent to $f(x)$; and the tangent becomes flatter as $x$ increases. (As we will see in Section 6.1, many production functions are convex at $x = 0$ and then become concave as the production function in Figure 5.3, so to be rigorous the present production function is not exactly generic.)

The relation between AP and MP is also illustrated in Figure 5.3. AP is defined by the ratio $y/x$. In graphic terms, this corresponds to the slope of the segment that extends from the origin to $(x^*, y^*)$, for a generic input level $x^*$. As can be seen, this slope is greater than the slope of the tangent (that is, MP is less than AP). Moreover, the slope of the $(x^*, y^*)$ segment is lower the greater the value of $x^*$ is (that is, AP is decreasing). In sum, for production functions like the one in Figure 5.3, average product is greater than marginal product and both average and marginal product are decreasing in output level.
PRODUCTION WITH MULTIPLE INPUTS

Normally firms use more than one input. In this case, the production function may be written as \( f(x_1, \ldots, x_n) \), where \( x_i \) stands for the quantity of input \( i \). For the purpose of this section, we consider two inputs: capital and labor. This is not to say that other inputs are not relevant: no matter how many ovens and oven operators you have, you cannot make bagels without flour. It’s just that, for the purposes of illustrating the main principles, it suffices to consider two inputs. Moreover, in many examples (e.g., consulting services) these are indeed the main inputs into production (in other words, paper and paper clips are a small fraction of the consulting firm’s operations). We follow the convention of denoting capital inputs by \( K \) and labor inputs by \( L \).

When there is more than one input, we cannot represent the production function in a two-dimensional graph. One alternative is to display the value of \( f(K, L) \) for each combination of \( K \) and \( L \) on a double-entry table. Table 5.2 considers one particular example. Each row corresponds to a different value of \( K \), from 1 to 6. Each column corresponds to a different value of \( L \), also from 1 to 6. The value of each cell is the output value corresponding to each combination of inputs. For example, 2 units of capital and 3 units of labor yield an output of 244. Before continuing, notice that the values in Table 5.2 satisfy the property of decreasing marginal returns. For example, if capital is set at \( K = 3 \), then the incremental output gain from each one-unit increase in labor input is given by 173, 71, 56, 46, 41, 37. See also Exercise 15.15 in the companion Exercises in Microeconomics book.

One interesting fact about Table 5.2 is that multiple combinations of \( K \) and \( L \) may yield the same output level. For example, \( K = 6 \) and \( L = 1 \) yield the same output as \( K = 2 \) and \( L = 3 \). Figure 5.4 takes this idea one step further and plots, on a graph with \( K \) on the vertical axis and \( L \) on the horizontal axis, a series of isoquants. An isoquant is a line on the \((L,K)\) map such that all points correspond to the same output level. Notice the close parallel with indifference curves, which we introduced in Section 3.2. Indifference curves connect points yielding the same utility to an economic agent (consumer, worker, etc). Isoquants, by contrast, connect points yielding the same output level. For example, the \( q = 224 \) isoquant in Figure 5.4 in-
5.1. PRODUCTION FUNCTION

TABLE 5.2
Production with two variable inputs

<table>
<thead>
<tr>
<th>K ↓ L →</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>141</td>
<td>173</td>
<td>200</td>
<td>223</td>
<td>244</td>
</tr>
<tr>
<td>2</td>
<td>141</td>
<td>200</td>
<td>244</td>
<td>282</td>
<td>316</td>
<td>346</td>
</tr>
<tr>
<td>3</td>
<td>173</td>
<td>244</td>
<td>300</td>
<td>346</td>
<td>387</td>
<td>423</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>282</td>
<td>346</td>
<td>400</td>
<td>447</td>
<td>489</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>316</td>
<td>387</td>
<td>447</td>
<td>500</td>
<td>547</td>
</tr>
<tr>
<td>6</td>
<td>244</td>
<td>346</td>
<td>423</td>
<td>489</td>
<td>547</td>
<td>600</td>
</tr>
</tbody>
</table>

cludes point A ($K = 3, L = 2$) as well as point B ($K = 1, L = 6$). As can be seen from Table 5.2, these two alternative combinations yield the same output level, namely $q = 224$. Similarly, points C ($K = 6, L = 2$) and D ($K = 3, L = 4$) belong to the same isoquant, this time the $q = 346$ isoquant.

The parallel between isoquants and indifference curves can be taken one step further. In Section 3.2 we introduced the concept of marginal rate of substitution, the rate at which an economic agent is willing to give up one good in order to obtain an additional unit of another good. Graphically, the MRS corresponds to the absolute value of the slope of the indifference curve containing each point under consideration. Similarly, the marginal rate of technical substitution (MRTS) measures how much extra $K$ the firm needs when it loses one unit of $L$ so as to keep the same output level; or, alternatively, how much capital a firm can save by employing an additional unit of labor so as to keep the same output level.

The shape of a production function’s isoquants indicates the degree to which the inputs are substitutes or complements. Consider the production function of commercial flights. At the risk of simplifying the problem, let us say you need one pilot and one plane in order to offer one flight. In this context, given that an airline owns one plane, having more pilots won’t help; you still can only offer one flight. Given that the airline employs one pilot, having more than one plane won’t help either; you still can only offer one flight. The pilot-plane example corresponds to the extreme of perfect complements.
**Leontief production functions.** In this case, isoquants are L-shaped as in the bottom panel of Figure 5.8. Once again, there is a parallel with indifference curves, where two goods may also be perfect complements. Examples include peanut butter and jelly, cars and tires, and so on. See also Exercise 3.12 in the companion book.

At the opposite end, consider McDonald’s production function. Suppose the fast-food chain uses both Texas and Nebraska beef as inputs. Moreover, for the purpose of this example, suppose that beef is the only ingredient into burgers (I am aware this is a strong assumption, but please bear with me). At the risk of offending the great states of Texas and Nebraska, it seems reasonable to assume that the quality of beef in these two states is similar. Therefore, the quantity and quality of McDonald’s burgers depends on the total quantity of beef, not on the particular proportions of beef from Texas or Nebraska. We thus have a case when the inputs are **perfect substitutes**. Perfect substitute inputs lead to straight isoquants (as in the top panel of Figure 5.8): one unit of output (burger) can be obtained with one unit of Texas beef and zero units of Nebraska beef; or one unit of Nebraska beef and zero units of Texas beef; or any combination therein. More generally,

*The closer substitutes (resp. complements) two production inputs are, the closer the isoquants are to straight lines (resp. L-shaped lines).*
5.2. INPUT MIX

Over the years, across countries and across industries we see significant variation in the capital-labor input mix. Take for example auto manufacturing. Comparing a GM factory from the 1960s to a Tesla factory of the 2010s we observe a significant increase in the capital/labor ratio. What determines the optimal mix of labor and capital as inputs? What is the effect of changes in the interest rate, the wage rate, technical progress (e.g., artificial intelligence), and so forth? These are some of the questions that economists address by using the framework developed in this chapter.

In the present context, a firm’s optimal problem can be thought of in two ways: (a) for a given cost level, what is the input mix that maximizes output level; and (b) for a given output level, what is the input mix that minimizes cost. We stress “in the present context” because there are many other choice problems a firm must consider. One of them, output pricing, will be considered in the next section.

Consider first the output maximization problem, which turns out to be very similar to the problem considered in Chapters 3 and 4. Let $C$ be the value of input costs to be spent by the firm. Given this, we can draw an isocost line, namely a line connecting all input combinations that cost $C$. Specifically, the cost of inputs is given by $C = wL + rK$ and the line defined by the above equation is referred to as isocost line. Each value of $C$ corresponds to a different isocost line.

An isocost line is essentially the same as a budget line. Recall that the latter consists of combinations of goods $x$ and $y$ that cost $I$, where $I$ is a consumer’s income. The $y$-axis intercept is given by $I/p_y$, whereas the $x$-axis intercept is given by $I/p_x$. Similarly, the intercept of the isocost line on the $L$ axis is given by $C/w$, where $w$ is the unit cost of labor; and the intercept of the isocost line on the $K$ axis is given by $C/r$, where $r$ is the unit cost of capital. It follows that the slope of an isocost line is given by $-w/r$. Different isocost lines correspond to horizontal shifts, that is, they all have the same slope.

The highest output attainable with an input budget $C$ corresponds to the point where one of the isoquants is tangent to the isocost line. In the top panel of Figure 5.5, this is given by the input mix $m^*$. Since the absolute value of the slope of the isoquants is given
FIGURE 5.5
Optimum input mix (top) and cost minimization (bottom)

by the MRTS and the absolute value for the slope of the isocost line is given by \(w/r\), we conclude that the optimal input mix is given by MRTS = \(w/r\). The intuition for the optimality rule is similar to what we found in the context of the study (Section 3.3), consumption and work decisions. If MRTS were different than \(w/r\) then the firm could do better: for the same input budget, it would be able to obtain a higher output. The argument is similar to that in Chapters 3 and 4.

The second approach to a firm’s optimal input mix is to derive the minimum cost possible for a given output level. The bottom panel of Figure 5.5 illustrates this process. Consider a given output level \(q\) and its associated isoquant \(q\). As mentioned earlier, there are multiple isocost lines, each corresponding to a different cost level \(C\). The bottom panel in Figure 5.5 depicts several of these isocost lines. The
optimal input mix is given by the point on the desired isoquant $q$ that is associated with the lowest isocost line, that is, the one closest to the origin. In the bottom panel of Figure 5.5, this is given by the input combination $m^*$, the point where an isocost line is tangent to the isoquant.

The following two problems lead to the same input mix: (a) maximize output for a given cost level; and (b) minimize cost for a given output level.

The two panels in Figure 5.5 look similar, and there is a reason for that. Regardless of whether we look for the highest output level attainable with a budget $C$ or we look for the lowest cost $C$ consistent with output $q$ we are led to the same rule: at the optimum input mix, the marginal rate of technical substitution is equal to the input cost ratio, that is, $\text{MRTS} = \frac{w}{r}$. This is analogous to the by now famous equality presented on p. 102 (the one typeset in a very large font size).

**DECLINING COST OF CAPITAL**

One of the more noticeable trends in modern economies, beginning with the industrial revolution, is the rapid decline of the cost of capital as a production input. In recent years, this trend has been very noticeable in digital industries. For example, Figure 5.6 documents the steep decline in cost of computing over the past decades. Up until 1940, the cost of computation was approximately constant. Since then, it has declined at an exponential rate. To be specific, the cost of a standard computation has declined at an average annual rate of
53% per year over the period 1940–2012! The exclamation mark reflects both the number 53 and the length of the period (72 years).

More generally, the cost of capital inputs keeps declining. Figure 5.7 illustrates the effect of a drop in the cost of capital inputs, from \( r_1 \) to \( r_2 \). Suppose that a given firm continues to operate with the same input budget \( C \). This is not entirely realistic: we would expect the firm to re-optimize its input and output decisions, leading probably to a different level of resources spent on acquiring inputs. However, for the purpose of illustrating the impact of a change in input costs, we will stick to this simplifying assumption.

A lowering of the cost of capital from \( r_1 \) to \( r_2 \) implies that the \( C \) isocost curve pivots around the \( L \) axis. Specifically, the \( L \)-axis intercept remains at \( C/w \), whereas the \( K \)-axis intercept increases from \( C/r_1 \) to \( C/r_2 \). If initially the optimum was given by \( m_1 \), the new optimum is now given by \( m_2 \). As expected, the new optimal solution corresponds to a higher output level, an increase from \( q_1 \) to \( q_2 \). Also as expected, the amount of capital used in production increases from \( K_1 \) to \( K_2 \). It is not clear in general what to expect regarding labor input levels. In Figure 5.7, the value of \( L \) remains unchanged with the decrease in the cost of capital. But this need not be the case in general. As we will see next, the effect of a change in the cost of capital on the use of labor as a production input depends greatly on the degree of substitutability between capital and labor, which in turn is
reflected in the shape of the isoquant curves.

Consider first the extreme case when capital and labor inputs are perfect substitutes. For example, one supermarket checkout counter can be manned by either a person or an automatic checkout machine. The top panel in Figure 5.8 illustrates this case. As mentioned earlier, when inputs are perfect substitutes isoquants are straight lines. Suppose that, initially, the cost of labor is given by \( w \), whereas the cost of capital is given by \( r_1 \). For a given level of total input costs \( C \), we have an isocost line stretching from \( C/r_1 \) on the vertical axis \( (K) \) to \( C/w \) on the horizontal axis \( (L) \). Since the isoquants are straight lines, we do not have a tangency point as in previous cases. Rather, the optimal solution is a “corner” solution, given by input mix \( m_1 \). In other words, with input costs \( w \) and \( r_1 \) the firm minimizes cost (or maximizes output) by using labor as the only input. Suppose that the cost of capital drops from \( r_1 \) to \( r_2 \) as shown in the top panel of Figure 5.8. Then the optimal input mix remains the same. In fact, since the firm uses no capital inputs, the change in the cost of capital has no effect on cost level or output level. However, there is a threshold level of the cost of capital beyond which the firm completely flips its input choice, from all labor and no capital to all capital and no labor. For example, if the cost of capital is given by \( r_3 \) then, with the same total input budget \( C \), the firm is able to produce a higher output level by using only capital than by using only labor. (Can you determine the
level of cost of capital at which the firm flips its input mix?)

Consider the opposite extreme: perfect complements. For example, an MRI machine requires three technicians to operate it and this proportion is not flexible. The perfect complements case is illustrated in the bottom panel of Figure 5.8. Initially the cost of labor is $w$ and the cost of capital $r_1$. For a given level of total input costs $C$, we have an isocost line stretching from $C/r_1$ on the vertical axis ($K$) to $C/w$ on the horizontal axis ($L$). Since the isoquants are L shaped, we do not have a tangency point as in previous cases. Rather, the optimal solution is at the kink of the L-shaped isoquant, that is, input mix $m_1$. In other words, with input costs $w$ and $r_1$ the firm minimizes cost (or maximizes output) by using the fixed proportion of labor and capi-
In October 2019, United Parcel Service (UPS) was awarded a certification that allows it to use drones on medical campuses. The certification allows UPS to fly drones beyond the visual line of sight in order to deliver health care supplies to various hospitals. UPS hopes this will be the first step to being able to deliver to homes and rural areas. According to UPS, delivery by drone is cheaper than the cost of a person driving a car.

The FAA has already granted a certificate to Wing, the drone-delivery unit of Google. However, the two certifications are different. Wing is allowed to use only one pilot and one drone at a time, while UPS is allowed to use several pilots and numerous drones simultaneously. Other countries have also begun drone deliveries of vital medical supplies. Zipline, for example, distributes blood in Rwanda using drones, whereas Swoop Aero delivers vaccines and other medical supplies in the Pacific.

What impact will drone technologies have on jobs? As often is the case, the most likely scenario is that the new technologies will both destroy jobs and create new jobs. Specifically, many truck delivery operations will be discontinued, and so many truck drivers will lose their jobs. At the same time, many drone driving jobs will be created.

Box 5.1: Delivery drones

In October 2019, United Parcel Service (UPS) was awarded a certification that allows it to use drones on medical campuses. The certification allows UPS to fly drones beyond the visual line of sight in order to deliver health care supplies to various hospitals. UPS hopes this will be the first step to being able to deliver to homes and rural areas. According to UPS, delivery by drone is cheaper than the cost of a person driving a car.

The FAA has already granted a certificate to Wing, the drone-delivery unit of Google. However, the two certifications are different. Wing is allowed to use only one pilot and one drone at a time, while UPS is allowed to use several pilots and numerous drones simultaneously. Other countries have also begun drone deliveries of vital medical supplies. Zipline, for example, distributes blood in Rwanda using drones, whereas Swoop Aero delivers vaccines and other medical supplies in the Pacific.

What impact will drone technologies have on jobs? As often is the case, the most likely scenario is that the new technologies will both destroy jobs and create new jobs. Specifically, many truck delivery operations will be discontinued, and so many truck drivers will lose their jobs. At the same time, many drone driving jobs will be created.

The two panels in Figures 5.8 correspond to extreme situations. If inputs are perfect substitutes, then a decrease in the cost of capital will eventually lead to a drop in the use of labor inputs — in fact, a rather drastic drop from labor-only production to capital-only pro-
duction. At the opposite end, if inputs are perfect complements, then a decrease in the cost of capital leads to greater use of both capital and labor inputs. Generally speaking, when we talk about the impact of new technology on jobs this is the critical question that needs to be addressed: to what extent are the new technologies a substitute or a complement to labor inputs.

**AI, ROBOTS AND JOBS**

The above discussion on the relation between capital and labor inputs might be summarized as follows:

*The impact of technology on jobs depends largely on the degree of substitutability between capital and labor inputs.*

What does this all imply regarding the impact of AI and robots on employment? The first general point is that one cannot generalize: In some cases, new AI technologies are a complements to labor, thereby increasing labor productivity. For example, medical support systems allow doctors to treat more patients and to do a better job with each one of them (i.e., more precise diagnostic). In other cases, robots are clearly labor substitutes, thereby eliminating jobs. For example, supermarket checkout machines, ATM machines, and production-line robots are nearly perfect substitutes for labor.

But even when new technologies replace existing jobs it’s not clear that, overall, they create unemployment. If history is a good example, technological progress has destroyed many jobs but never really created unemployment. This is not a universally accepted opinion, I should add. Albert Einstein, for example, believed that the massive increase in unemployment during the Great Depression was largely due to the increase availability of capital. In 1933 he stated that

> It cannot be doubted that ... the improvement in the apparatus of production through technical invention and organization has decreased the need for human labor, and thereby caused the elimination of a part of labor from the economic circuit.

In my opinion, Einstein was better at physics than at economics.
5.2. INPUT MIX

Figure 5.9
Returns to scale: increasing returns to scale (top) and decreasing returns to scale (bottom)

RETURNS TO SCALE

Earlier in this chapter (and in Chapter 3) we introduced the law of decreasing marginal returns: as you add more of a given input, you typically get a lower and lower additional output. One of the reasons for this pattern is that we are holding all other factors constant. Try, for example, adding more and more cooks (labor input) into one single kitchen (capital input).

What if we change all inputs at the same time, and in the same proportion? The answer to this question characterizes a firm’s **returns to scale**. There are basically three different possibilities: increasing, constant, or decreasing returns to scale.

We say that a production function exhibits **increasing returns to**
5.2. INPUT MIX

scale (IRS) if, when input levels are doubled, output more than doubles. In terms of isoquants, this implies that isoquants corresponding to constant output increases are “closer” at higher quantities of both inputs. This is illustrated in the top panel of Figure 5.9.

Firms subject to IRS are able to produce large output levels at a low cost. This implies that one firm is more efficient than many firms at producing the same total output level, a situation referred to as natural monopoly. An example might be given by electric utilities.

What causes IRS? One first factor is the presence of indivisibilities: some factors of production are only available in large sizes. For example, it’s difficult if not impossible — but certainly uneconomical — to produce a small nuclear power plant. Another source of IRS is given by specialization (already discussed in Section 1.2). As production scale increases, workers can specialize and focus on tasks that match their skills. Also, firms can use specialized machinery. Can you think of examples?

The opposite of increasing returns to scale is — you guessed it — decreasing returns to scale. We say that a production function exhibits decreasing returns to scale (DRS) if, when input levels are doubled, output increases by a factor of less than 2. In terms of isoquants, this implies that isoquants corresponding to constant output increases are “farther apart” at higher quantities of both inputs. This is illustrated in the bottom panel of Figure 5.9.

What causes DRS? An important factor is what we might refer to as management complexity. Suppose that GE increases by ten-fold in number of plants, lines of business, etc. Even a great CEO like Jack Welch would have a hard time managing such a gigantic firm. A similar problem is the difficulty of communicating through larger organizations and the emergence of “corporate red tape”, as I may attest based on my experience as a GM intern. Similarly, if Boeing wanted to double its size to scale it would need to double all inputs, including specialized engineers, some of which are very hard to find.

Finally, it can also be the case that, as we double a firm’s inputs, its output exactly doubles. We refer to this case as constant returns to scale (CRS). In terms of the isoquant map, this is the case when isoquants corresponding to a constant output increment are equidistant.
The term productivity is used frequently to describe a firm’s performance or the performance of its production factors. Unfortunately, it means different things to different people and as a result may be a source of confusion.

One first important concept is that of average labor productivity, or simply labor productivity. This is defined by \( q/p/L \), where \( q \) is number of units produced, \( p \) is the price at which each unit is sold, and \( L \) is the number of workers. For simplicity, suppose that price is equal to 1, that is, \( p = 1 \) (in the next section we deal with the issue of product pricing). Then labor productivity is simply \( q/L \). Consider the firm’s production function corresponding to the top panel in Figure 5.3. (This is the same as Table 5.2, as is reproduced here for your reading convenience.) If the firm has \( K = 2 \) units of capital and employs 4 units of labor, then labor productivity is given by \( 282/4 = 70.5 \). Were the firm to increase labor inputs to 5 units of labor, labor productivity would fall to \( 316/5 = 63.2 \). You should have expected labor productivity to decline. As we have seen before, declining MP implies declining AP, and when we fix the quantity of one input (e.g., capital), the marginal product of any other input declines as we use more of it.

A second important property of labor productivity is that it is increasing in \( K \). For example, if we keep \( L \) at 4 units but increase \( K \) from 2 to 3, then labor productivity increases from 70.5 to \( 346/4 = 86.5 \). This is important: as we compare labor productivity across firms or across sectors, we must take into account the relative levels of capital in each firm or sector. Even if we were to correct for inflation and other relevant factors, we would likely find that average labor productivity at a Tesla factory in 2020 is substantially higher than the average labor productivity of a Plymouth factory in 1960. It’s not so much the fact that Tesla workers are more skilled than Plymouth workers or that Tesla is better managed than GM (Plymouth’s parent company), rather that Tesla workers are combined with substantially more capital in 2020 than Plymouth workers were in 1960.

In this sense, a better measure of firm performance is how well it does with a given set of production factors. Suppose that the bottom panel of Figure 5.3 corresponds to the same product as the production function corresponding to the top panel. A firm that works ac-
5.2. INPUT MIX

TABLE 5.3
Two production functions with different levels of total factor productivity

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>141</td>
<td>173</td>
<td>200</td>
<td>223</td>
</tr>
<tr>
<td>2</td>
<td>141</td>
<td>200</td>
<td>244</td>
<td>282</td>
<td>316</td>
</tr>
<tr>
<td>3</td>
<td>173</td>
<td>244</td>
<td>300</td>
<td>346</td>
<td>387</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>282</td>
<td>346</td>
<td>400</td>
<td>447</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>316</td>
<td>387</td>
<td>447</td>
<td>500</td>
</tr>
<tr>
<td>6</td>
<td>244</td>
<td>346</td>
<td>423</td>
<td>489</td>
<td>547</td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>282</td>
<td>346</td>
<td>400</td>
<td>446</td>
</tr>
<tr>
<td>2</td>
<td>282</td>
<td>400</td>
<td>488</td>
<td>564</td>
<td>632</td>
</tr>
<tr>
<td>3</td>
<td>346</td>
<td>488</td>
<td>600</td>
<td>692</td>
<td>774</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>564</td>
<td>692</td>
<td>800</td>
<td>894</td>
</tr>
<tr>
<td>5</td>
<td>446</td>
<td>632</td>
<td>774</td>
<td>894</td>
<td>1000</td>
</tr>
<tr>
<td>6</td>
<td>488</td>
<td>692</td>
<td>846</td>
<td>978</td>
<td>1094</td>
</tr>
</tbody>
</table>

cording to the top production function and employs 2 units of labor and 3 units of capital produces an output of 244. With the same inputs \((K = 3, L = 2)\), the firm operating according to the bottom production function produces an output of 488, twice as much as the firm on the top panel. We would then say that the bottom firm’s total factor productivity is twice that of the top firm’s. Specifically, total factor productivity (TFP) measures a firm’s efficiency in the use of production inputs: it measures how much output a firm produces relative to other firms controlling for input levels.

These definitions and distinctions are important for a variety of reasons. For example, in Chapter 13 we will talk about international migration as a source of economic opportunity. Typically, workers in developing countries move to developed countries with a view to improve their economic condition. In terms of production functions, developed countries differ from developing countries in two ways. First, developed countries have a greater abundance of capital. Second, developed country firms tend to have greater total factor pro-
ductivity. A simplified description of a firm in a developing country would be the blue cell in the top panel of Figure 5.3, whereas a firm in a developed country would correspond to the red cell in the bottom panel. The same worker in the developed country has a much higher average and marginal productivity, partly because of the capital mix difference, partly because of the difference in TFP. You can see the attractiveness of migrating to a developed country.

5.3. OUTPUT LEVEL AND PRICE

So far we have been looking at a firm’s decisions regarding its inputs, in particular how to optimally combine inputs in order to produce a certain output level and minimum cost. But how should a firm determine its output level? Assuming that the firm faces a given downward-sloping demand that relates price and demand for the firm’s output, we can rephrase the problem as: How should a firm determine its price? In this section we focus on the problem of optimal choice of price (or output level).

ICE-CREAM PRICING

Consider a specific numerical example, in fact, one that is — as they say in Hollywood — “inspired by true events.” Rui, a young undergraduate economics major, once got a summer job selling ice-cream in Philadelphia. (There was such a student, by the way, and he took my microeconomics class. Not only that, but on his return from the summer job he made sure he let me know how useful the class had been for his summer job!) Rui operates his truck in a specific neighborhood of Philly, where he is the only vendor. After a few days of experimentation with different prices, Rui estimates that demand is given by \( q = 20 - 2p \), where \( p \) is price and \( q \) is quantity demanded. We will go through the details of consumer demand in Chapter 6. For now, suffice it to say that the demand curve faced by a firm relates the price it sets to the quantity demanded by consumers, that is, the maximum number of units the firm is able to sell at that price. We can also express the demand curve in inverse form. In the present case, inverse demand is given by \( p = 10 - q/2 \). Figure 5.10 illustrates the ice-cream demand curve faced by Rui. Rui’s costs are as follows.
5.3. OUTPUT LEVEL AND PRICE

Each hour he must pay $F = $15 for the truck rental. In addition, he must pay the ice-cream factory $c = $3 per unit sold. (Think of a unit as a box of 12 ice-cream bars.)

Given all this information, the question at hand is: what price should Rui set? Suppose Rui wants to maximize profit, which is given by

$$\pi = (p - c) q - F$$

Figure 5.10 plots a series of three isoprofit curves. An isoprofit curve is like an indifference curve, with the difference that instead of consumer utility it connects pairs $(p, q)$ corresponding to the same profit level. Isoprofit curves are downward sloping like indifference curves. The idea is that a firm can increase profit by increasing price or by increasing sales. In other words, price and output are substitute “goods” in “producing” profit.

As in the previous chapters, we have a problem of constrained optimization. The firm’s preference is to achieve an isoprofit curve as far from the origin as possible. The firm’s constrain is that its choice of $p$ and $q$ must fall on the demand curve. In other words, the demand curve corresponds to the frontier of the firm’s feasible set.

As in the previous chapters, the firm’s optimal choice corresponds to point $E$, the point where an isoprofit curve is tangent to the demand curve. In Figure 5.10, point $E$ corresponds to $q = 7$ and $p = 6.5$. Also as in the previous chapters, the firm’s optimal choice is
Ice cream truck at Columbus Circle, New York. If you had the choice, how would you price your ice cream?

given by the equality \( MRT = MRS \). What do these rates correspond to in the present case? The marginal rate of transformation is nothing but the slope (in absolute value) of the inverse demand curve, that is, \( |\Delta p/\Delta q| \).

The slope of the demand curve indicates the trade-off that the firm must face when balancing a price target with a sales target. If the firm wants to increase price, then it must accept selling a lower output level. Conversely, if the firm sets a higher sales target — that is, a higher value of \( q \) — then it must also set a lower price to go with it.

The firms’s marginal rate of substitution is a little more difficult to understand. The value of MRS indicates how much of a price decrease the firm is willing to accept in exchange for a one-unit increase in quantity sold. Since \( \pi = (p - c) q - F \), when the firm increases \( q \) by one unit its profit increases by \( p - c \). Moreover, when the firm decreases \( p \) by one unit its profit declines by \( q \). So, the greater the value of \( p - c \), the more the firm gains from increasing \( q \), which in turn implies that the firm is more willing to sacrifice a price decrease. Conversely, the greater the value of \( q \), the less willing the firm is to sacrifice a price decrease.

Putting it all together, we conclude that the firm’s marginal rate of substitution is given by \( MRS = (p - c)/q \). The greater the price-cost margin, that is, the greater \( p - c \), the more the firm is willing to increase \( q \) and lower \( p \) (steeper isoprofit curve). Conversely, the greater the current output level, \( q \), the less the firm is willing to increase \( q \) and lower \( p \) (flatter isoprofit curve). Finally, we conclude
that the rule for optimal price setting is that

$$\frac{\Delta p}{\Delta q} = \frac{p - c}{q}$$

or alternatively

$$\frac{p - c}{q} = \frac{1}{\left|\frac{\Delta q}{\Delta p}\right|}$$

In words, if demand is very sensitive to price changes — that is, if the slope $|\Delta q/\Delta p|$ is very high, then the seller optimally chooses a low margin $p - c$ (for a given value of $q$). This is intuitive: if a small price decrease leads to a large demand increase then the seller should decrease price and enjoy the large $\Delta q$ benefit at a relatively low cost from decreasing $p$ by $\Delta p$.

Figure 5.11 illustrates this idea. Specifically, we consider two different demand curves. Demand curve $D_1$ is relatively flat. This means that quantity demanded is very sensitive to price changes: a small change in price leads to a large change in quantity demanded. (This may seem a little confusing: Usually we think of a flat function as one where the dependent variable is not very sensitive to changes in the independent variable. The source of confusion is that sometimes we consider $p$ the dependent variable but most times we consider $q$, the variable on the $x$ axis, as the dependent variable. If you
are furious about this, please don’t blame me, blame Alfred Marshall, whose idea it was to represent $p$ on the vertical axis. That said, as we will see in Chapters 6 and 7, there is a “method” to Marshall’s “madness”). In contrast to $D_1$, demand curve $D_2$ is relatively steep. This means that quantity demanded is not very sensitive to price changes: a small change in price leads to a small change in quantity demanded. As Figure 5.11 shows, optimal price is higher when demand is given by $D_2$, that is, when demand is less sensitive to price changes.

**MARGINAL REVENUE AND MARGINAL COST**

There is more than one way to peel an orange — so they say — and this particular “orange” (optimal pricing) is sufficiently important to warrant additional approaches. Based on the information regarding demand and costs, Rui assembles the values listed in Table 5.4. The first two columns correspond to the demand curve $q = 20 - 2p$ (for a set of values of $p$). The third column shows total revenue for each price. This is simply price times quantity (first column times second column). The fourth column shows total cost: 15 plus 3 times the number of units sold (as given by the second column). The fifth and sixth columns will be discussed in detail below. Finally, the seventh column shows profit, the difference between the third and the forth columns.

Given all this information, what price should Rui set? Before continuing, notice that, since price and output are related by the demand curve (as shown in the first two columns of Table 5.4), it is the same thing to choose the optimal price or to choose the optimal output level. That is, even though the seller is assumed to set price and consumers choose quantity as a function of price, we can think of the seller as choosing the optimal quantity it wants consumers to buy and then setting the corresponding price. In what follows, we treat the seller’s decision as that of selecting an output level. Note also that, given the particular demand curve we consider, the sequence of declining prices in Table 5.4 corresponds to output increasing by units of 1 from row to row. This need not always be true, but it makes our life considerably easier.

So, to rephrase the earlier question: what level of unit sales should Rui optimally target? Economists like to think about these questions
by reasoning in terms of incremental, or marginal, decisions (cf Section 2.3). Specifically, let us first ask the question: is it better to sell one unit than to sell none (assuming the truck rental fee has already been paid)? Is it better to be in the first row (price equal to 10, zero sales), or the second one (price equal to 9.5, one unit of sales)?

In order to answer this question, we compute marginal revenue and marginal cost. The value of marginal revenue is shown in the fifth column of Table 5.4. For example, when setting price at 9.5, Rui is able to sell one unit. Compared to selling zero units (price equal to 10), this corresponds to a marginal revenue of 9.5, which is the difference between 9.5 (total revenue from selling one unit) and 0 (total revenue from selling no units). By the same token, the marginal revenue from selling 3 units instead of 2 is equal to 7.5 = 25.5 – 18; and so forth. Similarly to marginal revenue, we can also compute the values of marginal cost. Specifically, the marginal cost of selling one unit is given by 3 = 18 – 15. As can be seen in Table 5.4, this is also the marginal cost for all other units. (We will return to the important concept of marginal cost in Chapter ??.)

How do the concepts of marginal revenue and marginal cost help determine the optimal sales target? When considering the choice between selling zero units and selling one unit, Rui compares a
marginal revenue of 9.5 with a marginal cost of 3. Since 9.5 is greater than 3, it is better to sell one unit than to sell none.

Next we compare selling two units to selling one. The marginal revenue of the second unit is 7.5, whereas the marginal cost is only 3. Rui is therefore better off by selling two units than by selling only one. Continuing with this reasoning, we conclude that it is optimal to sell 7 units (by setting a price equal to 6.5). In fact, at this output level, a further increase to 8 would imply a marginal revenue of only 2.5, whereas the marginal cost would be 3.

The fact that a price of 6.5 and a sales target of 7 correspond to the optimal solution could also be gotten by simply looking at the rightmost column: the value of profit is maximal precisely where price equals 6.5. However, the marginal revenue versus marginal cost reasoning helps derive an important rule in economics: the level of output should be chosen so that the value of marginal revenue is as close to marginal cost as possible.

This may seem strange: if we want to maximize profit, then surely we want the difference between revenues and costs to be as high as possible. The solution to this apparent paradox is that one thing is the difference between revenues and costs; and a different thing is the difference between marginal revenue and marginal cost.

Figure 5.12 illustrates the marginal revenue-marginal cost approach to optimal pricing. The demand curve and the marginal cost curves are the same as in Figure 5.10. The novel element in Figure 5.12 is the marginal revenue curve, which we plot from the values in Table 5.4. The value of $q$ at which marginal revenue is equal to marginal cost is $q = 7$, the value we previously determined as the optimal choice of $q$. Once we have the optimal value of $q$, we can simply determine the optimal $p$ by looking up the point on the demand curve corresponding to $q = 7$. As expected, we get $p = 6.5$.

The ice-cream example is a little special in that we assumed constant marginal cost and a linear demand. The bottom panel on Figure 5.12 considers a more general case.

One feature that is common to both panels is that the marginal revenue curve falls below the (inverse) demand curve. Recall that marginal revenue is defined as the revenue from selling one extra unit. For a price-taking firm, marginal revenue is simply equal to price. By contrast, a price maker needs to lower price in order to sell one extra unit. This implies a loss of margin in all units sold.
Marginal revenue is lower than price

DEMAND ELASTICITY

How high should a firm set price? It depends! — specifically, it depends on how sensitive demand is to price. We can see this in Equation (5.1), which shows that, for a given \( q \), the higher the slope (in absolute value) of the demand curve, \( |\Delta q/\Delta p| \), the lower the optimal margin \( p - c \). More generally, optimal pricing requires the knowl-
Suppose that world oil demand decreases by 1.3 million barrels a day when price increases from $50 to $60 dollars per barrel. Would you consider the demand for oil to be very sensitive or not very sensitive to changes in price? Consider a second example: The demand for sugar in Europe decreases by 1 million tones per day when average retail price increases from €.80 to €.90 per kilo. Would you consider the demand for sugar to be very sensitive or not very sensitive to changes in price? You’re reaction to these questions might be: I have no idea! There is a reason for this: most people have no idea whether a million barrels is a lot or not. More generally, it’s hard to say if a given slope value is a large number or a small number, unless you are very familiar with the industry in question and the units with which \( p \) and \( q \) are measured. To make matters even more, comparing the demand for sugar in Europe with the worldwide demand for oil in terms of price sensitiveness implies comparing apples and oranges, so to speak. Again, the problem is that, by measuring the slope of the demand curve, we are stuck with units: barrels, dollars, kilos, euros, and so on.

To address these concerns, economists frequently make use of the concept of **price elasticity of demand** (or simply elasticity), which is normally denoted by the Greek letter \( \varepsilon \). Elasticity is defined as the percent change in demand divided by the percent change in price:

\[
\varepsilon \approx \frac{\% \Delta \text{ quantity}}{\% \Delta \text{ price}}
\]

The percent change in \( q \) is given by \( \Delta q \) (the variation in \( q \)) divided by \( q \) (the value of \( q \)). Therefore,

\[
\varepsilon \approx \left( \frac{\Delta q}{q} \right) / \left( \frac{\Delta p}{p} \right)
\]

Finally, we can rewrite this as

\[
\varepsilon \approx \left( \frac{\Delta q}{\Delta p} \right) \left( \frac{p}{q} \right)
\]

Let us go back to the ice-cream example. From Table 5.4, we know that when \( p = 6.5 \), \( q = 7 \); whereas when \( p = 6 \) we get \( q = 8 \).
### TABLE 5.5
Examples of elasticity values

<table>
<thead>
<tr>
<th>Product and market</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norwegian salmon in Spain</td>
<td>-0.8</td>
</tr>
<tr>
<td>Norwegian salmon in Italy</td>
<td>-0.9</td>
</tr>
<tr>
<td>Coffee in the Netherlands</td>
<td>-0.2</td>
</tr>
<tr>
<td>Natural gas in Europe (short-run)</td>
<td>-0.2</td>
</tr>
<tr>
<td>Natural gas in Europe (long-run)</td>
<td>-1.5</td>
</tr>
<tr>
<td>US luxury cars in US</td>
<td>-1.9</td>
</tr>
<tr>
<td>Foreign luxury cars in US</td>
<td>-2.8</td>
</tr>
<tr>
<td>Basic cable TV in US</td>
<td>-4.1</td>
</tr>
<tr>
<td>Satellite TV in US</td>
<td>-5.4</td>
</tr>
<tr>
<td>Ocean shipping services (worldwide)</td>
<td>-4.4</td>
</tr>
</tbody>
</table>

Consider then a change in price from 6.5 to 6. In terms of the above notation, we get $\Delta p = 6 - 6.5$ and $\Delta q = 8 - 7$. We can therefore estimate the value of elasticity at $p = 6.5$ as

$$
\epsilon = \left( \frac{8 - 7}{6 - 6.5} \right) \frac{1 \times 6.5}{-0.5 \times 7} = \frac{-6.5}{3.5} \approx -1.86
$$

In fact, we know from p. 166 that $\Delta q / \Delta p = -2$. Therefore, we can also estimate the value of $\epsilon$ when $p = 6.5$ as

$$
\epsilon = -2 \left( \frac{6.5}{7} \right) \approx -1.86 \quad (5.2)
$$

Estimating the value of the price elasticity of demand effectively amounts to estimating the demand curve. As we will see in Section 6.2, this can be a tricky business. So, if someone gives you an estimate of $\epsilon$ you should accept it with a healthy dose of skepticism.

Table 5.5 lists estimates of $\epsilon$ for various products. Notice that all values are negative, reflecting the fact that (nearly always) demand curves are downward sloping.

One note before we conclude this subsection: There is nothing in economic theory that implies a specific functional form for market demand. Commonly used functional forms include: (a) linear
5.3. OUTPUT LEVEL AND PRICE

ELASTICITY GALORE

When we talk about elasticity *tout court* we mean the price elasticity of demand, that is, the concept defined on p. 174. However, there are other elasticity concepts that economists frequently consider. First, we have the *income elasticity of demand*, normally denoted by $\epsilon_y$ and defined as

$$
\epsilon_y \equiv \frac{\Delta q / q}{\Delta y / y}
$$
TABLE 5.6
Elasticity value and product characteristics
Notation: $\varepsilon$ = price elasticity, $\varepsilon_{ij}$ = cross-price elasticity, $\varepsilon_y$ = income elasticity

<table>
<thead>
<tr>
<th>condition</th>
<th>nature of demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\varepsilon</td>
</tr>
<tr>
<td>$</td>
<td>\varepsilon</td>
</tr>
<tr>
<td>$\varepsilon_{ij} &gt; 0$</td>
<td>goods $i$ and $j$ are substitutes</td>
</tr>
<tr>
<td>$\varepsilon_{ij} &lt; 0$</td>
<td>goods $i$ and $j$ are complements</td>
</tr>
<tr>
<td>$\varepsilon_{ij} = 0$</td>
<td>goods $i$ and $j$ are independent</td>
</tr>
<tr>
<td>$\varepsilon_y &lt; 0$</td>
<td>the good is inferior</td>
</tr>
<tr>
<td>$\varepsilon_y &gt; 0$</td>
<td>the good is normal</td>
</tr>
<tr>
<td>$0 &lt; \varepsilon_y &lt; 1$</td>
<td>the good is a necessity</td>
</tr>
<tr>
<td>$\varepsilon_y &gt; 1$</td>
<td>the good is a luxury</td>
</tr>
</tbody>
</table>

where $y$ is income level. It measures the sensitivity of quantity demanded with respect to a change in consumer income.

Second, we have the **cross-price elasticity of demand**, normally denoted by $\varepsilon_{ij}$, which is defined as

$$
\varepsilon_{ij} \equiv (\Delta q_i / q_i) / (\Delta p_j / p_j)
$$

where $i$ refers to good $i$ and $j$ refers to good $j$. It measures the sensitivity of quantity demanded of good $i$ with respect to a change in the price of good $j$.

Economists have a plethora of names to classify products according to the values of the demand elasticities. Table 5.6 lists the main entries into this jargon set. First, based on the value of the price elasticity of demand, we say that a demand is elastic if the absolute value of the demand elasticity is greater than 1. Conversely, if the absolute value of the demand elasticity is less than 1 we say the demand is inelastic. As Table 5.5 suggests, an example of an elastic demand curve is cable TV, whereas coffee provides an example of an inelastic demand. Note however that, since the value of elasticity may vary with price, it’s possible that a demand is elastic for high price levels and inelastic for low price levels.

Regarding the cross-price elasticity, here are some examples to illustrate each possibility: Coke and Pepsi are substitute products,
whereas cars and tires and complement products. It’s fun to think of examples of independent products. Firestone tires and Ben&Jerry ice cream comes to mind.

Finally, with respect to the income elasticity of demand, two of the terms in Table 5.6 should be familiar from Chapter 4: A good is called an inferior good if consumers buy less of it when income increases. This implies that $\Delta q < 0$ following $\Delta y > 0$, which in turn implies $\epsilon_y$ is negative. Whether a good is normal or an inferior depends on the particular consumer and their income level. For most people, spam would be an example of an inferior good. Also, just as a given demand may be elastic for some price levels and inelastic for other ones, so the income elasticity of demand varies with income level. $50$ bottles of wine are a normal good for me. I dream of the day when my income is so high that $50$ bottles of wine will be an inferior good for me. Finally, as the name suggests most goods are normal goods. For example, if your income increases you are likely to buy more trips to Hawaii, thus trips to Hawaii would be a normal good.

**ELASTICITY RULES**

Equipped with the concept of elasticity, we are now ready to revisit the problem of optimal pricing. Recall from Equation 5.1 that

$$\frac{p-c}{q} = \frac{1}{|\Delta q/\Delta p|}$$

Multiplying both sides by $q/p$ (which is legit if $p > 0$) we get

$$\frac{p-c}{p} = \frac{1}{|\Delta q/\Delta p|} \times \frac{p}{q}$$

Since $|\Delta q/\Delta p| \times \frac{p}{q} = -\Delta q/\Delta p \times \frac{p}{q} = -\epsilon$, the above may be rewritten as

$$m = \frac{1}{-\epsilon}$$

(5.3)

where

$$m \equiv \frac{p-c}{p}$$

(5.4)

denotes the firm’s margin. In words,
If a firm is pricing optimally then its margin equals the inverse of the price elasticity (in absolute value).

Figure 5.14 provides graphical intuition for the so-called elasticity rule given by (5.3). Consider a price change from $p'$ to $p''$, that is, a price change of $\Delta p$. Since the seller’s profit is given by $(p - c) q$, this price increase leads to a gain given by area $G$, which in turn is given by $\Delta p \times q''$. The price increase also leads to a decrease in sales, from $q'$ to $q''$. This corresponds to a loss of $p' \times \Delta q$, which corresponds to area $L$.

The condition that the gain from a price increase is greater than the loss corresponds to $G > L$, which in turn is equivalent to

$$q \Delta p > (p - c) (-\Delta q)$$

$$\frac{-q \Delta p}{p \Delta q} > \frac{p - c}{p}$$

$$m < \frac{1}{-\epsilon}$$

In other words, the seller should increase the price if $m < 1/(-\epsilon)$. By a similar argument, the seller should decrease price if $m > 1/(-\epsilon)$. It follows that if the seller is setting an optimal price then it must be that $m = 1/(-\epsilon)$, the elasticity rule.

Let us return to the ice-cream pricing example. We didn’t use the elasticity rule to find $p^*$, but nevertheless the elasticity rule holds.
5.3. OUTPUT LEVEL AND PRICE

At $p = p^*$. In fact, at $p = 6.5$, margin is given by

$$m = \frac{6.5 - 3}{6.5} = .5385$$

As shown in Equation (5.2), at $p = 6.5$ elasticity is given by $-2 \times 6.5/7$. It follows that

$$\frac{1}{-\epsilon} = \frac{1}{-2 \times 6.5/7} = .5385$$

Bingo! We get the same .5385 value!

There are multiple ways of expressing the elasticity rule. One that can be particularly useful corresponds to re-writing Equation (5.3) as follows

$$\frac{p - c}{p} = \frac{1}{-\epsilon}$$

$$p - c = \frac{p}{-\epsilon}$$

$$p (1 + 1/\epsilon) = c$$

$$p = \frac{c}{(1 + 1/\epsilon)}$$

$$p = \left(\frac{\epsilon}{1 + \epsilon}\right) c$$

This version of the elasticity rule is particularly helpful if, given the values of $\epsilon$ and $c$, we want to determine optimal $p$.

There are also multiple ways of expressing the difference between price and cost. Specifically, a common alternative to the use of margin $m$ is markup $k$, defined as

$$k \equiv \frac{p - c}{c}$$

The elasticity rule corresponding to the firm’s markup is given by

$$k = \frac{1}{-\epsilon - 1}$$

(Practice: derive this rule from Equation 5.3.)

As an illustration, consider the following example. A pharmaceutical company is launching a new drug which is protected by
5.3. OUTPUT LEVEL AND PRICE

It’s estimated that the price elasticity of demand is constant and given by \(-1.5\). The seller’s cost is $10 per 12-dose package (our unit of account). What’s the profit maximizing price? From (5.5) we get

\[
p = \frac{-1.5}{1 + (-1.5)} \times 10 = 3 \times 10 = 30
\]

What are values of margin, markup at optimal price? From (5.4) and (5.6) we get

\[
m = \frac{30 - 10}{30} = \frac{2}{3}
\]

\[
k = \frac{30 - 10}{10} = 2
\]

respectively. Finally, let us check that the elasticity rules hold. From the right-hand sides of (5.3) and (5.5) we get

\[
\frac{1}{-e} = \frac{1}{1.5} = \frac{2}{3}
\]

\[
\frac{1}{-(e - 1)} = \frac{1}{1.5 - 1} = 2
\]

respectively. As expected, we get the values of \(m\) and \(k\) previously computed.

ADDITIONAL NOTES ON ELASTICITY AND PRICE

What if \(-1 < e < 0\)? Applying the elasticity rule \(p = c / \left(1 + \frac{1}{e}\right)\), we get a negative value for \(p\), which does not make much sense (how does a firm maximize profit by setting a negative price?). Let’s think about the economics of the situation. If \(e\) is less than 1 in absolute value, then an increase in price leads to an increase in revenue: the decrease in quantity is very small compared to the increase in price. Since an increase in price implies a decrease in quantity (small as it may be), it also implies a decrease in cost (for cost is increasing in output). We thus conclude that, if \(e\) is less than 1 in absolute value, then an increase in price leads to an increase in revenue and a decrease in cost; it thus leads to an increase in profit. It follows that, whenever \(-1 < e < 0\), it is optimal for the seller to increase price. It can never be optimal for a monopoly seller to operate on a segment of its demand curve where \(|e| < 1\), that is, where demand is inelstic.
In practice, we do observe goods with inelastic demand (see Table 5.5 for examples). Does this imply a violation of the elasticity rule? Maybe yes, maybe not. Consider for example ConEdison, the electrical utility serving New York City. Consumer demand for electricity is inelastic, that is, $-1 < e < 0$; still, ConEdison does not increase price as the elasticity rule would suggest. The reason is not that ConEdison would not want to increase price, rather that regulation prevents it from doing so.

Consider now the case of milk sold in Manhattan. This is another example of a good with inelastic demand. Absent price regulation as in electricity, why isn’t the price of milk higher? This time the reason is that, while the market price elasticity is less than 1 (in absolute value), the demand elasticity faced by each seller is considerably greater than 1 (in absolute value). In fact, to the extent that milk is a relatively homogeneous product, a small price increase by a small producer would reduce its demand to zero.

Finally, if the market demand for milk is inelastic, one might ask why don’t milk sellers get together and jointly increase price (so that buyers have no alternative but to buy a similar quantity at a higher price). This time the reason is that such an agreement would most likely be illegal, as we will see in Section 8.2.
KEY CONCEPTS

production function  average product  marginal product

decomposing marginal returns  concavity  isoquant

marginal rate of technical substitution (MRTS)

perfect complements  Leontief production function

perfect substitutes  isocost  returns to scale

increasing returns to scale  natural monopoly

decreasing returns to scale  management complexity

constant returns to scale  productivity  labor productivity

total factor productivity  isoprofit  marginal revenue

price elasticity of demand  income elasticity of demand

cross-price elasticity of demand  margin  elasticity rule
CHAPTER 6

SUPPLY AND DEMAND

In Chapter 2 we mentioned that economics is not just a behavioral science but also a social science. The behavioral element was patent in Part II of this book, devoted to the study of economic decision making. Parts III and IV introduce a different dimension of economics analysis, namely the study of the interaction between economic agents. In particular, Part III focuses on markets, the economic institution par excellence. We begin in this chapter with the study of firm and market supply, beginning with looking at the firm’s cost function. Section 6.2 focuses on the other side of supply and demand, beginning with the important concept of consumer willingness to pay.

6.1. COST FUNCTION AND SUPPLY

Up until now, we have characterized a firm by its production function $F(K, L)$. A more frequent, alternative characterization is given by the firm’s cost function. The firm’s cost function, typically denoted by $C(q)$, shows the least total cost of inputs the firm needs to pay in order to produce output $q$; that is, the cost of producing $q$ assuming the firm does so efficiently. The cost function presupposes that the firm chooses its optimal input mix (see Chapter 5).
FROM PRODUCTION FUNCTION TO COST FUNCTION

In order to get a better idea of the relation between a firm’s production function (introduced in Chapter 3) and its cost function, it helps to consider the one-input case. The left panel of Figure 6.1 depicts a firm’s production function when there is one input $x$ and one output $y$. In Chapter 3 we stressed the fact that marginal product is typically declining, that is, the production function is concave. The top panel of Figure 6.1 considers a more general case, the case when, for low levels of $x$, the production function is convex.

The justification for the convex portion of the firm’s production function is that at low levels of inputs (e.g., labor) an increase in input levels allows for efficiencies (e.g., division of labor) to the point that output increases more than proportionately with respect to input. For example, suppose that you need to dig up two holes on the sidewalk of Washington Square in order to fix a pipe. (If you were at NYU anytime from 2014–2019 this would have been a very common sight and we’re glad we’re done with it—knock on wood.) If only one worker (say, Flavio) is assigned to the job, then he is going to take a long time to do it: each time he fills a bucket inside the hole, he needs to come out of the hole and empty the bucket in a large container. Suppose however that there are two workers, Flavio and Anna. Then one of the workers (say, Anna) can stay outside of the
hole and take care of the job of carrying the bucket from Flavio and empty it in the large container. In this way each of the two workers specializes in one task and their joint contribution is greater than what they would accomplish if each were working separately.

While a production function may be convex for low levels of input, eventually the law of decreasing marginal product likely kicks in and the production function becomes concave, as shown on the top panel of Figure 6.1. Specifically, for \( x > x^o \) we see that \( f(x) \) is a concave function.

Since there is only one input, the cost function is relatively easy to derive. (Actually, not necessarily easy, but certainly easier than in the multiple-input case.) Suppose the firm wants to produce \( q \) units of output. Since \( q = f(x) \), this requires \( x = f^{-1}(q) \) input units, where \( f^{-1} \) is the inverse function of \( f \). For example, if \( q = \sqrt{x} \), then \( x = q^2 \), that is, \( f(x) = \sqrt{x} \) and \( f^{-1}(q) = q^2 \).

Since the firm must pay \( p_x \) per input unit, it follows that the cost of output level \( q \) is given by \( C(q) = p_x x = p_x f^{-1}(q) \). This cost function is plotted on the bottom panel of Figure 6.1. The shape of the inverse function \( f^{-1} \) corresponds to “flipping” \( f \) about the 45° line (the dashed line). Since \( C \) equals \( f^{-1} \) times a scalar (\( p_x \)), the shape of \( C \) is the same as the shape of \( f^{-1} \). (In this particular case we assume \( p_x = 1 \) for simplicity. We can always change money units so that this is the case.)

Since \( C \) corresponds to flipping \( f \) about the main diagonal and \( f \) is convex for \( q < q^o \), it follows that \( C \) is concave for \( q < q^o \). Conversely, \( f \) is concave for \( q > q^o \), whereas \( C \) is convex for \( q > q^o \). In other words,

If a firm’s one-input production function is concave (resp. convex) then the corresponding cost function is convex (resp. concave).

In the multiple input case, the cost function is defined as the lowest cost the firm must incur in order to produce output \( q \) including all possible input combinations that lead to \( q \). Specifically, consider the case when there are two inputs, \( K \) and \( L \). Then we look for the lowest isocost curve that includes a point in the isoquant \( q^* \) (a specific value of \( q \)), as we did in Chapter 5. The cost level corresponding to such isocost is the value of \( C(q^*) \). If we do so for every possible value \( q \) then we have the cost function \( C(q) \).
The optimal input mix may vary as we vary the level of output $q$. For example, if you have a very small farm then it does not pay to own a tractor. If you have a very large farm, however, then the capital/labor mix is likely to tilt in the direction of capital intensity. In other words, when the firm’s production function includes multiple inputs the relation between the production function and the cost function is not as clear-cut as in the one-input case. However, the idea remains true: if the production function is concave, then the cost function is likely convex.

### MARGINAL COST AND AVERAGE COST

As mentioned earlier, the firm’s cost function, typically denoted by $C(q)$, shows the least total cost of inputs the firm needs to pay in order to produce output $q$; that is, the cost of producing $q$ assuming the firm does so efficiently. The cost function $C(q)$ leads to a series of related cost concepts:

- **fixed cost** ($FC$): the cost that does not depend on the output level.
- **variable cost** ($VC$) that cost which is not fixed.
- **total cost** ($TC$): the sum of fixed cost and variable cost.
- **average cost** ($AC$) (also known as “unit cost”): total cost divided by output level.
- **average fixed cost** ($AFC$): fixed cost divided by output level.
- **average variable cost** ($AVC$): variable cost divided by output level.
- **marginal cost** ($MC$): the cost of one additional unit. In other words, the total cost of producing $q + 1$ units minus the total cost of producing $q$ units of output.

Two notes before continuing. First, strictly speaking, the above definition of $MC$ correspond to the concept of incremental cost. The rigorous definition of marginal cost is the derivative of total cost with respect to the output level. Second, in order to distinguish total cost from variable or fixed cost, sometimes we denote total cost by $TC$. In other words, $TC$ is simply $C$. Similarly, in order to distinguish average total cost from average variable or average fixed cost, we
The shape of a firm’s cost function depends on the shape of the firm’s production function, which in turn depends on the technology that transforms inputs into outputs.

Neil Turner

Krispy Kreme doughnut factory.

sometimes denote average total cost by $ATC$. In other words, $ATC$ is simply $AC$.

The particular structure of a firm’s cost function depends on its technology. Here are some examples:

- **bagels**: modest fixed cost (space), relatively constant marginal cost (labor and materials)
- **electricity generation**: large fixed cost (plant), initially declining marginal cost (large plants are more efficient, and many plants have startup costs)
- **music CDs**: large fixed cost (recording), small marginal cost (production and distribution)

As an example, Table 6.1 presents a specific cost function. For each output level (first column) the fourth column shows the total cost. This can be split into fixed cost (second column), that is, the portion of the cost that does not depend on output level, and variable cost (third column).

Marginal cost, as mentioned earlier, is given by the cost of one additional unit. Since we do not have the value of total cost unit by unit, we estimate marginal cost by dividing incremental cost by incremental output level. For example, we estimate the marginal cost at $q = 10$ as $(39.2 - 30.0) / (10 - 0) = .92$. Strictly speaking, marginal cost is the derivative of cost with respect to output level, that is, it corresponds to an infinitesimal change in $q$. Also, as in other cases there is some arbitrariness in the choice of $C(q + 1) - C(q)$ or $C(q) - C(q - 1)$ as the difference we use in order to estimate marginal cost at $q$. 
6.1. COST FUNCTION AND SUPPLY

### TABLE 6.1
Cost function

<table>
<thead>
<tr>
<th>q</th>
<th>FC</th>
<th>VC</th>
<th>TC</th>
<th>MC</th>
<th>AFC</th>
<th>AVC</th>
<th>ATC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td>0.0</td>
<td>30.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>9.2</td>
<td>39.2</td>
<td>0.92</td>
<td>3.0</td>
<td>0.92</td>
<td>3.92</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>14.8</td>
<td>44.8</td>
<td>0.56</td>
<td>1.5</td>
<td>0.74</td>
<td>2.24</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>20.4</td>
<td>50.4</td>
<td>0.56</td>
<td>1.0</td>
<td>0.68</td>
<td>1.68</td>
</tr>
<tr>
<td>40</td>
<td>30</td>
<td>29.6</td>
<td>59.6</td>
<td>0.92</td>
<td>0.75</td>
<td>0.74</td>
<td>1.49</td>
</tr>
<tr>
<td>50</td>
<td>30</td>
<td>46.0</td>
<td>76.0</td>
<td>1.64</td>
<td>0.6</td>
<td>0.92</td>
<td>1.52</td>
</tr>
<tr>
<td>60</td>
<td>30</td>
<td>73.2</td>
<td>103.2</td>
<td>2.72</td>
<td>0.5</td>
<td>1.22</td>
<td>1.72</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>114.8</td>
<td>144.8</td>
<td>4.16</td>
<td>0.43</td>
<td>1.64</td>
<td>2.07</td>
</tr>
</tbody>
</table>

![Figure 6.2](image)

**FIGURE 6.2**
Total cost function

Average cost, as mentioned earlier, is given by the ratio between total cost and output level. For example, if \( q = 40 \) then average cost is given by \( 59.6/40 = 1.49 \). The value of average cost can be divided into average fixed cost and average variable cost. Continuing with \( q = 40 \), we have \( AVC = 29.6/40 = .74 \) and \( AFC = 30/40 = .75 \). Notice that \( AC = AFC + AVC \).

Assuming that the firm can choose any value of \( q \) (that is, not just multiples of 10 as in Table 6.1), Figures 6.2 and 6.3 depict the various cost function concepts. Figures 6.2 includes total cost and variable cost as a function of output level \( q \). Figures 6.3 includes marginal cost, average cost (divided into average fixed cost, average variable
6.1. COST FUNCTION AND SUPPLY

Focusing first on Figure 6.2, we see that if \( q = 50 \) then \( VC = 46 \), whereas \( TC = 76 \). If instead \( q = 0 \), then \( VC = 0 \) (by definition) and \( C = 30 \), knowing that \( C(0) = FC \). Focusing now on Figure 6.3, we observe an interesting property relating marginal and average cost curves:

- when \( MC \) is below \( AVC \) (\( q < q_1 \)), \( AVC \) is falling
- when \( MC \) is above \( AVC \) (\( q > q_1 \)), \( AVC \) is rising
- when \( MC \) is below \( ATC \) (\( q < q_2 \)), \( ATC \) is falling
- when \( MC \) is above \( ATC \) (\( q > q_2 \)), \( ATC \) is rising

As a result, the \( MC \) curve crosses the \( AVC \) and the \( ATC \) curves at their respective minima. (There is a close parallel between these relations and the relation between \( AP \) and \( MP \) discussed in Chapter 5.)

Frequently (and in Figure 6.3) the average cost function is U-shaped. This results from the tension between two “forces”, as it were. To the extent that the firm has positive fixed costs, the greater the output level, the lower the average cost. In other words, a higher output level allows the firm to spread the fixed cost over more units. This effect is responsible for the declining portion of the U (that is, the left part of the average cost function).

There is however another effect. As we saw in previous chapters, marginal product tends to decline beyond a certain input level (imagine adding a 14th worker to work on digging up a hole on the sidewalk of Washington Square). In term of the cost function, declining marginal product corresponds (broadly speaking) to an increasing marginal cost curve. If marginal cost increases at a fast rate, eventually it becomes higher than average cost, at which point average cost becomes increasing. This then is responsible for the increasing portion of the U of average cost.

FIRM SUPPLY

Knowing how much it costs to produce output \( q \), we now turn to the decision of which value \( q \) to choose. As an illustration, consider a very simple example, that of a small, price-taking t-shirt factory. By price-taking we mean that the output price (as well as the input prices) are taken by the firm as fixed (that is, as given). The idea is
6.1. COST FUNCTION AND SUPPLY

![Graph showing cost functions: min ATC, min AVC, ATC, AVC, AFC, MC.]

**FIGURE 6.3**
Derived cost functions

![Graph showing marginal cost (MC) and average cost (AC) vs. quantity (q).]

**FIGURE 6.4**
T-shirt factory marginal cost and average cost

that the firm is so small that its decision of how much output to sell or how much to purchase of each input does not have an impact on the respective market prices. (Similarly, as a consumer you are most likely a price taker: no matter how many lattes you drink at Starbucks (within reason) the price is not going to change. If you drank something like 13 million lattes then Starbucks might start thinking about charging a different price, but it’s unlikely you will buy or drink 13 million lattes.)

Suppose that, in order to produce t-shirts, a manager leases one
machine at the rate of $20 per week. The machine must be operated by one worker. The hourly wage paid to that worker is as follows: $1 during weekdays (up to 40 hours), $2 on Saturdays (up to 8 hours), and $3 on Sundays (up to 8 hours). (Lest you think this is an unfair wage, keep in mind this factory operates in Kabralstan, where the local dollar is worth $10 US.) Finally, suppose that the machine — which is operated by the worker — produces one t-shirt per hour. Assuming that current output \( q \) is 40 t-shirts per week, we have that:

- The fixed cost is given by the machine weekly lease. We thus have \( FC = $20 \).
- The variable cost is given by 40 t-shirts times one hour per t-shirt times $1 per hour, which equals $40.
- The average cost is \( \frac{20 + 40}{40} = $1.50 \).
- The marginal cost is $2. In fact, producing the 41st t-shirt in a given week would imply asking the worker to work on Saturday, which would be paid at the hourly rate $2; and producing a t-shirt requires one hour of work.

These cost values were computed for a particular output level. However, both average cost and marginal cost depend on the output level. By computing the values of marginal cost and average cost for each output level, we get the marginal cost and average cost functions (as before). Figure 6.4 depicts these functions for the particular case of the T-shirt factory. The more general case is given by Figure 6.5.

What is the use of all of these cost concepts? Suppose that Benetton, the sole buyer of t-shirts from our small factory, is offering a price of \( p = $1.80 \) per t-shirt. Moreover, Benetton is willing to buy at that price as many t-shirts (within reason) as the factory wants to sell (that is, the t-shirt factory is a price-taking firm). Given this offer, should the factory operate on Saturday?

At the current output of \( q = 40 \) t-shirts a week, average cost is given by $1.50 (see above). This means that, at \( p = $1.80 \), the factory is making money. It might seem that, for this reason, it is worth it to operate on Saturdays as well: “if you are making money at the current output level, produce more and you will make more money.” As it turns out, this is the wrong way to think about it. What is relevant for the decision of whether or not to operate on Saturdays is the
comparison between price and *marginal* cost, not the comparison between price and average cost. And since marginal cost of operating on Saturdays is $2, it is not worth pursuing that strategy given that the selling price is only $1.80.

In other words, although the factory is making money at output level $q = 40$ (because price is greater than average cost), profits would be lower if output were increased (because price is lower than marginal cost); the factory would lose money at the margin. (By “lose money at the margin” I mean lose money by producing an additional — a marginal — unit of output.)

Suppose now that Benetton (still the sole buyer) offers a price $p = $1.30 per t-shirt. No matter what the output level is, price falls below average cost. (Check this.) That is, no matter how much the factory produces, it will lose money. In fact, $p < AC$ implies that $p \times q < AC \times q$, that is, revenues $(p \times q)$ are less than total cost $(AC \times q = C)$. It follows that the optimal decision would be not to produce at all. (This comparison is based on the assumption that the firm has still not paid for the weekly machine lease; more later.) To summarize,

*marginal cost is the appropriate cost concept to decide how much to produce, whereas average cost is the appropriate cost concept to decide whether to produce at all.*
The t-shirt factory example is a bit special in that there are only two factors of production and there isn’t much flexibility in production. In general, the marginal cost and average cost functions would be continuous functions — or nearly continuous functions — as shown in Figure 6.5. In this figure, \( p^o \) denotes the minimum of the average cost function. For prices below this minimum, a price-taking firm would prefer not to produce at all. For values of \( p \) greater than \( p^o \), the optimal output level for a price-taking firm is given by the marginal cost function. For example, if \( p = p' \), then optimal output is \( q' \). More generally, a price-taking firm’s **supply function** is given by the marginal cost function for values of price greater than the minimum of average cost.

Optimal supply by a price-taking firm is a particular instance of a more general rule of optimal behavior: Suppose that — as we assume throughout — the firm wants to choose output level so as to maximize its profit (or, if you will, so as to maximize firm value). Profit is given by revenues minus costs. The marginal approach to optimal behavior — which appeared in all chapters from Chapter 2 to 5 — asks the following question: given the currently planned output level, what happens to firm profits if the firm decides to increase output by one unit?

We know that increasing output by one unit implies an increase in cost given by marginal cost \( MC \). We also know (from Chapter 5) that there is an equivalent concept for revenues: called marginal revenue. Specifically, **marginal revenue** is the added revenue the firm makes from selling one additional unit. In the present case — that is, the case of a price-taking firm — marginal revenue is simply given by price: each additional unit I sell I get the price I sell it for and that’s the additional revenue I get.

Since profit is given by revenues minus costs, increasing output by one unit changes profits by marginal revenue minus marginal cost. Or, considering the present case, by price minus marginal cost. If price is greater than marginal cost, then the firm is better off by increasing output by one unit: what it gets in extra revenue is more than what it pays in extra cost. If price is lower than marginal cost, then the firm is worse off by increasing output by one unit. (We can say more: the firm is actually better off by decreasing output by one unit.)

Similar to the marginal approach to price and output setting con-
6.1.COST FUNCTION AND SUPPLY

sidered in Section ??, the only point where the firm does not have a profitable adjustment is when price is exactly equal to marginal cost. At this point, neither a lower output nor a higher output would increase firm profits.

Figure 6.6 illustrates the above points. Optimal output is given by the point where price equals marginal cost. Specifically, assuming that current price is \( p^* \), optimal output is given by \( q = q^* \). Suppose that the firm were to set \( q = q_1 < q^* \) instead. Then the firm would be leaving money on the table: each of the units from \( q_1 \) to \( q^* \) costs less than \( p^* \) to produce (marginal cost is below \( p^* \)). Therefore, the firm is missing out on profitable sales. Suppose that the firm were to set \( q = q_2 > q^* \) instead. Then the firm would again be leaving money on the table: each of the units from \( q^* \) to \( q_2 \) costs more than \( p^* \) to produce (marginal cost is greater \( p^* \)). Therefore, the firm is engaging in unprofitable sales: reducing output from \( q_2 \) would lead to higher profit.

In the special case of the t-shirt factory, to the extent that marginal cost varies in a discontinuous way, we may end up in a solution where price falls between two values of marginal cost. Specifically, for \( q = 40 \), the marginal cost that is saved by reducing output by one unit is $1, whereas the marginal cost that needs to by paid to increase output to \( q = 41 \) is $2.
SHORT RUN AND LONG RUN

Continuing with the t-shirt factory example. Suppose that Benetton’s announcement that they are only willing to pay $1.30 arrived on Monday morning, after the factory manager had already committed to lease the machine for the week. What should the firm do?

To the extent that the fixed cost has already been paid, it is effectively a sunk cost: no matter what the firm does this week the $20 will need to be paid. As we saw in Section 2.3, sunk costs should be ignored in optimal decision making. The decision of whether or not to operate — and, if so, how much to produce — should be taken irrespective of the value of fixed cost.

In terms of cost functions, this means that, in the short run, the
relevant average cost concept is that of average variable cost. The top panel in Figure 6.7 shows the t-shirt factory’s average variable cost curve. As can be seen, its minimum is at $1. It follows that, if Benetton offers $1.30, then the firm should accept the offer and produce \( q = 40 \). We thus conclude that, in the short-run, the factory’s supply curve is given by the marginal cost curve for price values above $1 (the minimum of the average variable cost curve). This is shown in the top panel of Figure 6.7, which depicts the firm’s short-run supply curve.

In the long run, nothing is fixed. In particular, the firm must decide whether or not to renew the machine lease. In this context, assuming the $1.30 price remains valid, the firm is better off by simply shutting down. More generally, the bottom panel in Figure 6.7 shows the firm’s long-run supply curve.

A more general case is depicted in Figure 6.8. This case is more general in the sense that we do not have “jumps” as in the t-shirt factory example. It is still somewhat special in that we assume that the firm only has one technology available (and thus one possible cost structure). In this context, the only relevant difference between the short-run and the long-run is whether fixed costs have been paid or not. (In other words, more generally we could assume that, in the long run, the firm has the option of changing its technology, which, as we have seen before, corresponds to a production function \( f \), which in turn corresponds to a cost function \( C \).)

The short-run supply curve is given by the the portion of the supply curve lying above the minimum of the average variable cost (\( AVC \)). This corresponds to the red and the blue portions of the marginal cost curve. The long-run supply curve, in turn, corresponds to the portion of the supply curve lying above the minimum of average cost (\( AC \)). This corresponds to the blue portion of the marginal cost curve.

The above discussion also helps clarify the economics’ notion of short run and long run. The short run is defined as the period of time during which some factors of production are fixed. In the t-shirt factory example, the machine lease lasts for a week, and so the relevant period to define short run is one week. Naturally, in other industries the period corresponding to short run differs. In fact, in most industries it’s likely to be considerably longer than one week. The long run, by contrast, is the period of time required for all inputs
6.1. COST FUNCTION AND SUPPLY

For readers with a STEM background, the supply function is a difficult graph to read. Usually we think of functions as having the independent variable on the horizontal axis and the dependent variable on the vertical axis. For example, the function \( y = f(x) \) has \( x \) on the horizontal axis and \( y \) on the vertical axis.

The firm supply function is given by \( q = S(p) \), where \( p \) is on the vertical axis and \( q \) on the horizontal axis. This can be confusing, but hopefully you will get used to it. Moreover, while the supply curve gives \( q \) (horizontal axis) as a function of \( p \) (vertical axis), we can also read the curve the other way around: for a given \( q \) (horizontal axis), what is the value of \( p \) (vertical axis) on the supply curve? This corresponds to the inverse of the supply curve. We don’t usually refer to the supply curve in this way, but it may help to think about it for a minute.

The firm’s supply curve is determined by the equality \( p = MC \). Therefore, the inverse supply curve is nothing else but the marginal cost curve! In other words, \( S^{-1}(q) \) is nothing but \( MC(q) \). What is then the meaning of the inverse supply function? For a given value of \( q \), the inverse supply function (which is really the marginal cost
function) gives me the lowest price that the firm would require to sell that output unit. In fact, if the firm chooses $q$ optimally then I know that if $p$ is lower than $MC(q)$ then that particular $q$th unit will not be supplied, whereas, if $p$ is greater than $MC(q)$ then that particular $q$th unit will be supplied. In sum,

The supply curve may be read in two different ways. First, the direct way: for a given price, it gives the output level that the firm is willing to supply. Second, the inverse way: for a given output unit, it gives the lowest price such that the firm would want to supply that unit.

**SUPPLY CURVE AND SUPPLY FUNCTION**

One final word on terminology. A price-taking firm’s decision of how much to supply is determined, as we’ve seen, by the going price at which the firm can sell. However, there is a host of other factors that influence the firm’s decision. For example, if the cost of a particular input decreases then the firm’s marginal cost shifts downwards, which in turn results in a higher optimal output level for the same given price.

In this context, it is helpful to distinguish the supply function, which includes all factors influencing a firm’s decision, from the supply curve, where the single independent variable is price. Formally, the supply function is given by $S_f(p \mid x_1, x_2, \ldots)$, where $p$ is price and $x_1, x_2, \ldots$ represent input costs and other factors which influence a firm’s decisions. The supply curve, in turn, corresponds to $S(p) = S_f(p \mid \bar{x}_1, \bar{x}_2, \ldots)$, where $\bar{x}_1, \bar{x}_2, \ldots$ represent specific values of the variables $x_1, x_2, \ldots$ When representing the supply curve on a $(q, p)$ graph we implicitly assume specific values of $x_1, x_2, \ldots$ If the value of any of these variables changes, then the value of the supply function $S_f(p \mid x_1, x_2, \ldots)$ changes accordingly. In terms of the $(q, p)$ graph, we observe a shift in the supply curve $S(p)$.

**MARKET SUPPLY**

Having derived the firm’s supply curve, we now turn to the market supply curve. This is rather simple: the market supply is given by the (horizontal) sum of all firms’ supply curves. In other words, for
TABLE 6.2
From firm supply to market supply

<table>
<thead>
<tr>
<th>Price</th>
<th>Firm A</th>
<th>Firm B</th>
<th>Firm C</th>
<th>Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>6</td>
<td>3</td>
<td>20</td>
</tr>
</tbody>
</table>

TABLE 6.3
Individual firm marginal cost (and supply) functions

<table>
<thead>
<tr>
<th>Firm</th>
<th>$MC$</th>
<th>inverse $MC$</th>
<th>min $AC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{1}{2} + \frac{1}{2} q$</td>
<td>$1 + 2p$</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>$-1 + q$</td>
<td>$1 + p$</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>$2 + q$</td>
<td>$-2 + p$</td>
<td>5</td>
</tr>
</tbody>
</table>

each possible price $p$, we determine each firm’s supply $q_i(p)$; and the market supply is simply the sum of these values: $q_1(p) + q_2(p) + ...$

Consider the values in Table 6.2. The second to fourth columns give the value of supply by firms A, B and C. (Note that the assumption of price-taking firms is rather strong in a market with only three firms. We’re considering a small number of firms for simplicity sake. Normally, the markets where the price-taking assumption holds would include a considerably larger number of sellers.) For each price level (first column), we compute market supply by adding each firm’s supply. For example, if $p = 3$ then market supply is equal to $7 + 4 + 1 = 12$.

Alternatively, we may have estimated individual firm marginal cost curves (and thus supply functions). For example, consider the values in Table 6.3. Each row corresponds to a different firm. The first column indicates the firm’s marginal cost schedule. For example, Firm 1’s marginal cost is given by

$$MC(q) = -\frac{1}{2} + \frac{1}{2} q$$

We are also told that the minimum of the Firm 1’s average cost is given by 1. What is then Firm 1’s supply curve? Recall that the sup-
6.1. COST FUNCTION AND SUPPLY

The supply curve corresponds to the marginal cost curve but is read as $q$ as a function of $p$, whereas the $MC$ curve reads as cost as a function of $q$. So, in order to obtain the firm’s supply curve we need to set $p = MC$ and invert the marginal cost function:

$$p = -\frac{1}{2} + \frac{1}{2} q$$

or simply

$$q = 1 + 2p$$

This expression may be found on the third column of Table 6.3 and corresponds to that Firm 1’s supply curve (for values of $p$ greater than 1, the minimum of its average cost).

Figure 6.9 shows Firm 1’s supply curve (noted $q_1$). For values of $p$ lower than 1, Firm 1 supplies zero (better just shut down). For values of $p$ greater than 1, Firm 1 supplies $q = 1 + 2p$. By a similar process, we can also derive and graph the other firms’ individual supply curves:

$$q_1 = 1 + 2p$$
$$q_2 = 1 + p$$
$$q_3 = -2 + p$$

There is one added complication: each individual firm’s supply only kicks in when price is greater than the minimum of their average cost.
6.1. COST FUNCTION AND SUPPLY

MC is constant up to capacity. If \( p \leq MC \), supply equals zero. If \( p \geq MC \), then supply equals capacity.

FIGURE 6.10
Firm supply with capacity constraints

(which we can find on the fourth column of Table 6.3). So, for \( p \) less than 1 no firm supplies at all. For \( p \) greater than 1 but lower than 3, only Firm 1 supplies, in which case market supply \( Q \) is given by \( Q = q_1 \). For \( p \) greater than 3 but less than 5, both Firms 1 and 2 are active, in which case \( Q = q_1 + q_2 \). Finally, for \( p > 5 \), all firms are active and market supply is given by \( Q = q_1 + q_2 + q_3 \). Figure 6.9 shows market supply \( Q \) as a blue line with multiple segments. For \( p < 1 \), \( Q = 0 \). For \( 1 \leq p < 3 \), \( Q = q_1 = 1 + 2p \). For \( 3 \leq p < 5 \), \( Q = q_1 + q_2 = 2 + 3p \). Finally, for \( q \geq 5 \), \( Q = q_1 + q_2 + q_3 = 4p \).

A third possible case corresponds to firms that are subject to capacity constraints. As an illustration, consider the case of electricity generation in California. Table 6.4 lists the main generation plants as of 2000 (listed alphabetically). This includes a variety of types of plants: hydroelectric, nuclear, natural gas, etc. Different technologies have different cost structures. For example, hydroelectric plants have a very low variable cost (no fuel is required) but a high fixed cost.

As an approximation, we may assume that marginal cost is constant up to capacity, that is, marginal cost is equal to average variable cost. In graphic terms, this corresponds to a marginal cost function as in Figure 6.10. For any value of output lower than the firm’s capacity \( K \), marginal cost is constant and given by \( MC \). At \( q = K \), marginal cost shoots up to infinity, reflecting that it is impossible to produce
### 6.1. COST FUNCTION AND SUPPLY

#### TABLE 6.4
California electricity generation plants

<table>
<thead>
<tr>
<th>Unit name</th>
<th>Capacity (MW)</th>
<th>Fuel cost ($/MWH)</th>
<th>Var O&amp;M ($/MWH)</th>
<th>Total ($/MWH)</th>
<th>O&amp;M ($/Day)</th>
<th>Start cost ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALAMITOS 3-6</td>
<td>1900</td>
<td>48.00</td>
<td>1.50</td>
<td>49.50</td>
<td>20,000</td>
<td>34,000</td>
</tr>
<tr>
<td>ALAMITOS 7</td>
<td>250</td>
<td>83.00</td>
<td>1.50</td>
<td>84.50</td>
<td>0</td>
<td>8,000</td>
</tr>
<tr>
<td>BIG CREEK</td>
<td>1000</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>40,000</td>
<td>0</td>
</tr>
<tr>
<td>CONTRA COSTA 4&amp;5</td>
<td>150</td>
<td>58.00</td>
<td>0.50</td>
<td>58.50</td>
<td>8,000</td>
<td>16,000</td>
</tr>
<tr>
<td>CONTRA COSTA 6&amp;7</td>
<td>700</td>
<td>54.00</td>
<td>0.50</td>
<td>54.50</td>
<td>8,000</td>
<td>16,000</td>
</tr>
<tr>
<td>COOLWATER</td>
<td>650</td>
<td>58.00</td>
<td>0.50</td>
<td>58.50</td>
<td>4,000</td>
<td>12,000</td>
</tr>
<tr>
<td>DIABLO CANYON 1</td>
<td>1000</td>
<td>7.50</td>
<td>4.00</td>
<td>11.50</td>
<td>75,000</td>
<td>15,000</td>
</tr>
<tr>
<td>EL SEGUNDO 1&amp;2</td>
<td>400</td>
<td>60.00</td>
<td>1.50</td>
<td>61.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>EL SEGUNDO 3&amp;4</td>
<td>650</td>
<td>54.00</td>
<td>1.50</td>
<td>55.50</td>
<td>4,000</td>
<td>12,000</td>
</tr>
<tr>
<td>ELLWOOD</td>
<td>300</td>
<td>96.00</td>
<td>0.50</td>
<td>96.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENCINA</td>
<td>950</td>
<td>56.00</td>
<td>0.50</td>
<td>56.50</td>
<td>4,000</td>
<td>18,000</td>
</tr>
<tr>
<td>ETIWANDA 1-4</td>
<td>850</td>
<td>56.00</td>
<td>1.50</td>
<td>57.50</td>
<td>16,000</td>
<td>20,000</td>
</tr>
<tr>
<td>ETIWANDA 5</td>
<td>150</td>
<td>85.00</td>
<td>1.50</td>
<td>86.50</td>
<td>16,000</td>
<td>20,000</td>
</tr>
<tr>
<td>HELMS</td>
<td>800</td>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>40,000</td>
<td>0</td>
</tr>
<tr>
<td>HIGHGROVE</td>
<td>150</td>
<td>68.00</td>
<td>0.50</td>
<td>68.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HUMBOLDT</td>
<td>150</td>
<td>65.00</td>
<td>0.50</td>
<td>65.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HUNTERS POINT 1&amp;2</td>
<td>150</td>
<td>66.00</td>
<td>1.50</td>
<td>67.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>HUNTERS POINT 4</td>
<td>250</td>
<td>98.00</td>
<td>1.50</td>
<td>99.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>HUNTINGTON BEACH 1&amp;2</td>
<td>300</td>
<td>52.00</td>
<td>1.50</td>
<td>53.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>HUNTINGTON BEACH 5</td>
<td>150</td>
<td>75.00</td>
<td>1.50</td>
<td>76.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>KEARNY</td>
<td>200</td>
<td>105.00</td>
<td>0.50</td>
<td>105.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LONG BEACH</td>
<td>550</td>
<td>72.00</td>
<td>0.50</td>
<td>72.50</td>
<td>4,000</td>
<td>4,000</td>
</tr>
<tr>
<td>MANDALAY 1&amp;2</td>
<td>300</td>
<td>53.00</td>
<td>1.50</td>
<td>54.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>MANDALAY 3</td>
<td>150</td>
<td>75.00</td>
<td>1.50</td>
<td>76.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
<tr>
<td>MOHAVE 1</td>
<td>750</td>
<td>17.00</td>
<td>2.50</td>
<td>19.50</td>
<td>30,000</td>
<td>30,000</td>
</tr>
<tr>
<td>MOHAVE 2</td>
<td>750</td>
<td>17.00</td>
<td>2.50</td>
<td>19.50</td>
<td>30,000</td>
<td>30,000</td>
</tr>
<tr>
<td>MORRO BAY 1&amp;2</td>
<td>335</td>
<td>54.00</td>
<td>0.50</td>
<td>54.50</td>
<td>10,000</td>
<td>20,000</td>
</tr>
<tr>
<td>MORRO BAY 3&amp;4</td>
<td>665</td>
<td>50.00</td>
<td>0.50</td>
<td>50.50</td>
<td>10,000</td>
<td>20,000</td>
</tr>
<tr>
<td>MOSS LANDING 6</td>
<td>750</td>
<td>45.00</td>
<td>1.50</td>
<td>46.50</td>
<td>10,000</td>
<td>26,000</td>
</tr>
<tr>
<td>MOSS LANDING 7</td>
<td>750</td>
<td>45.00</td>
<td>1.50</td>
<td>46.50</td>
<td>10,000</td>
<td>26,000</td>
</tr>
<tr>
<td>NORTH ISLAND</td>
<td>150</td>
<td>85.00</td>
<td>0.50</td>
<td>85.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OAKLAND</td>
<td>150</td>
<td>74.00</td>
<td>0.50</td>
<td>74.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ORMOND BEACH 1</td>
<td>700</td>
<td>52.00</td>
<td>0.50</td>
<td>52.50</td>
<td>16,000</td>
<td>26,000</td>
</tr>
<tr>
<td>ORMOND BEACH 2</td>
<td>700</td>
<td>52.00</td>
<td>0.50</td>
<td>52.50</td>
<td>16,000</td>
<td>26,000</td>
</tr>
<tr>
<td>PITTSBURGH 1-4</td>
<td>650</td>
<td>56.00</td>
<td>0.50</td>
<td>56.50</td>
<td>10,000</td>
<td>38,000</td>
</tr>
<tr>
<td>PITTSBURGH 5&amp;6</td>
<td>650</td>
<td>50.00</td>
<td>0.50</td>
<td>50.50</td>
<td>10,000</td>
<td>38,000</td>
</tr>
<tr>
<td>PITTSBURGH 7</td>
<td>700</td>
<td>82.00</td>
<td>0.50</td>
<td>82.50</td>
<td>10,000</td>
<td>38,000</td>
</tr>
<tr>
<td>POTRERO HILL</td>
<td>150</td>
<td>85.00</td>
<td>0.50</td>
<td>85.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>REDONDO 5&amp;6</td>
<td>350</td>
<td>56.00</td>
<td>1.50</td>
<td>57.50</td>
<td>16,000</td>
<td>24,000</td>
</tr>
<tr>
<td>REDONDO 7&amp;8</td>
<td>950</td>
<td>52.00</td>
<td>1.50</td>
<td>53.50</td>
<td>16,000</td>
<td>24,000</td>
</tr>
<tr>
<td>SAN BERNADINO</td>
<td>100</td>
<td>72.00</td>
<td>0.50</td>
<td>72.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SOUTH BAY</td>
<td>700</td>
<td>60.00</td>
<td>0.50</td>
<td>60.50</td>
<td>2,000</td>
<td>8,000</td>
</tr>
</tbody>
</table>
more than capacity. In this context, the difference between the various plants is that they correspond to different electricity generation technologies, which in turn correspond to different values of $K$ and $MC$.

The supply function of firms with a cost structure as in Figure 6.10 is simple: if price is below $MC$, then supply zero; if price is greater than $MC$, then supply up to capacity. What about market supply? As mentioned earlier, market supply is given by the (horizontal) sum of all individual supply curves. In the present context, we proceed as follows:

- Order all plants by marginal cost level, from lowest to highest.
- For each price, determine the set of plants with marginal cost lower than that price.
- Add up the capacity levels of all such plants.
- The total value obtained in this way corresponds to the market supply at that price.
- Doing this for all price levels, we obtain the market supply function.

Figure 6.11 shows the result of this process. The first segment of the supply curve corresponds to hydroelectric plants such as Big Creek. These are plants with very low marginal costs (maybe even zero), which explains why the supply curve is close to the horizontal axis. The medium segment of the supply curve corresponds to nuclear power plants such as Diablo Canyon. There are several plants like Diablo, all with relatively similar marginal cost levels. This in turn leads to a relatively "flat" portion of market supply: for a relatively small change on price, many new plants come in line, thus producing a large increase in supply. Finally, the high price segment corresponds to fuel-burning plants such as Kearny. There is quite a variety of fuel-burning technologies (and fuel types). Moreover, these plants tend to have relatively small capacity. This results in a portion of the supply curve that is relatively steep, that is, it takes a considerable increase in price to induce some of these plants to come in line.
In the previous sections we examined the foundations of market supply. We now turn to the foundations of market demand. Specifically, the goal is to present the basic elements from which the market demand curve is derived. We also touch on the issue of estimating the demand curve based on historical data.

6.2. WILLINGNESS TO PAY AND DEMAND

In Chapters 3 and 4 we saw numerous examples where an agent (a consumer, a worker, a student) optimizes the trade-off between two goods. Consider now Margarida’s choice between consumption of a good $q$ and consumption of all other goods, which we denote by $\tilde{q}$. Let the price of $q$ be given by $p$. As to the price of $\tilde{q}$, assuming we measure $\tilde{q}$ as the amount spent in all goods other than $q$, its price is simply given by 1 (it costs exactly $1 to spend $1 in $\tilde{q}$). Finally, let income be given by $I$.

Figure 6.12 illustrates this problem. Suppose that initially the price of $q$ is given by $p_1$. Then the consumer’s budget line is given by the line marked $b_1$. This is a line with intercept $I$ on the $\tilde{q}$ axis and $I/p_1$ on the $q$ axis. Given Margarida’s preferences, we conclude that her optimal choice is given by $c_1$, which corresponds to consuming $q_1$ of $q$.

Suppose now that we cut $p$ by 50%. Since the intercept on the
6.2. WILLINGNESS TO PAY AND DEMAND

$q$ axis is given by $I/p$, we have a rotation in Margarida’s budget line around the vertical intercept $(0, I)$ and such that the horizontal intercept is twice as far from 0 as the initial one. In other words, $l/p_2 = 2l/p_1$. Given Margarida’s preferences, we conclude that her new optimal choice is given by $c_2$, which corresponds to consuming $q_2$ of $q$.

We now have two observations regarding Margarida’s behavior. When $p = p_1$, she chooses to purchase (and consume) $q_1$ units of $q$. When $p = p_2$, she chooses to purchase (and consume) $q_2$ units of $q$. We can plot these points in Figure 6.13, where we continue to plot $q$ on the horizontal axis but now we plot its price, $p$, on the vertical axis.

We could repeat this exercise (change the value of $p$, determine Margarida’s optimal choice $c$, transfer the corresponding value of $q$ from Figure 6.12 to Figure 6.13, together with the value of $p$) for many values of $p$. The result would be a relation between $p$ and $q$ like the line in Figure 6.13. This we call Margarida’s demand curve for $q$. Specifically, an individual’s demand curve for good $q$, denoted $D(p)$, gives their choice of $q$ when price is given by $p$.

I already mentioned this in Section 6.1 (in the context of the supply curve) and will do it again: The consumer demand curve is given by $q = D(p)$, where $p$ is on the vertical axis and $q$ on the horizontal axis. This can be confusing, so keep that in mind.
6.2. WILLINGNESS TO PAY AND DEMAND

In Chapters 3 and 4, we saw that points \( c_1 \) and \( c_2 \) in Figure 6.12 are derived from the \( \text{MRS} = \text{MRT} \) condition. The value of MRT reflects Margarida’s ability to trade-off consumption of \( q \) for consumption of \( \bar{q} \). It’s simply given by the price ratio. In the present context, since the price of \( q \) is (by normalization) equal to 1, the price ratio is simply the price of \( q \), that is, \( p \).

The value of MRS, in turn, reflects Margarida’s willingness to trade-off consumption of \( q \) for consumption of \( \bar{q} \). Generally speaking, MRS measures how much the individual is willing to give up of what’s on the vertical axis in order to obtain one additional unit of what’s on the horizontal axis. In the present case, we measure consumption of “not \( q \)” on the vertical axis (\( \bar{q} \)), and this we measure in $. Therefore, in the present context, MRS measures how much money (how many $) Margarida is willing to give for one additional unit of \( q \). For this reason, in the present context MRS corresponds to Margarida’s willingness to pay for \( q \).

Also in the present context, the optimality condition \( \text{MRS} = \text{MRT} \) comes down to \( \text{MRS} = p \). In other words, at Margarida’s optimal solution her willingness to pay for an extra unit of \( q \) (MRS) is exactly equal to the price of \( q \) (MRT). This makes sense: if Margarida’s willingness to pay for \( q \) were different from \( p \), then she could do better. Specifically, if MRS is greater than \( p \) then Margarida is better off by...
getting more of \( q \). Conversely, if MRS is lower than \( p \) then Margarida is better off by getting less of \( q \).

Let us now turn to Figure 6.13. As explained on page 207, for a given value of \( p \), an individual’s demand curve, denoted \( D(p) \), gives their choice of \( q \). But this function \( D(p) \) can also be read in the reverse: for a given value of \( q \), \( D^{-1}(q) \) gives a particular value of \( p \). What is the meaning of this value of \( p \)? As we saw in the previous paragraphs, Margarida’s optimal behavior implies that \( p \) equals her willingness to pay for \( q \), so \( D^{-1}(q) \) is Margarida’s willingness to pay for the \( q \)th unit. More generally, the values of \( D^{-1} \) correspond to Margarida’s inverse demand curve.

For each unit of an individual’s consumption plan, the inverse demand curve, denoted \( D^{-1}(q) \), gives their willingness to pay for such unit.

For example, suppose that Raj’s (daily) demand for coffee is given by \( q = 10 - 2p \). This means that, if \( p = 3 \), then Raj would be willing to purchase \( 10 - 2 \times 3 = 4 \) cups of coffee. From the direct demand curve (or simply “demand curve”), \( q = 10 - 2p \), we can derive Raj’s inverse demand by solving it with respect to \( p \). We get \( p = 5 - \frac{1}{2}q \).

We can now answer questions such as: How much is Raj willing to pay for a second cup of coffee each day? We get the answer by plugging \( q = 2 \) into the inverse demand: \( 5 - \frac{1}{2} \times 2 = 4 \). How much is Raj willing to pay for a fourth cup of coffee each day? Again, we can get the answer by plugging \( q = 4 \) into the inverse demand: \( 5 - \frac{1}{2} \times 4 = 3 \).

Note that, for \( p = 3 \), we could have saved ourselves the above computational work: We already knew that, at \( p = 3 \), Raj demands 4 cups of coffee. Therefore, it must be the case that Raj is willing to pay $3 for the 4th cup of coffee.

DEMAND CURVE AND DEMAND FUNCTION

One final word on terminology. A consumer’s decision of how much to buy is determined, as we’ve seen, by the good’s price. However, there is a host of other factors that influence the consumer’s decision. For example, if the good is a normal good and the consumer’s income
increases then the consumer purchases more of the good for the same given price.

In this context, it is helpful to distinguish the demand function, which includes all factors influencing a consumer’s decision, from the demand curve, where the single independent variable is price. Formally, the demand function is given by \( D_f(p \mid x_1, x_2, ...) \), where \( p \) is price and \( x_1, x_2, ... \) represent income, price of substitute products, and other factors that influence a consumer’s decisions. The demand curve, in turn, corresponds to \( D(p) = D_f(p \mid \bar{x}_1, \bar{x}_2, ...) \), where \( \bar{x}_1, \bar{x}_2, ... \) represent specific values of the variables \( x_1, x_2, ... \).

When representing the demand curve on a \((q, p)\) graph we implicitly assume specific values of \( x_1, x_2, ... \). If the value of any of these variables changes, then the value of the demand function \( D_f(p \mid x_1, x_2, ...) \) changes accordingly. In terms of the \((q, p)\) graph, we observe a shift in the demand curve \( D(p) \).

**MARKET DEMAND**

In Section 6.1 we saw how to get from individual firm supply curves to the market supply curve. Something similar happens with demand curves: we derive the market demand curve by adding up (horizontally) all individual consumers’ demand curves.

Market demand is downward sloping for two reasons: First, as price decreases, existing consumers demand more of the good. Second, as price decreases, new consumers “enter” the market and contribute to market demand. (There is some similarity here with respect to the reasons why the supply curve is increasing. See Chapter 6.)

As mentioned in Chapter 4, the law of demand establishes that the substitution effect is always negative, that is, the substitution effect of a price increase is negative (or, conversely, the substitution effect of a price decrease is positive). Sometimes people interpret the law of demand as meaning that the demand curve is downward sloping. However, that is not necessarily the case: if the income effect of a price change is sufficiently large then demand curves can be positively sloped (and the good in question is called a Giffen good).

The same ideas apply to the market demand curve: It is normally downward sloping, and frequently we (somewhat imprecisely) refer to this as the law of demand. In this sense, the “law of demand” is more an empirical regularity than a logical proposition.
Finally, note that factors which influence the demand of one or many consumers (e.g., income levels) also influence market demand.

As an illustration of how market demand can be obtained from individual demand curves, consider the values in Table 6.5. Each row corresponds to a different price, from 1 to 5. Columns 2 to 4 show the quantity demanded by three different individuals (A, B and C). Finally, Column 5 gives total market demand, the sum of all individual demands.

Normally, the market is made up of more than three consumers — more like three thousand or three million. In this context, working with individual demand curves for each of three million consumers is not practical. What we can do, however, is to start from individual demand for certain types of consumer, and then derive the market demand based on the number of consumers of each type. This is particularly helpful when we have information about population demographics. In the next subsection, I develop an example of such aggregation process.

**EXAMPLE: DEMAND FOR MINIVANS**

Suppose we want to estimate the demand for minivans at the state level, that is, for each of the 50 US states. Suppose the demand for minivans for a typical US household depends on whether the household includes children under 18 or not. Specifically, suppose it was estimated that the demand of a household without children \( q_0 \) and...
with children \((q_1)\) are given by

\[
q_0 = 0.8 - \frac{1}{50} p \\
q_1 = 1 - \frac{1}{80} p
\]

Figure 6.14 plots these demand curves. Note that for positive prices \(p\), the values of \(q_0\) and \(q_1\) are less than 1. It may seem odd for a household to purchase a fraction of a minivan. The correct interpretation is that the values of \(q_i\) \((i = 0, 1)\) measure the probability that a household of a given type purchases a minivan.

For a specific household, many factors come into play when deciding whether to purchase a minivan. It would be very difficult to predict the choice of any particular household. However, as we consider market demand, which corresponds to thousands of households, by virtue of the law of large numbers we will be able to predict the aggregate number of purchases reasonably well. For example, if a household of type \(a\) buys with probability 38% and there are 200 such households then I expect that, in the aggregate, \(0.38 \times 200 = 76\) minivans will be purchased.

Assuming (for simplicity) that there are only two types of buyer (households with children and households with no children) we are ready to derive the market demand for each state. Consider specifically the states of New Mexico and West Virginia. (For simplicity, we restrict the analysis to New Mexico and West Virginia. Naturally,
TABLE 6.6
2000 US Census: families and children, by state

<table>
<thead>
<tr>
<th>Concept</th>
<th>New Mexico</th>
<th>West Virginia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Households (thousand)</td>
<td>466.5</td>
<td>504.1</td>
</tr>
<tr>
<td>Households without children (thousand)</td>
<td>231.5</td>
<td>291.0</td>
</tr>
<tr>
<td>Households with children (thousand)</td>
<td>235.0</td>
<td>213.1</td>
</tr>
<tr>
<td>Households with children (%)</td>
<td>50.4</td>
<td>42.3</td>
</tr>
</tbody>
</table>

the same information can be obtained for all other states.) From the 2000 Census, we obtain the values in Table 6.6. In order to derive the total demand for minivans, we compute the product of the demand curve by households with no children times the number of households with no children and to this we add the product of the demand curve by households with children times the number of households with children.

One trick we must take into account is that, for high values of $p$, the demand by households with no children is equal to zero, in which case only the second product above applies. Specifically, by looking at Figure 6.14 we see that, if $p > 80$, then both types demand zero. If $p$ falls between 40 and 80, then type 1 has positive demand but type 0 has zero demand. Finally, if $p$ is less than 40 then both types have positive demand. Taking this into consideration, we derive the market demand in New Mexico as

$$Q_{NM} = \begin{cases} 0 & p > $80 \\ 235.0 \times (1 - p/80) & $40 < p \leq $80 \\ 231.5 \times (0.8 - p/50) + 235.0 \times (1 - p/80) & p \leq $40 \end{cases}$$

whereas for West Virginia we get

$$Q_{WV} = \begin{cases} 0 & p > $80 \\ 213.1 \times (1 - p/80) & $40 < p \leq $80 \\ 291.0 \times (0.8 - p/50) + 213.1 \times (1 - p/80) & p \leq $40 \end{cases}$$

Figure 6.15 plots the resulting market demand curves. For simplicity, we restrict to the 24-30,000 price range, where most minivans are priced. Notice that, overall, the demand in New Mexico is not very different than the demand in West Virginia. At first this may seem
surprising, considering that there are 504 thousand households in West Virginia and only 467 thousand in New Mexico. However, because the percentage of families with children is higher in New Mexico than in West Virginia, the totals end up being similar.

Another interesting observation is that, at a price of $28,000, total demand is the same in New Mexico and West Virginia. However, the market demand in New Mexico is steeper — that is, less sensitive to price changes — than the demand in West Virginia. This makes sense: to the extent that there are more households with children in New Mexico, we expect New Mexico households to be more “dependent” on minivans, and so their demand less sensitive to price changes.

DEMAND ESTIMATION

In the previous exercise (statewide demand for minivans) we started from given demand curves for a typical household (with or without children). Where do these demand curves come from? In an ideal world, firms would know the demands for their products. In practice, it’s not so easy. One reason is that it’s hard to get reliable market data: how much was bought by whom and at what price.

Another difficulty with demand estimation is that it’s inherently difficult to tease out the effect of price from the effect of other vari-
ables, especially when the latter might be changing at the same time as price (or, even worse, when they are not known to us).

Consider Figures 6.16 and 6.17. Both plot annual data of prices and quantities. Figure 6.16 corresponds to gasoline market in the US, whereas Figure 6.17 corresponds to NY Mets ticket sales. If one eyeballs the historical observations, a pattern emerges: years when prices were higher are also years when quantity sold was higher. One may naively suppose that the demand for gasoline and the demand for Mets tickets are upward sloping. This would be great news for the Mets, who could easily solve the problem of empty seats: simply raise prices until the park is full!

Alas, a historical positive relation between price and quantity does not mean that demand is positively sloped. (As mentioned in Section ??, correlation is not the same as causality.) What happens is that over time the demand curve has shifted (and, to some extent, the supply curve too, but primarily the demand curve). And this results in a series of points that look more like a supply curve than a demand curve, but is really neither one nor the other.

The point is that estimating the demand curve from historical data is a tricky business. To see this, consider Figure 6.18, which shows a series of demand curves and supply curves over a number of periods and a series of observations (data points). Suppose that the demand curve is fixed at \( D_3 \) (i.e., over time demand remains at
6.2. WILLINGNESS TO PAY AND DEMAND

$p (\$)\n\begin{array}{c}
10 & 15 & 20 & 25 & 30 & 35 & 40 \\
1998 & D & 2000 & q (\text{per game}) (000)
\end{array}$

**FIGURE 6.17**
Example: demand for NY Mets tickets

As the supply curve shifts between $S_1$ and $S_2$, we observe equilibrium points $A$ and $D$. Based on these, we are able to correctly identify the shape of the demand curve. Suppose instead that the supply curve is fixed at $S_1$. As the demand curve shifts between $D_3$ and $D_4$, we observe equilibrium points $A$ and $C$. Based on these, we are able to correctly identify the shape of the supply curve.

The problem is that, in practice, market data such as that found in Figures 6.16 and 6.17 result from a combination of demand-curve and supply-curve shifts; and if we are not careful we may end up failing to identify either. I have colleagues who provide this service for a reasonable fee, but you should be aware of the main challenges in estimating the demand curve from actual data, including the problem of both $S$ and $D$ shifting around. Econometricians refer to this as the **identification problem**.

While I don’t expect you to become an expert in the identification problem, it helps to look at a specific example: estimating the slope of the demand for gasoline. From August 23 to August 31, 2005, the states of Louisiana and Mississippi suffered a major natural disaster: Hurricane Katrina. New Orleans and other parts of the South experienced winds as strong as 174 mph. More than 1,800 human lives were lost. In addition to human loss, Katrina also implied considerable economic costs. At an estimated $125 billion in damage, Katrina is considered the costliest Atlantic hurricane ever.
large chunk of the US gasoline refining capacity was affected by the hurricane. In other words, Katrina implied a major supply shock to the US gasoline market.

Figure 6.19 shows two time series: US gasoline prices (left scale) and an index of supply disruption (right scale). Time $t$ represents the onset of Katrina: August 23, 2005. We notice an immediate spike in supply disruption. It corresponds to the many refineries in Louisiana that were hit by the hurricane. We also note a spike in price during the period of supply disruption. The idea of demand identification is to take advantage of this sudden shift in the supply curve to estimate the slope of the demand curve. The underlying assumption is that, while demand was affected by the hurricane (e.g., many drivers in Louisiana were unable to drive), the impact of Katrina on supply was considerably higher than the impact on demand. In terms of Figure 6.18, the events of Katrina corresponds to a shift in supply keeping demand constant, something like points $A$ and $D$. This allows us to estimate the slope of $D_3$.

Solving the supply-demand identification problem requires observing variables that shift one curve but not the other.

Once we’ve estimated the slope of the demand curve, we can go back to the left panel of Figure 6.16 and plot the various demand curves over the years. It’s not that a higher price implies a higher quantity...
6.2. WILLINGNESS TO PAY AND DEMAND

![Graph showing gasoline price and US gasoline supply disruption over time.](image)

**FIGURE 6.19**
Gasoline supply and prices around hurricane Katrina

demanded, as the correlation between \( p \) and \( q \) might suggest. Rather, what happens is that several factors (especially consumer income) shift the demand schedule over the years (as can be seen in Figure 6.16). Considering that the supply curve did not change as much as the demand curve did, the points where supply equals demand (the data values we keep a record of) show a positive correlation as seen in Figure 6.16.

Figure 6.17 (ticket sales at the NY Mets) shows a similar pattern. In this case, the shifts in the demand curve year after year are likely caused by team performance. Specifically, if the Mets perform well in year \( t \), then we observe a rightward shift in the demand for tickets (including season tickets) in year \( t + 1 \). Seeing how the team increases in popularity, the Mets management decides to increase ticket prices. In terms of observable data, we see an increase in price and an increase in number of ticket sales.

As we mentioned in Section ??, a correlation between two different variables (price and quantity, in this case) does not necessarily imply a causal relation. It’s not that an increase in \( p \) causes an increase in \( q \). Rather, a third variable (team popularity in the Mets case) causes a shift in demand and supply, which in turn leads to both an increase in \( p \) and an increase in \( q \).
Over the years, as the price of gasoline increased in the US, total sales of gasoline increased as well. This positive correlation does not imply that the demand for gasoline is positively sloped, rather that there were other factors (namely income) shifting the demand for gasoline to the right at the same time as price increased.

**ALTERNATIVE PATHS TO ESTIMATING DEMAND**

Statistical analysis is not the only avenue for demand estimation. An alternative approach is to obtain data by means of surveys. One problem with this method is that we are never sure how accurate the responses are going to be: until your own money is at stake you don’t have an incentive to think hard. For example, in Consumer Reports’ 2018 Automotive Fuel Economy Survey Report, we learn that

- 85% of Americans agree that “Ammakers should continue to improve fuel economy for all vehicle types.”
- 78% of Americans agree that “Making larger vehicles such as SUVs or trucks more fuel-efficient is important.”
- 74% of Americans agree that “Increasing average fuel economy from 25 miles per gallon (mpg) today to 40 MPG by 2025 is a worthwhile goal.”

However, according to the International Energy Agency average fuel efficiency of US cars is about 27.4 miles per gallon, whereas France, Germany and the UK show values of 44.4, 40.0 and 40.6, respectively. If fuel efficiency is so important to US drivers, why don’t they switch to more fuel efficient cars (which do exist in the US market)? In sum, one big problem with surveys is that consumers don’t always “walk the talk”.

Still another approach to estimating demand is to do experiments in markets. Thus catalog companies used to send out catalogs to different customers in which some of the prices are different. The idea is that knowing a customer’s ZIP code we know a lot about their
demographic characteristics (income, education, etc). If you sell at two different prices to different consumers in the same ZIP code, you are able to test the effect of changing price (assuming that the other determinants of consumer demand are the same for the consumers receiving different prices). With the advent of the Internet, these catalog strategies no longer work as well as they did before. Moreover, these experiments run the risk of alienating customers: What if you find out you got the high price?

Finally, if you do not have historical data to estimate demand or the resources to experiment with price changes, you should at least have an idea of whether demand is more or less sensitive to price changes depending on the characteristics of the good in question. In that spirit, here are some rules of thumb that might help in this process: First, demand for luxuries tends to be more price sensitive than demand for necessities. Compare, for example, food and Armani suits.

Second, demand for specific products (e.g., the iPhone) tends to be more price sensitive than demand for a category as a whole (smartphones). Why is this so? Because when the price of a specific product rises, people are willing to buy fewer units. Some of this reduction leads to purchases of other products in the same category (e.g., Samsung phones), some to a reduction in the category as a whole (smartphones). Only the latter induces price-sensitivity of the category as a whole.

Finally, demand is less sensitive to price changes in the short run than in the long run. A good example is gasoline demand. Can you see why? Suppose, for example, that the government plans to levy a 100% tax on gasoline for the next 3 years. In the next day or week consumers would probably still drive their cars, but in the longer term their demand for gas could change for many reasons: they might buy more fuel-efficient cars, carpool, or take the bus or train to work; perhaps some would work from home. As a result, the quantity of gasoline demanded at the new price would gradually decrease.

The following demand curves tend to be more sensitive to price changes: luxuries (vs necessities); specific products (vs category); long run (vs short run).
Consider the specific example of Facebook. Facebook users do not pay to use Facebook, so it’s impossible to use historical $p$ and $q$ data to estimate the demand curve. Is there a demand curve? As we saw earlier, we can “read” the demand curve as $q$ as a function of $p$ or, in inverse form, $p$ as a function of $q$. You may not know what your willingness to pay for Facebook is, but surely there is one. Sometimes, when I ask the question “How much would you be willing to pay to have a Facebook account” (or an Instagram account or a TikTok account), people tell me “I have no idea”. This is not true. You do have an idea. Is it more than one billion dollars? Probably not. So, we already know that your willingness to pay for a Facebook account is less than one billion dollars! You do have some idea.

Can we get a more precise estimate by asking Facebook users? In the previous subsection we learned the pitfalls of estimating valuations by asking people (how do you know they will tell the truth?). But suppose that, in addition to asking for a number, you also ask for a commitment on the part of the user, so that they have some skin in the game. Specifically, suppose that I ask Facebook users “How much would you be willing to receive in order to stop using Facebook?” Suppose moreover that it is understood that the answer is also a promise: After all of the answers are in, the experimenter will announce the “price” of turning off Facebook, and every user who announce a valuation lower than that “price” receives the “price” and turns off Facebook as promised.

This type of experiment has been performed a number of times. The idea is that, if users are rational economic agents, then their answer should be equal to their willingness to pay. Why? Suppose that my willingness to pay — or my best estimate thereof — is equal to $50. What if I say that I would only turn off Facebook for $100? Then if the announce price is $90 I will effectively be excluded from the offer, for 100 is greater than 90. I could have gotten $90 for something that would only cost me $50 (turning off Facebook). What if instead I say that I would turn Facebook off for $30? Then there is the risk that the price is $40, in which case I am paid $40 for something that will cost me $50 (turning off Facebook). In sum, my best bet is actually to be honest.

Figure 6.2 shows the result of two such surveys (source and sour-
6.2. WILLINGNESS TO PAY AND DEMAND

On the vertical axis we measure the offer proposed to each consumer (dollars paid if user accepts to deactivate her or his account for a period of four weeks). On the horizontal axis we measure the results, specifically the probability that the offer is rejected. Why do we plot probability of rejection? Because demanding Facebook corresponds to rejecting the offer to de-activate the account. In other words, to use Facebook is equivalent to not de-activating Facebook. It’s an awkward double negative but it’s an accurate one.

Based on the average responses, we can then estimate the inverse demand for Facebook. As expected, the demand is downward sloping: the more you pay users to quit Facebook, the fewer keep their Facebook account active. Note that the two curves correspond to different samples (one in the US and one in Europe) and follow slightly different experiments. Given that, it’s remarkable how similar the results are.
## KEY CONCEPTS

<table>
<thead>
<tr>
<th>cost function</th>
<th>cost function</th>
<th>fixed cost</th>
<th>variable cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>total cost</td>
<td>average cost</td>
<td>average fixed cost</td>
<td></td>
</tr>
<tr>
<td>average variable cost</td>
<td>marginal cost</td>
<td>supply function</td>
<td></td>
</tr>
<tr>
<td>marginal revenue</td>
<td>short-run supply curve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>long-run supply curve</td>
<td>short run</td>
<td>long run</td>
<td></td>
</tr>
<tr>
<td>supply function</td>
<td>supply curve</td>
<td>demand curve</td>
<td></td>
</tr>
<tr>
<td>willingness to pay</td>
<td>inverse demand curve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>demand function</td>
<td>demand curve</td>
<td>law of demand</td>
<td></td>
</tr>
<tr>
<td>law of large numbers</td>
<td>identification</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 7

EQUILIBRIUM AND EFFICIENCY

Having introduced the supply and demand framework (Chapter 6), this chapter puts it all to work. The goal is to understand how price and output level are determined in competitive markets (the law of supply and demand); how price and output levels are affected by exogenous events (e.g., a new tax, a natural disaster, a change in the price of an input or of a substitute product); and what properties such market equilibrium has. The chapter is divided into several parts: Section 7.1 introduces the concept of competitive markets and develops the framework of comparative statics (the analysis of how exogenous shocks affect price and output level). Section 7.2 defines the important concepts of producer and consumer surplus, and includes an important result in microeconomics, The First Welfare Theorem. Finally, Section 7.3 deals with public policy that interferes with the market mechanism (e.g., rent control).

7.1. COMPETITIVE MARKETS

In the previous chapters I have been referring to competitive markets, stressing the feature of many sellers on the supply side and many buyers on the demand side. At this point it will help to provide a more precise definition of what we mean by competitive markets. First, we assume that all firms produce the same product (as in commodity markets), that is, we make the homogeneous product
assumption. This implies that firms compete head to head on price, the only relevant variable. Moreover, firms are so small that no particular firm has an effect on price. In other words, firms are price takers.

Second, we assume well-defined property rights. At this point, this may not seem a very important point, but as we will see in Chapter 11 this turns out to be quite crucial. For now, suffice it to say that property rights over Planet Earth are not well defined: we all own the planet and nobody owns the planet, and this is one of the reasons why pollution and climate change are a big problem and an unregulated free-market solution won’t solve it. We’ll get back to this in Chapter 9.

Third, we assume that all agents (firms and consumers) are well aware of all prices and product characteristics and contract conditions. No hidden fees and so forth. Formally, we assume perfect information.

How reasonable is this? More so in some cases than in others. The characteristics of many commodity industries seem to match (at least approximately) the assumptions of the competitive model. More generally, the idea is to think of the perfect competition model as a benchmark that particular industries can be compared to.

That said, there are many industries where the above assumptions clearly do not hold. Internet search does not have a multitude of small-size, price-taking sellers: it’s Google and little else. Recent developments notwithstanding, the car industry — and we, car drivers — have not taken into account the enormous harm that gasoline engines inflict on the planet, a violation of the property rights of present and future generations. And no consumer that I know of has ever read the “terms of agreement” in any website; we simply click on “I agree”. It is therefore clear that we do not have perfect information regarding prices and product characteristics and contractual terms and conditions.

In Chapters 8, 9 and 10 we will deal with each of the above departures from the model of competitive markets. In this chapter we will stick with this reference point.
FIGURE 7.1
The market mechanism

THE MARKET MECHANISM

Figure 7.1 shows the basic diagram of supply and demand. We will repeat this figure numerous times throughout the rest of the semester, so it’s good that you get well acquainted with it. Recall that, as per Alfred Marshall, we measure price on the vertical axis and quantity (or output) on the horizontal axis.

As mentioned in Chapter 6, it helps to distinguish between the supply function and the supply curve; and between the demand function and the demand curve. In the supply and demand chart (Figure 7.1), price and quantity are treated as endogenous variables, that is, variables to be determined. In fact, the whole point of the model is to study the determinants of $p$ and $q$. All of the other variables included in the supply and demand functions are set at fixed values, that is, the particular supply and demand curves represented in Figure 7.1 correspond to particular values of those variables. If we change these values (as we will later) then we have a shift in either the supply curve or the demand curve or both.

The central idea of the model of supply and demand is that the price of a product is the result of the interaction between buyers (demand) and sellers (supply). We refer to the point at which the supply and demand curves intersect as the market equilibrium. In Figure 7.1 this is given by point $(q^*, p^*)$, where the supply curve crosses the demand curve. We thus say that $p^*$ is the equilibrium price and $q^*$ the
We say this outcome is an equilibrium in the sense that none of the market participants have an incentive to change their behavior: buyers are buying what they want at that price (the point is on the demand curve) and sellers are selling what they want (the point is on the supply curve).

If the price were higher than the equilibrium price, then fewer people would want to buy than to sell. We would be in a situation of excess supply (or surplus). This is illustrated in Figure 7.2, where price $p_1$ is higher than the equilibrium price $p^*$. If we were in this situation, then the excess of sellers would tend to drive the price down. The idea is that disgruntled sellers — that is, sellers who are unable to sell — would look for buyers and offer a lower price than the buyer is currently paying.

Conversely, if the price was lower than the equilibrium price, then fewer people would be willing to sell than to buy. We would be in a situation of excess demand (or shortage). This is illustrated in Figure 7.2, where price $p_2$ is lower than the equilibrium price $p^*$. If we were in this situation, then the excess of buyers would cause the price to increase. The idea is that disgruntled buyers — that is, buyers who are unable to buy — would look for sellers and offer a higher price than the seller is currently selling for.

The tendency of price to move in the direction of the equilibrium output level.

FIGURE 7.2
Excess supply and excess demand
7.1. COMPETITIVE MARKETS

When price deviates from the market clearing price, disgruntled buyers or disgruntled sellers make revised offers that push price in the direction of the equilibrium price.

Jeremy Kemp

Chicago Board of Trade corn pit (1993).

Price — also known as market-clearing price — is frequently referred to as the law of supply and demand. In other words,

*Price tends to move in the direction of the equilibrium price (where supply equals demand).*

The law of supply and demand is not a law in the sense common to natural and exact sciences. However, the analogy can be useful. Consider, for example, the system formed by a pendulum hanging from a ceiling. Such system has an equilibrium: the vertical position. This is the point where the downward gravitational force is exactly compensated by the upward pull exerted by the string (by Newton’s third law). Whenever the pendulum is not in the vertical position, a net force points in the direction of the rest point. (Not to worry, I will not be asking physics questions in the final.) By analogy, we can think of supply and demand as “forces” that push price up or down depending on whether the current price is below or above the equilibrium price. The equilibrium price, in turn, is the “rest point” of the system, the situation where the net force acting on price is zero.

As an example, consider the gold market. The top panel in Figure 7.3 shows the time series of gold price on a monthly frequency since 1970. Over the years, the price of gold has fluctuated considerably. This does not imply that price has been different from the equilibrium price, that is, the variations in price do not imply that we’ve been off equilibrium for decades. What happens is that, over time, both the supply of gold and the demand for gold are subject to shifts due to a variety of reasons, which in turn implies that the equilibrium values of p and q change.
The bottom panel of Figure 7.3 shows the daily price of gold during 2019. This time we see considerably less variation, which suggests that demand and supply have been relatively stable during the year 2019, i.e., $1,500 per ounce is the price such that market demand balances market supply. (Real time data — see for example Kitco — shows some within-day price variation, but not a very significant one.) Specifically, on October 25, 2019, the going price of gold was 1513.45 dollars per ounce. If for some reason many sellers had set an ask price of 1600, then there would have been too few buyers willing to buy at that price (excess supply). It would then have been in some sellers’ interest to reduce the ask. This is the equivalent of the
gravitational force in the pendulum example: if price is above the equilibrium price, then the “natural” market forces tend to bring it down to the market-clearing level. The same applies to situations when price is lower than its equilibrium level.

As the gold example suggests, in the real world there are many exogenous factors (production technology, input costs, tastes, income, and so forth) shifting the supply and demand curves at all times. For this reason, the equilibrium point itself is constantly changing. In this sense, it is not very appropriate to talk about equilibrium as a “rest point.” Nevertheless, finding the equilibrium is a helpful way of understanding in which direction we expect price to change in a given situation. We next turn to this issue.
COMPARATIVE STATICS

Often we are concerned about changes in market conditions. If the demand or supply curve shifts due to changes in various underlying factors, then a new equilibrium price is established. The term **comparative statics** is used by economists to describe the exercise of looking at what happens to the market equilibrium if an exogenous factor changes. We do this by shifting the supply or demand curves and noting the resulting change in the equilibrium price and quantity. The most basic principles of comparative statics are that

- (a) A rightward shift of the demand curve leads to an increase in quantity and an increase in price.
- (b) A rightward shift of the supply curve leads to an increase in quantity and a decrease in price.

(For the opposite shifts in demand or supply, just change the signs.) So, if I ask you what effect will event Y have on price and quantity in market X, the question to ask yourself is what Y implies in terms of shifts in the demand and supply curves of good X. For example, suppose that the price of raw materials falls. As illustrated in the top panel of Figure 7.4, this implies that $S$ shifts from $S_1$ to $S_2$ (a downward shift in the supply curve). If price were to stay at its original level $p_1$ then we would have a situation of excess supply. Therefore, price adjusts to the new and lower equilibrium level $p_2$. Suppose instead that consumer income increases. As illustrated in the bottom panel of Figure 7.4, this corresponds to a rightward shift in $D$, from $D_1$ to $D_2$. If price were to stay at its original level $p_1$ we would have a situation of excess demand. Therefore, price adjusts to the new and higher equilibrium level $p_2$.

**EXAMPLE: TAIWAN’S 1999 EARTHQUAKE**

Consider a real-world example: the September 1999 earthquake in Taiwan. What impact would you expect it to have on the world market for DRAM (dynamic random access memory), both in terms of price and in terms of quantity? Before answering that question, two important pieces of information: First, the relevant market for
7.1. COMPETITIVE MARKETS

DRAM is the world. Second, Taiwan is one of the world’s leading producers, accounting for about 10% of world output.

An earthquake is likely to shut down a series of factories (it did), which implies a northwest (or leftward) shift in the supply curve, like the shift from $S_2$ to $S_1$ in the top panel of Figure 7.4. The shift in supply leads to an increase in price and a decrease in quantity.

Figure 7.5 shows the world price of DRAM around the date of the Taiwan earthquake. Just before the earthquake, prices were around $7. Just after the earthquake, in a matter of days, prices sky-rocketed to values as high as $25. This seems a little too much, considering that only a fraction of Taiwan’s factories were affected, and Taiwan in turn represents less than 10% of world supply. One possible interpretation is that the earthquake also induced a sharp increase in speculative demand (brokers or computer manufacturers who stock-piled in anticipation of further price increases). This interpretation is consistent with the fact that prices declined considerably in the months after the initial spike.

SOME RULES OF THUMB

Before continuing, here are some rules of thumb that may help in the process of comparative statics. First, it’s important to distinguish be-
between shifts in the supply (or demand) curve and movements along the supply (or demand) curve. Consider the top panel in Figure 7.6. For illustration purposes, think of it as reflecting the supply of copper. Three periods are considered, from 1 to 3. At time 1, the supply curve is given by $S_1$ and price is given by $p_1$. It follows that the quantity supplied is given by $q_1$. Now suppose that, at time 2, an increase in demand by Chinese manufacturers pushes price up to $p_2$. The supply function is still the same, so we have a movement along the supply curve, leading to a higher quantity supplied, $q_2$. Finally, at time 3 a new copper mine is discovered. This shifts the entire supply schedule from $S_1 = S_2$ to $S_3$. China continues to push demand outward, so that the new equilibrium point is at $E_3$. The result is that, even though price did not change, we have an increase in quantity supplied, this time caused by a shift in the supply curve. (It is a bit of a fluke that the shift in demand and supply curves leads to a change in $q$ and a zero change in $p$. I consider this case for illustration purposes only.)

In sum, the change from $E_1$ to $E_2$ corresponds to a movement along the supply curve, whereas the change from $E_2$ to $E_3$ corresponds to a shift of the supply curve. The rule is simple: if it’s the price that changes, we have a movement along the curve; if it’s something else that changes, we have a shift of the curve itself.

The same applies to the demand curve. This is illustrated in the bottom panel of Figure 7.6. For illustration purposes, think of $D$ as the demand for Coke. Initially, demand is at $D_1$ and price is given by $p_1$, resulting in a quantity demanded equal to $q_1$. Suppose that, at time 2, the Coca-Cola Company decides to jack up price to $p_2$. Assuming that demand remains the same as at time 1, we have a movement along the demand curve, bringing quantity demanded down to $q_2$. Finally, suppose that, at time 3, PepsiCo decides to decrease the price of Pepsi. I know this is not your case, but there are consumers who do not have very strong feelings regarding Pepsi or Coke. Given that, the demand for Coke shifts to the left to $D_3$, the result of a big consumer shift from Coke to Pepsi. Now, even if the price of Coke remains the same ($p_3 = p_2$) we will observe a decrease in the quantity demanded of Coke, a decrease that results from the leftward shift in the demand for Coke.
 Movements along the supply (resp. demand) curve of $x$, while other factors remain the same, corresponds to changes in the price of $x$. Shifts in the supply (resp. demand) curve of $x$ result from changes in factors other than the price of $x$.

Figure 7.7 illustrates another important rule of thumb. Let us start with the top panel. When we say that there is an increase in supply, or that the supply function expands, we mean that the supply curve shifts in the south east direction, for example from $S_1$ to $S_2$. This shift in the supply function can be thought of as a movement to
7.1. COMPETITIVE MARKETS

Shifts in supply (top) and shifts in demand (bottom)

the right or a downward movement. I know, this seems confusing: a downward shift of the supply function corresponds to an increase in supply! Hopefully the confusion disappears if we pay closer attention to what is actually happening. Let us first consider a rightward movement. This is easier to interpret. For example, a new firm enters the market or an existing firm builds a new factory or goes on a hiring spree. All of these changes imply that, for a given price, the market is willing to supply a higher total quantity. This makes a lot of sense.

Consider now the case when one of the firm’s inputs becomes cheaper. For example, due to technological progress the cost of robots declines, or the government lowers import tariffs on steel (one of the
firm’s inputs). This implies that the lowest price that each supplier requires to supply a given output unit is now lower. Recall that the firm’s supply curve (and thus the market supply curve) reflects the firm’s marginal cost. If marginal cost is lower, then the height of the firm’s supply curve is lower. This in turn corresponds to an expansion in supply. In fact, if there is a reduction in the lowest price that a firm requires to supply, it follows that, for a given price, the firm is willing to supply more (check $S_2$ vs $S_1$).

The same is true for shifts in the demand curve, as illustrated in the bottom panel of Figure 7.7. Consider, for example, the demand for movie tickets at the only movie theater in Hill Valley, CA. If a bunch of people move to Hill Valley, then we have a shift of the demand curve to the right: for a given theater ticket price, more people will come to the movies. Alternatively, suppose that the theater owner invests on a major uplift of the theater facilities and starts exhibiting really popular movies. Then, even if the number of residents of Hill Valley does not change, their willingness to pay for a theater ticket increases. This corresponds to an upward shift of the demand curve. Similarly to a rightward shift, this upward shift corresponds to an expansion of the demand curve. In sum, an expansion of the demand curve may be interpreted as an increase (rightward shift) of the (direct) demand curve; or as an increase (upward shift) of the inverse demand (i.e., willingness to pay).

Once you get into the mechanics of comparative statics, it should come naturally to interpret real-world events as shocks to demand and supply curves, which in turn lead to adjustments in price and transaction volumes. Consider some recent events in the honey industry, as described in Box 7.1 (source: Cyril Morong). The left column includes a series of quotes from a Wall Street Journal article on the honey industry. The right column, in turn, includes a series of comments on how to interpret these events in terms of the model of supply and demand.

**COMPARATIVE STATICS: THE ANALYTICAL APPROACH**

Sometimes we are able to estimate the shape of the supply and the demand curves. In these cases, we may be able to go beyond the qualitative approach considered up to now and actually put a number on the prediction of what happens to price and quantity. Con-
How a news article on the honey industry reflects many of the concepts introduced in this chapter.

<table>
<thead>
<tr>
<th>Article quote</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honey prices are starting to sting. Global honey prices are at their highest levels in years, due to a new wave of consumer demand for natural sweeteners ...</td>
<td>Demand increases because tastes or preferences increased with the opposite happening for sugar.</td>
</tr>
<tr>
<td>... and declining bee populations that are hampering mass production.</td>
<td>Supply decreases.</td>
</tr>
<tr>
<td>In addition, it is being used more as an ingredient in shampoos, moisturizers and other personal-care products that companies market as naturally made.</td>
<td>Another increase in demand due to tastes.</td>
</tr>
<tr>
<td>U.S. retail prices averaged $7.66 a pound in May, up 9% from a year earlier. Those prices have risen by about two-thirds in the last decade. Americans consumed 596 million pounds of honey in 2017, or an average of nearly two pounds per person — up 65% since 2009.</td>
<td>If demand shifts right, we expect both price and quantity to increase</td>
</tr>
<tr>
<td>It has been touted by celebrities — including tennis star Novak Djokovic — for its health benefits and numerous scientific studies have shown it can help heal wounds, ulcers and burns.</td>
<td>Maybe this is part of the reason tastes increased.</td>
</tr>
<tr>
<td>Global honey production has been relatively stable over the past five years.</td>
<td>But if supply shifted left that could cancel out the demand increase and leave quantity the same.</td>
</tr>
<tr>
<td>In the U.S., honey production peaked in 2014 and has fallen 15% since then.</td>
<td>If supply shifted more to the left than demand shifted to the right, total Q falls — maybe the increased American quantity means less for consumers elsewhere.</td>
</tr>
</tbody>
</table>

Consider the following example from the market for gasoline. Suppose supply and demand functions are given by \((p \text{ in dollars}, q \text{ in million})\)
Once you get into the mechanics of comparative statics, it should come naturally to interpret real-world events as shocks to demand and supply curves which lead to adjustments in price and transaction volumes.

\[ Q_D = 150 - 50p \]
\[ Q_S = 60 + 40p \]

The initial market equilibrium results from the equality of supply and demand:

\[ 60 + 40p = 150 - 50p \]
\[ p = (150 - 60)/(40 + 50) = 1 \]
\[ q = 60 + 40 = 100 \]

So, the initial equilibrium price is $1 per gallon and a total quantity of 100 million gallons is sold.

Now suppose the government imposes a tax of \( t = 56.25 \) cents per gallon. Specifically, suppliers now receive the sale price but must pay 56.25 cents per gallon sold to the tax authority. It follows they get a net price of \( p \) minus 56.25 cents. This implies that the new supply curve is given by

\[ Q_S = 60 + 40(p - .5625) \]

The idea is that firm supply is determined by the net price the firm receives. Even though consumers pay \( p \) at the pump, each gas station must pay 56.25 cents to the tax authority, thus netting \( p - .5625 \). So, instead of \( Q_S = 60 + 40p \), the original supply curve, we get a new one where net price \( p - .5625 \) substitutes for \( p \). The new market
7.1. COMPETITIVE MARKETS

The equilibrium (with a 56.25 cents/gallon tax) is given by

\[ 60 + 40(p - 0.5625) = 150 - 50p \]
\[ p = \frac{(150 - 60 + 22.5)}{(40 + 50)} = 1.25 \]
\[ q = 60 + 40(p - t) = 60 + 40(1.25 - 0.5625) = 87.5 \]

Figure 7.8 illustrates the above calculations. We conclude that, as a result of the tax, producers get less, consumers pay more, and the government reaps \(0.5625 \times 87.5 = \$49.21875\) billion.

COMPARATIVE STATICS: THE NUMERICAL APPROACH

On September 14, 2019, drone attacks sparked fires at two Saudi Aramco oil facilities in Abqaiq (about 37 miles southwest of Dhahran). As a result, Saudi Arabia’s Interior Ministry announced that the government would shut-in half of its national oil production. Roughly 5 million barrels of oil per day, or 5% of global crude production, were suddenly taken off the market.

(Exercise “Oil supply function”, included in the companion Exercise Book, considers a simplified depiction of the world oil market. Assuming that each country has constant marginal cost up to capacity and that each country is a price taker, we derive the supply curve marked \(S_1\) on the top panel in Figure 7.9.)
As a result of the September 2019 attack, suppose that Saudi Arabia stops producing oil altogether. What impact would we expect that to have on the oil price? One way to answer the question is to compute the new market supply curve. Clearly, it’s shifting to the left, but by how much? If the oil price is less than Saudi Arabia’s cost, then quantity supplied is not affected. If the oil price is greater than Saudi Arabia’s cost, then quantity supplied decreases by the amount of Saudi Arabia’s capacity. The top panel in Figure 7.9 depicts the new market supply of oil, $S_2$, following this reasoning. Assuming that demand is given by $D$, then the market equilibrium moves from $E_1$ to $E_2$. The bottom panel in Figure 7.9 shows the actual daily spot price during the month of September 2019. As expected, price did in-
In October 2019, Saudi Arabia oil output was halved as some of its facilities were attacked by drones.

crease after the September 14 attack. (Prices are only quoted during weekdays. Since the attack took place on a Saturday, the first price observed after the shut-down was that for Monday, September 16.)

**EFFECT ON PRICE AND EFFECT ON VOLUME**

By now we know how shifts in supply and demand lead to variations in quantity and price. An additional question is whether the shifts in supply and demand lead primarily to movements in price or primarily to movements in quantity. In other words, going beyond the sign of the changes in $p$ and $q$, we are now also interested in the size of such changes.

If you work through an example using the supply and demand diagram, you’ll see that the impact on price or quantity depends on the slopes of the supply and demand curves. Specifically, the effect of a shift in the supply curve depends on the slope of the demand curve. (This sounds a little strange, but it’s true because the demand curve hasn’t shifted, so the change in equilibrium is a movement along the demand curve.) If the demand curve is steep, then price will change more than quantity in reaction to a shift in supply. By contrast, if the demand curve is flat then the impact of a shift in supply will be felt predominantly on quantity. For shifts in demand, the impact depends on the slope of the supply curve. The four panels in Figure 7.10 illustrate the four possible cases (demand or supply shifts with elastic or inelastic supply and demand, respectively). Can you think of examples that fit each of these cases?
7.1. COMPETITIVE MARKETS

The more sensitive demand (resp. supply) is to changes in price, the more the effect of a supply (resp. demand) shift will be felt on output as opposed to price.

Thus a key ingredient to any market analysis is an assessment of the slopes of the supply and demand curves: how sensitive the decisions of buyers and sellers are to changes in price. Consider the California electricity market. The capacity of local power plants can’t be changed much without building new ones. Moreover, the high-voltage lines to bring in power from other states have limited capacity. Hence the supply curve is very steep (vertical?) and the impact of an increase in demand (the result of growth of the California economy) is reflected almost entirely in the price. In fact, during 2000 and 2001, a combination of increasing demand and limited supply sent wholesale prices up by a factor of 8 or 9!
To conclude the discussion on price effects vs volume (or quantity) effects, let us go back to the case of gold. Figure 7.11 shows the daily values of gold price and trade volume. One particularly notable feature of the data is that the volume of trade is considerably more volatile than price. Why? One possible explanation in terms of the supply and demand framework is that both the supply and demand curves are very sensitive to price (very “flat”), so that small shocks to either schedule lead to major movements in volume of trade but not in price.

Figure 7.12 illustrates this possibility. Notice that the vertical price scale varies from 1400 to 1600. Therefore, the supply and demand schedules are considerably “flatter” than they look at first. Suppose there is a shift in the demand schedule from $D_1$ to $D_2$. This is a relatively small shift. In terms of vertical distance, we are talking about a drop in $40$ per ounce, a relatively small fraction of the $1500$ initial equilibrium price. However, to the extent that the supply curve is very sensitive to price changes, the shift in the demand curve produces a dramatic movement in the equilibrium value of $q$. Specifically, at $E_2$ the volume of trade is given by $18$ million troy ounces, a considerable drop from the initial $30$ million.
7.2. GAINS FROM TRADE AND EFFICIENCY

In this section we define the concepts of producer and consumer surplus. We also discuss how they can be estimated and how they can help understand the concept of value, in particular value in use. We then come to one of the central results in microeconomics, namely that in competitive markets efficiency is maximized at equilibrium.

**PRODUCER SURPLUS**

In Chapter 6 we saw that price-taking firms optimally set an output level such that price equals marginal cost. Suppose that this were not the case. Specifically, suppose that \( p > MC \), that is, price is greater than marginal cost. If that were the case, then by selling an extra unit the firm would receive an extra revenue of \( p \) and pay an extra cost \( MC \), which implies the firm would make an extra profit of \( p - MC \) on that unit. If the firm can increase its profit with respect to the current output level, then it is not choosing an optimal output level. It follows that \( p > MC \) cannot be an optimal solution. A similar argument applies to \( p < MC \), which in turn implies that it must be \( p = MC \) at the optimal output level.

The fact that the firm sells up to the point when \( p = MC \) does not mean that the firm is selling all of its output at cost. Specifically, suppose that \( q = q^* \) when \( p = p^* \), where \( p^* \) is the going market
price. To the extent that marginal cost is increasing in output level, each unit that the firm sells up to \( q = q^* \) is sold at a price greater than marginal cost. This is illustrated in the top panel of Figure 7.13, where we see that \( MC \) is below \( p \) for all values of \( q \) lower than \( q^* \). (This need not be the case always, that is, it is theoretically possible that the first units sold by the firm have a marginal cost greater than sale price.) Specifically, the first unit is sold at a (variable) profit \( p^* - MC(1) \), where \( MC(1) \) is the marginal cost of the first unit. The second unit is sold at a (variable) profit \( p^* - MC(2) \), where \( MC(2) \) is the marginal cost of the second unit. And so forth.

If we add all of these individual-unit profit values, then we obtain the firm’s overall variable profit. Specifically, the firm’s **producer surplus** is equal to the firm’s variable profit, the sum of the variable profit of all units sold. Graphically, in the top panel of Figure 7.13 the firm’s variable profit corresponds to the area limited below by the the \( MC \) curve and above by the price level.

Since the producer surplus corresponds to the firm’s variable profit, we can also measure it as the product of (price minus average cost) times output:

\[
(p^* - AVC(q^*)) \cdot q^* = p^* q^* - VC(q^*)
\]

where \( q^* \) is the output corresponding to \( p^* \) and \( AVC \) the average variable cost. In terms of the top panel of Figure 7.13, this corresponds to the area of a rectangle with height \( p^* - AVC(q^*) \) and length \( q^* \). It follows that the area of this rectangle is equal to the area between \( MC \) and \( p \), that is, the area \( A + B + C \) is equal to the area \( B + D \).

Just as we aggregate individual supply functions to obtain the market supply function, we can also aggregate each firm’s surplus to obtain the market producer surplus. This is illustrated in the bottom panel of Figure 7.13. As we can see,

*The industry producer surplus is given by the area limited below by the market supply curve and above by the price level, ranging from zero to the value of supply corresponding to the going price.*

(Note that I use the notation \( Q \) for industry output so as to distinguish it from \( q \), the individual firm’s output. In a competitive market, we would expect there to be many sellers and the value of \( Q \) to be measured in larger units than the value of \( q \).)
CONSUMER SURPLUS

Suppose that before going to watch a movie you stop at a pizzeria and place your order. Pizza comes at a dollar a slice (I know, I should update this example). Imagine the maximum price you would be willing to pay for one pizza slice. Perhaps three dollars, especially if you are very hungry and there is no alternative eatery in the neighborhood. Consumers don’t usually think about this explicitly; all they need to know is that they are willing to pay at least one dollar for that pizza slice. But, for the sake of argument, let us suppose that the maximum you would be willing to pay is three dollars.

How about a second slice of pizza? While one slice is the minimum necessary to survive through a movie, a second slice is an op-
It makes sense to assume you would be willing to pay less for a second slice than for the first slice; say, one dollar and 50 cents. What about a third slice? For most consumers, a third slice would be superfluous. If you are going to watch a movie, you might not have the time to eat it, anyway. If you were to buy a third slice, you would probably only eat the toppings and little else. You wouldn’t be willing to pay more than, say, 20 cents.

Putting all of this information together, we have your demand curve for pizza. The top panel in Figure 7.14 illustrates this. On the horizontal axis, we have the number of pizza slices you buy. On the vertical axis, we measure the \textit{willingness to pay}, that is, the maximum price (in dollars) at which you would still want to buy.
As discussed in Chapter 6, there are two things we can do with a demand curve. First, knowing what the price is (one dollar per slice), we can predict the number of slices bought. This is the number of slices such that willingness to pay is greater than or equal to price. Or, to use the demand curve, the quantity demanded is given by the point where the demand curve crosses the line $p = 1$. In the present case, this corresponds to two slices.

A second important use of the demand curve is to measure the consumer’s willingness to pay. You would be willing to pay up to three dollars for one slice of pizza. That is, had the price been $3, you would have bought one slice of pizza. Happily, you only paid $1 for that first slice. Since the pizza is the same in both cases, you are $3 - $1 = two dollars better off than you would be had you bought the slice under the worst possible circumstances (or not bought it at all).

Likewise, you paid 50 cents less for the second slice than the maximum you would have been willing to pay. Your total surplus as a consumer is thus $2 + 50c = $2.50, two dollars from the first slice and 50 cents from the second one. In other words, buying two slices of pizza for $1 each generated value for you as a consumer, specifically the equivalent of $4.50 of gross value or $2.50 of net value (for you spend $2 in the process).

This net value has a name: we call it **consumer surplus**. It is defined as the difference between willingness to pay and price actually paid. Recall, from Chapter 6, that willingness to pay is simply the value of the inverse demand curve (i.e., the value of $p$ for a given value of $q$). It follows that **consumer surplus is given by the difference between the inverse demand curve and price**. Adding this up for all units purchased by the consumer, we have the area limited above by the consumer’s inverse demand curve and below by price.

Just as we aggregate individual demand functions to obtain the market demand function, we can also aggregate each consumer’s surplus to obtain the market consumer surplus. This is illustrated in the bottom panel of Figure 7.14, where market demand is represented by a continuous line.

---

**Consumer surplus is given by the area under the (inverse) demand curve and above the price paid by the consumer, ranging from zero to**
7.2. GAINS FROM TRADE AND EFFICIENCY

Water: large surplus, small expenditure

Diamonds: small surplus, large expenditure

FIGURE 7.15
Application: water and diamonds

\[ Q, \text{ the total quantity demanded.} \]

Specifically, if price is \( p^* \), then quantity demanded is \( Q^* \) and consumer surplus is given by the shaded area.

THE PARADOX OF VALUE

We first introduced the paradox of value in Section 2.3 (specifically, p. 76). We will now revisit it in a more formal way. To refresh your memory, the water-diamonds paradox stems from the question, Which of the two has greater value: water or diamonds? The answer is that it really depends on what notion of value you consider (market value or value in use).
Figure 7.15 helps understand the distinction. On the top panel we have the market for water. The willingness to pay for water is very, very high when it comes to the first units you consume. You can survive without watering your lawn, maybe without showering, but I can assure you cannot survive without drinking water. Fortunately for us (in the US), water is relatively abundant, so that the market price is small. We thus live in an equilibrium where the total expense in water (light-shaded area) is relatively small compared to the (user) value created by consumer purchases (total of light- and dark-shaded areas).

By contrast, the demand for diamonds (bottom panel) is relatively more flat. If you were to ask consumers to pay a lot for diamonds eventually they would substitute something else for diamonds (jade?). Moreover, the market value of diamonds is very high, that is, they are sold for a very high price. It follows that the total expense in diamonds (red-shaded area) is relatively large compared to the (user) value created by consumer purchases (total of red- and blue-shaded areas).

In other words, the market value of diamonds is higher than the market value of water, but the value in use of water is higher than the value in use of diamonds. To put it in one sentence, I’d rather live without diamonds than without water, but I’d rather own DeBeers than ConEdison.

**ESTIMATING CONSUMER SURPLUS**

How do you find the value of consumer surplus, really? I get this question a lot. It’s easy to understand where producer surplus comes from. After all, it’s just variable profit, a concept we’re all familiar with (at least in theory). Producer surplus, or profit, is the value that sellers create for their shareholders.

When it comes to consumers, there is no accounting profit to speak of and no shareholders to distribute it to. Still, consumers do create value (for themselves) when they pay less than what they would be willing to pay. If you feel happy when you made a purchase, one reason is that you just made a “profit”, that is, you just created net value (the difference between your willingness to pay and price).
Can you put a number on the consumer’s “profit”? Yes: the difference between willingness to pay and price is a specific dollar value. Can you estimate this dollar value? To the extent that you can estimate a consumer’s or the market demand curve, you can invert it to find willingness to pay, and so compute the value of consumer surplus. In Section 6.2 we discussed the pitfalls of estimating market demand. They also apply when it comes to estimating consumer surplus.

**GAINS FROM TRADE**

We now come to one of the central results in microeconomics. One may agree or disagree with the material in this section, but one can-
not underplay the importance that it has had in economic and po-
litical thought for the past two centuries or so. Before getting into
the result itself, we discuss one of the central concepts in economics,
namely the concept of gains from trade (from which the concept of
efficiency is derived).

Perhaps it’s not immediately obvious, but trade creates value, that
is, creates surplus. If trade is voluntary, this has to be true, or peo-
ple wouldn’t do it. Consider a specific example: Jane is a big fan of
apples (the fruit, not the computer). She would be willing to pay up
to $10 for a pound of golden delicious. It costs Old McDonald 50
cents per pound to grow apples, and the current sale price is $3 per
pound. In this situation, when Jane buys her first pound of golden
delicious she makes a “profit” of $10 – $3 = $7, which we refer to as
consumer surplus. Old McDonald, in turn, makes a variable profit
of $3 – $0.5 = $2.5 from selling Jane one pound of golden delicious, a
value we refer to as producer surplus. All in all, the transaction cre-
ates a total value of $9.5, the difference between Jane’s value for the
one pound of golden delicious and Old McDonald’s production cost.

The crucial point is that the moment Old McDonald produces this
particular pound of apples he does not create as much value as when
he sells it to Jane, whose valuation equals $10, far more than it costs
Old McDonald to produce it. Naturally, when growing apples the
farmer has in mind that there are consumers like Jane who value the
fruit highly; but without trade such value is not realized.

The top panel of Figure 7.16 generalizes the concept. For buyers,
the (inverse) demand curve represents their willingness to pay. The
difference between the demand curve and market price (area $CS$) is
thus surplus to buyers. Similarly, the (inverse) supply curve mea-
sures the price at which sellers are willing to sell. The difference be-
tween price and the supply curve (area $PS$) is thus surplus to sellers.

Total surplus generated by trade, the sum of areas $A$ and $B$, mea-
sures the increase in economy-wide value that results from produc-
tion and trade: going back to our example, it measures how much
better the economy is with the existence of golden delicious apples
and the fact that there is a market where they can be traded, i.e., a
means for transferring this particular product from those who have
it to those who value it.
THE FIRST WELFARE THEOREM

Why do economists (and many politicians) wax lyrical about markets? One reason — and there are other perhaps more compelling ones — is that competitive markets are efficient, meaning that

In a competitive market the equilibrium levels of output and price correspond to the maximum total surplus.

To economists this is a sufficiently important and striking result that it has won the designation First Welfare Theorem. To rephrase the previous wording, the First Welfare Theorem states that gains from trade are maximized at the competitive equilibrium.

The opening paragraph in this subsection makes an allusion to the concept of efficiency. As justice is to law and health to medicine, efficiency is a central concept in economics, in particular microeconomics. We may distinguish different types of efficiency. The first version, allocative efficiency, requires that resources be allocated to their most efficient use. We may then rephrase the First Welfare Theorem by stating that competitive markets are efficient in this sense, meaning that resources are properly allocated. Any output level different from the market level $q^*$ would result in a lower total surplus (i.e., in lower gains from trade). In other words, it would result in an inefficient allocation of resources.

The possibility of an inefficient outcome is illustrated by the bottom panel of Figure 7.16. To continue with the example of the market for apples, suppose that only $q'$ pounds of golden delicious are transacted. There are various reasons why trade volume is $q'$ rather than $q^*$. One reason is that a regulator dictates a price ceiling $p_1$. Another reason is that a regulator dictates a price floor of $p_2$. This may not make a lot of sense, but as we will see in the next section there may be reasons to do so. For now, it’s worth keeping in mind that efficiency does not necessarily imply optimality. In other words, there may be reasons why we as a society prefer an allocation that is not efficient but is better along some other dimension.

If, because of a price ceiling $p_1$ or because of a price floor $p_2$, market output is artificially kept at a level $q'$ lower than the equilibrium (and efficient) level $q^*$, then there are a number of disgruntled buyers and sellers who would be willing to trade but do not do so because
the price is either too high or too low. Specifically, suppose we are in $E_1$. Then there are a series of buyers — those with valuations corresponding to the demand curve from $q'$ to $q^*$ — who would be willing to pay more than the sellers would ask but won’t do so because the price is too high ($p_2$). Suppose instead that we are in $E_2$. Then there are a series of sellers — those with marginal cost corresponding to the supply curve from $q'$ to $q^*$ — who would be willing to sell for less than the buyers would be willing to pay but won’t do so because the price is too low ($p_1$).

In both cases — in equilibrium $E_1$ and in equilibrium $E_2$ — total output equals $q'$, which is lower than the (unregulated) market equilibrium and efficient level $q^*$. In total, area $L$ measures the loss of value (i.e., loss in total surplus, i.e., loss in gains from trade) due to the inefficient output level. We refer to this area as the **deadweight loss** due to the deviation from the (unregulated) market equilibrium.

So, another way to rephrase the First Welfare Theorem is to state that, in a competitive market, deadweight loss is minimized at the unregulated market equilibrium.

**EXAMPLE: EUROPEAN AIRLINE DEREGULATION**

The European airline industry provides a useful illustration of the First Welfare Theorem at work. Until the early 1990s, European airline markets were highly regulated. For example, if you wanted to travel from Rome to Paris you had two options: Alitalia or Air-France. Worse than that, there was no competition between these state-owned airlines: fares were set by an open agreement between the governments of Italy and France. As a result, airfares were expensive. Very expensive. A 1987 study estimated that, controlling for flight distance, Europeans had to pay between 34 and 65 cents per mile — about three times as much as Americans.

In 1992, the European Union (EU) Regulation 2408/92 dictated that by 1997 any European carrier could offer service on any intra-European route. Gone the effective cartel agreements, the number of airlines and flights expanded during the 1990s and 2000s. By 1990, the number of European flights per week was just under 60,000. By 2000, the number was already greater than 100,000. Inevitably, the supply expansion had an effect on prices. By 2010, for similar 600-mile flights, Europeans were paying 11 cents per mile, whereas
Americans were paying 23 cents per mile. A quarter of a century earlier (in 1986, to be precise) these fares were 52 and and 15 cents per mile, respectively.

To summarize, in the 24 years from 1986 to 2010, we estimate a drop in European fares (for 600-mile flights) from about 52 to about 11 cents per mile, whereas in the US the fares increased from 15 to 23 cents! Since 2010, matters have become even worse for Americans, on account of a major consolidation of US airlines. More on this in Chapter 8.

Figure 7.17 illustrates the above narrative. The top panel describes the market equilibrium during the 1980s. Due to bilateral cartel agreements, fares were set at $p_1$ (high level) and the number of flights and passengers was $q_1$ (relatively small level). By contrast, the bottom panel describes the market during the late 1990s. As a result of the 1992 liberalization policy, the number of airlines and flights on most European routes increased. The number of passengers flown increased from $q_1$ (top panel) to $q_2$ (bottom panel). In this new equilibrium ($E_2$) fares are given by $p_2$, a lower value than $p_1$.

Liberalization of the European airline industry increased market efficiency. First, we have the area $C$ in the top panel of Figure 7.17, which corresponds to the deadweight loss from “missing trades”, that is, passengers who would be willing to pay more than it costs to fly them but not as much as the airlines were asking them to pay. In addition to allowing, say, British Airways to fly from Paris to Rome, the European liberalization process also allowed new airlines to spring up. Ryanair, EasyJet, and many other newly-created low-cost airlines increased the competitive pressure. As the name suggests, these airlines are considerably more “lean” — that is, have considerbly lower marginal cost — which allows them to offer lower fares. In terms of the graph on the bottom panel of Figure 7.17, entry by low-cost airlines may be represented by a shift of the supply curve from $S$ to $S'$. This shift implied an additional drop in fares and an increase in the volume of passengers flown — and an additional increase in total surplus, measured by area $D$.

In addition to entry by “leaner” producers, the shift from $S$ to $S'$ may also be explained by the effort that incumbent airlines had to make to keep up with competition from low-cost airlines. Economists refer to this efficiency source as productive efficiency, which we may define as the proximity of a firm’s cost to the truly
lowest possible cost. Admittedly, this is not a very rigorous definition: Up until now, we have assumed that firms choose the best combination of production inputs so as to produce output $q$ at the lowest possible cost. Well, this is one of those cases when theory and practice don’t necessarily coincide. Many firms — especially firms subject to little competition — may become “lazy” and produce in a way that does not minimize production cost (think CEO perks, shorter hours, cutting corners, etc). In other words, firms not subject to competition frequently show lower levels of productive efficiency. Harvey Leibenstein referred to this phenomenon as **X-inefficiency**.
7.2. GAINS FROM TRADE AND EFFICIENCY

European airline liberalization brought fares down by about 80%.

THE INVISIBLE HAND

The First Welfare Theorem may be rephrased with reference to the famous (or infamous) “invisible hand” of the marketplace, one of the many pioneering ideas developed by Adam Smith. Though buyers and sellers may have disparate and conflicting preferences and abilities, the price mechanism ensures that those consumers who are willing (and able) to pay the most for each good get it, and those firms for whom it is cheapest to produce those goods produce it. Moreover, in the equilibrium of a competitive market there are no remaining consumers whose willingness to buy one additional unit is greater than what it would cost any firm in the economy to produce it. That is, all trades such that willingness to pay is higher than cost take place.

We usually don’t think about it this way, but it’s a truly phenomenal achievement: a decentralized, virtually costless system (the price system) manages the allocation of multiple resources among millions of agents in an efficient way (in the sense of allocative efficiency). Unlike central planning, which requires a very “heavy” structure to collect and process information, the price system accomplishes the efficient allocation of resources in a very simple, “lean” way. Specifically, all that each buyer has to do is answer the question: Is my willingness to pay for \( x \) greater than price? If so, then I will buy, if not then I won’t. Similarly, all that each seller has to do is answer the question: Is my cost (i.e., willingness to sell) of \( x \) lower than price? If so, then I will sell, if not then I won’t. The beauty of it is that this rule is both optimal for the individual buyer, the individual seller, and optimal for society as a whole insofar as we’re interested in maximizing gains from trade.
This is what Adam Smith meant by the “invisible hand of the marketplace.” We may agree or disagree with the accuracy and relevance of the idea. However, we cannot understate the importance that it has had in economic thought — and in political thought as well — for more than two centuries. Within the field of economics, these ideas were developed and popularized in the 20th century by economists such as Frederich Hayek and Milton Friedman. Under these and related authors, the First Welfare Theorem becomes part of the larger edifice of economic and political thought that includes liberalism, individualism and libertarianism. What these currents of thought have in common is their accent on individual freedom — including individual economic freedom — often in opposition to the role of government as a market regulator or as an agent of solidarity.

**SURVIVAL OF THE FITTEST**

When discussing the effects of liberalization of the European airline industry, we saw that putting an end to the bilateral cartels brought fares down (and increased the number of flights). In terms of the “invisible hand” narrative, this was the sign to European travelers with mid-level willingness to pay that it was OK to fly. In fact, their willingness to pay was considerably higher than the marginal cost of carrying them. Therefore, the increase in volume of air traffic created value.

We also saw that an additional source of value was the entry of low-cost airlines such as Ryanair and easyJet. One thing I did not mention but played an important role was the exit of a variety of (not so efficient) airlines, such as SwissAir or Sabena (Belgium’s national airline). In other words, an additional benefit from increased competition is the selection of more efficient producers. In this sense, more than Adam Smith’s “invisible hand”, the appropriate metaphor seems to be Charles Darwin’s “natural selection” (only the more efficient firms survive). As Warren Buffett aptly put it, “only when the tide goes out do you discover who’s been swimming naked.” In the present context, this might be rephrased as “only when the tide of perfect competition takes place do you discover which firms are not clothed with the garb of productive efficiency.” And no, I don’t expect this sentence to help me towards a Pulitzer prize, but you get the idea.
Maximizing allocative and productive efficiency is no doubt a good thing. However, a few observations or caveats are worth highlighting. First, the First Welfare Theorem is a statement about efficiency rather than equity, that is, it concerns the size of total surplus, not its distribution. In particular, note that in calculating total surplus we value consumer and producer surplus equally. Presumably, we attach value to profits because they are eventually returned to the firm’s shareholders. However, inasmuch as shareholders tend to be wealthier than consumers, many would argue that firm profits ought not to be weighted as much as consumer surplus. For example, institutional arrangements or regulations that raise consumer surplus by the equivalent of $20 million and decrease firm profits by $25 million may be judged to improve welfare even at a cost of $5 million in terms of total surplus.

Another important note is that the First Welfare Theorem (as stated above) is about static efficiency, that is, the optimal allocation of resources given the current set of products and production technologies. To continue with the airlines examples: Competitive markets lead to the efficient number of flights, but do they also lead to the right level of investment in new transportation technologies or investment in new energy sources? It is difficult to measure these dynamic effects, even more difficult to compare them to measures of static efficiency. Unfortunately, this creates a bias: economist focus too much on efficiency largely because they can measure it. Unfortunately, this is not an innocuous bias.

Last but certainly not least, an important caveat is that the First Welfare Theorem applies to competitive markets, which, as outlined earlier in the chapter, correspond to some fairly strong assumptions. When producers or consumers are large enough to affect market prices; or when products are differentiated; or when there is less than perfect information about price and quality; or when entry into the industry is restricted; or when property rights are not properly established — then the First Welfare Theorem does not necessarily hold. More on this in the next three chapters.

Given all of these caveats, one may ask the question: Why pay any attention at all to the First Welfare Theorem? If we are to be rigorous, there isn’t any truly competitive market in the world, so isn’t all of
this much ado about nothing? One possible defense of the economics approach is that the competitive market model provides a reference point: To the extent that a real-world market is close to a competitive market, one might say that efficiency is close to maximized under equilibrium conditions. In his famous 1877 novel, *Anna Karenina*, Russian writer Leo Tolstoy famously noted that

> All happy families are alike; each unhappy family is unhappy in its own way.

As we will see in the next three chapters, there are many reasons why real-world markets are not quite the competitive market described earlier in this chapter. Still, it is helpful to understand the behavior of competitive markets, the “happy family” of economics. But there’s more: Even the characterization of competitive markets as a “happy family” is open to debate. One must remember that efficiency is only a partial — at times very partial — performance measure.

Parts IV and V of this book deal with these two observations. Part IV focuses on market failures, that is, markets that, for one reason or another, are not competitive markets. As we will see, this typically implies that the market equilibrium is not efficient. Part V, in turn, explores optimality dimensions that go beyond efficiency: free markets may be very efficient, but this may serve some better than others.

**NON-MARKET LOGIC**

Among the many assumptions underlying the First Welfare Theorem and the role of the invisible hand of the marketplace is the assumption that economic agents are individual, logic, selfish maximizers who only respond to extrinsic, material motivation. No one believes this is entirely true, but most accept it as a useful starting point. But is this a good approximation? As usual, the answer is — it depends, specifically, it depends on the particular setting.

Consider the following study based on Israeli daycare centers. A common 21st-century problem in daycare services is that parents are often late to pick up their children at the end of the day. At a number of Israeli daycare centers, late arrivals used to be punished with a cold stare or a similar form of social censure. But seeing how parents continued to arrive late, some daycare centers agreed to test a new
policy, namely to charge a ten shekel fine for lateness (about $3 at the time). The results of this experiment were quite surprising (at least to some). Instead of reducing lateness, the fines led to an increase in the frequency of late pickups: on average, the rate doubled with respect to treatment group (that is, the group of daycare centers not imposing fines).

One possible explanation for this pattern is that, before the fine was introduced, most parents were on time because they felt that it was the right thing to do, or because they were concerned with the social punishment they suffered from the daycare staff if they were late. By contrast, once the fine is introduced, may parents think of lateness as a possibility they can “buy” by paying the fine. In other words, once there is a “market” for lateness parents no longer see punctuality as a moral obligation.

One important learning point from this experiment is that incentives, in particular market incentives, are not the sole source of extrinsic motivation. In fact, the introduction of market incentives may crowd out other sources of incentives: as we increase monetary incentives we decrease incentives based on our sense of decency. To put it differently

> Human behavior in a social context is determined by a variety of "logics" in addition to the logic of the market.

We return to this issue at other points in the book. (This Israeli daycare center has proven a rather controversial one in the economics profession. The interested reader might be interested in a methodological criticism and the authors’ response to it.)

7.3. PRICE CONTROLS

Efficiency is not the only goal society should be concerned with. There may be many good reasons why governments take actions that effectively disturb the “natural” equilibrium of an otherwise competitive market. In this section we consider four particularly important cases: taxation, rent controls, minimum wage, and price gouging.
TAXES

There are many reasons why government levy taxes and there are many different types of taxes. Historically, taxes were first and foremost a form of raising income to cover for government expenditures. In this context, it may be useful to estimate the efficiency cost of raising $1 of revenue. This we do next. Specifically, let us return to the example considered in Section 7.1, a tax on gasoline consumption. To recap, the supply and demand functions are given by (p in dollars, q in million gallons a day)

\[ Q_D = 150 - 50p \]
\[ Q_S = 60 + 40p \]

The initial (pre-tax) market equilibrium results from the equality of supply and demand:

\[ 60 + 40p = 150 - 50p \]
\[ p = (150 - 60) / (40 + 50) = 1 \]
\[ q = 60 + 40 = 100 \]

So, initial equilibrium price is $1 per gallon and a total quantity of 100 million gallons is sold.

Now suppose the government imposes a tax of \( t = 56.25 \) cents per gallon. Specifically, suppliers now receive the sale price but must pay 56.25 cents per gallon sold to the tax authority. It follows that they receive a net price of \( p \) minus 56.25 cents. This implies that the new supply curve is given by

\[ Q_S = 60 + 40(p - .5625) \]

The new market equilibrium (with a 56.25 cents/gallon tax) is given by

\[ 60 + 40(p - .5625) = 150 - 50p \]
\[ p = (150 - 60 + 22.5) / (40 + 50) = 1.25 \]
\[ q = 60 + 40(p - t) = 60 + 40 (1.25 - .5625) = 87.5 \]

All this we saw in Section 7.1. We may now ask: what is the excess burden — that is, the loss in allocative efficiency — created by the
tax? Figure 7.8 illustrates the answer to this question. As a result of the tax, buyers lose part of their surplus and sellers, too, lose part of their surplus.

Part of these losses in surplus are compensated by the gain to the tax authority. To the extent that these funds are used to benefit buyers and sellers (e.g., better infrastructure or a better mass transit system) we should not include them as efficiency losses. However, there is a part of the loss in total surplus which is not compensated by an increase in tax revenues, namely the area \( L \) in Figure 7.18.

Since both the demand and the supply curve are linear, the deadweight loss corresponds to the area of a triangle. Specifically,

\[
L = \frac{1}{2} \times (100 - 87.5) \times .5625 = 3.515625
\]

We can also compute the loss in consumer surplus due to the tax. Consumers lose the area given by the trapezoid \([ABDC]\), which is given by

\[
\frac{1}{2} \times (87.5 + 100) \times (1.25 - 1) = 23.4375
\]

Regarding producer surplus, we have a loss of given by the area of the trapezoid \([CDFE]\), that is

\[
\frac{1}{2} \times (87.5 + 100) \times (1 - .6875) = 29.296875
\]
As to the government, a tax of 56.25 cents generates a tax revenue given by the area of the [ABFE] rectangle, that is,

$$87.5 \times 0.5625 = 49.21875$$

Gains and losses must add up. Specifically, the loss in total surplus must equal tax revenue plus deadweight loss. Let’s check

$$23.4375 + 29.296875 = 52.734375$$

$$49.21875 + 3.515625 = 52.734375$$

There are certainly good reasons to levy a tax — or, more generally, to take actions which move us away from the competitive market equilibrium. That said, it’s important to know this comes at a cost, namely a cost in terms of efficiency loss. In the present case, the loss is given by $L = 3.515625$, which represents a little over 7% of the tax revenue. Another important observation regarding the gasoline tax is that, while it is nominally paid by the seller, its effect is felt both by seller and buyer. In fact, the loss in surplus is about the same for seller and buyer. This and other points regarding taxation will be examined in greater detail in Section 12.3.

### IMPORT TARIFFS

Trade creates value — this is, to a great extent, the central theme of this chapter. In particular, international trade creates value. Much of the growth experienced by China and other economies in the past decades can be attributed to the value-creation nature of international trade.

In this context, an **import tariff** — or an **import quota** — can have a value-destruction effect. Can we measure this? Essentially, it corresponds to the deadweight loss derived before in this section. Consider the bottom panel of Figure 7.16. Suppose the supply corresponds to exports from China to the US, whereas the demand corresponds to US demand for these imports. Absent any trade regulation, the equilibrium level of exports/imports is given by $q^*$. Suppose however that an import tariff is set such that consumer price increases to $p_2$. Alternatively, suppose that an import quota is set such that not more than $q'$ can be imported. Regardless of the specific nature of the policy, we observe a decrease in imports and thus
a reduction in the number of trades. These were all trades for which the willingness to pay by US consumers was higher than the production cost by Chinese manufacturers. When we add these up we come to the area $L$, a dollar value measuring the “lost trades” due to the import tariff or import quota.

If import tariffs destroy value, why are there import tariffs at all? First, sometimes import tariffs are part of a broader game played by countries. Specifically, tariffs may serve as a tool for “punishing” a foreign country for what it has done or has failed to do. Second, an import tariff drives foreign producers out of the domestic market, which in turn opens the way for less efficient domestic producers. Thus domestic consumers pay a higher price — and many trades are “lost” — but in return domestic jobs are created or saved. This may be a very inefficient way of creating or maintaining jobs at home, but it’s one governments may opt to — especially if subject to intense lobbying by the protected domestic industry.

The issue of **lobbying** is sufficiently important to justify a brief digression. Economic policy is part of overall government policy and the rationale for specific measures is to be found in the process of policy-making rather than the economic efficiency calculations considered in this chapter. A particularly important aspect of the policy-making process is the **special interests** paradigm. An import tariff on steel, for example, implies an enormous benefit for a small number of agents (domestic steel producers) and a small cost for a large number of agents (domestic consumers of products which incorporate some steel). The incentive for a domestic steel company to lobby for a tariff is thus very high, whereas an individual consumer’s incentive to lobby against a tariff is very low. This results in an unbalanced political process whereby special interests are given disproportionate weight.

If you think about it, this phenomenon (the benefit is concentrated in a small number of agents, the cost is spread among a large number of agents) is more general and applies to other instances of public policy. In the next section we consider another specific example of special interests and the effect they might have on market regulation.
7.3. PRICE CONTROLS

RENT CONTROL

A well known 1990 survey of economists found that 93.5 percent — the highest percentage of any question — was in agreement or qualified agreement with the statement that “A ceiling on rents reduces the quantity and quality of housing available.” The argument regarding quantity is simple: Suppose that the bottom panel of Figure 7.16 represents the rental housing market. Suppose that the government sets a maximum rent (price) of $p_1$. Then the supply of rental dwellings drops from $q^*$ to $q'$. The argument regarding quality is not as straightforward, but the idea is that if landlords earn less from renting out a building, then the incentives to maintain the building are commensurably lower.

Notwithstanding this broad agreement regarding the negative effects of rent control, the policy is still very common in many cities throughout the world. As of 2019, there are approximately 200 cities in the United States with some type of rent regulation. Moreover, a variety of jurisdictions are considering or implementing new rent regulation.

There are a number of arguments in favor of rent regulation. The most common one is that rent control helps the poor, which is particularly important at a time of rising income inequality. This is a highly debatable argument, as we will see below in this section and later in Chapter 12. A second argument, less based on economics and more based on social and psychological considerations, is that we should recognize the legitimate interest of long-term tenants in remaining in their homes. Even if a house is rented, it is still a family home, and its tenants have a reasonable expectation of remaining in it on terms similar to those they have enjoyed in the past. A third, related argument is that there is a social interest in diverse and stable neighborhoods. For example, some complain that most artists had to leave New York’s Greenwich Village as rents skyrocketed to unaffordable levels, and that the Village lost a lot with the exodus.

There have been many empirical studies on the effects of rent controls, and the evidence is mixed. One problem with many studies is the usual problem with empirical social science: distinguishing between correlation and causality in a world where lots of things change at the same time. For this reason, a sudden change in legislation which took place in Massachusetts provides a useful testing
In November 1994, landlords succeeded in placing an initiative on the Massachusetts ballot to ban rent control statewide. In November 1995, the proposal was approved statewide (51% in favor, though in the cities that had rent control — Boston, Brookline, and Cambridge — the vote was overwhelmingly against). One study shows that — not surprisingly — rents increased following the repeal of rent control. Moreover, both the quantity and the quality of rental housing supplied increased (as predicted by the basic economics of supply and demand).

One year before Massachusetts, another important ballot was (also narrowly) approved: San Francisco repealed an exemption on rent control, that is, it moved in the opposite direction of Massachusetts by extending the reach of rent regulations. Specifically, all multi-family structures with four units or less built in 1979 or earlier became subject to rent control (whereas before 1994 they were exempt from rent control). Like Massachusetts, the change in legislation provides a unique window on the effects of rent regulation. A study looking at very granular data — basically at the unit level — reaches the following conclusions. First, the probability that a beneficiary of rent control remains in their unit is about 20 percent higher than the control group. This difference in persistence rates is particularly significant for older tenants and tenants who had lived in their rent-controlled unit for longer. Second, the data suggests that, through this persistence effect, rent controls contribute to racial diversity. The reason is that minorities are disproportionately represented among rent-control tenants.

Third, the switch to rent control also had an effect on the supply side. The law allows landlords to convert a building into a condo, thus removing its units from the rental market. The data shows a 15 percent decline in available rentals among the previously rent-control-free units. This decline in supply likely led to a long-term increase in rents (following the basic forces of supply and demand). Moreover, the newly converted units attracted in-migrants with income levels about 18 percent higher than previous tenants.

Overall, the evidence vindicates both the arguments in favor and against rent control (and in fact the study has been used as supporting evidence for both sides of the argument). The immediate effect is, to a great extent, the intended effect: keeping poorer and minority tenants in the neighborhood. However, the long-term effects likely
work against those that the policy is supposed to protect. To some extent, rent controls benefit the present generation of disadvantaged people but hurt those who come later.

**MINIMUM WAGE**

Similar to rent control, minimum wage policies tend to be big dividers, both across the political spectrum and between economists and non-economists. The neoclassical economics argument is essentially the same as the argument against rent control (in this regard, neoclassical economics is rather boring): By setting a minimum wage, the equilibrium quantity of labor decreases, thus creating a deadweight loss corresponding to hires that should take place but do not take place. However, unlike the case of rent controls — which largely confirm the economics prediction — the evidence on minimum wage is largely at odds with the prediction based on the bottom panel of Figure 7.16.

One particularly well-known study looked at employment in New Jersey and Pennsylvania. In 1992, the minimum wage in New Jersey increased from $4.25 to $5.05 per hour, while in the adjacent state of Pennsylvania it remained at $4.25. For an economics researcher, this provides a unique opportunity to identify causality by means of a difference-in-differences research design. The idea is to compare the changes in employment in two “identical” establishments, one on each side of the border. If the two establishments are sufficiently close to the border, then the case can be made that the minimum wage is the only relevant difference in the environments they face. Moreover, by comparing changes in employment rates we
control for a host of factors that might have affected employment rates and are not related to minimum wage. The study concludes that New Jersey’s increase in minimum wage slightly increased or had no effect on employment in New Jersey restaurants (the set of establishments considered in the study), which is at odds with what the supply-and-demand model would predict.

A number of subsequent studies, using different data sets or different methodologies or both, reached conclusions closer to the neoclassical economics prediction: an increase in minimum wage leads to a decrease in employment. However, it is remarkable how the surprising result from the 1992 New Jersey case has been extended to other states and years. For example, a recent study shows that increases in minimum wage in New York state have had no effect on employment. New York state — among other states — has committed to push the state minimum wage to $15. Currently it’s set at $12.50. Meanwhile, other states — such as Pennsylvania — have stuck to the federal level of $7.25 (which in turn has remained constant for many years). Similarly to New Jersey in the 1990s, this new study finds no losses in employment at New York state restaurants close to the Pennsylvania border when compared to their cross-the-border counterparts.

Additional research into specific establishments suggests businesses cover higher labor costs by taking a hit to profits, improving productivity or simply raising prices. This in turn suggests that the industry may not be as close to the competitive market reference point as the efficient markets view suggests. We will return to this in the next chapter.

**PRICE GOUGING**

In the wake of unexpected negative supply shocks — or positive demand shocks — prices tend to surge to abnormally high levels. For example, during the COVID-19 pandemic the price of a 3M mask on Amazon increased from less than $10 in 2019 to — in some cases — more than $300 in March 2020. Many efficient-markets economists find such price hikes normal and in fact efficient: it’s the law of supply and demand at work. By contrast, most of the world accuses sellers of price gouging and clamors for laws protecting buyers from unfair pricing. For example, in the aftermath of Hurricane Sandy,
which struck the US in October 2012, New Jersey authorities filed civil suits accusing seven gas stations and one hotel of price gouging. Meanwhile, Libertarian TV personality John Stossel was inviting his viewers to “hug a price gouger today.” The contrast could not be greater.

Figure 7.19 (source) shows the online price of 3M masks sold during the first months of 2020. Specifically, the figure plots the ratio between the price set throughout 2020 and the average price set by Amazon during 2019 (a natural reference point for prices during a non-crisis period). Three groups of sellers are considered: Amazon, who is both seller and the platform “host” of other sellers; “incumbent” sellers, that is, sellers who already sold before January 15, when the first US case was announced; and “entrant” sellers, that is, sellers who began selling after January 15. Several features are noticeable. First, despite the enormous spike in demand caused by COVID-19 fears, Amazon barely increased its price. In addition to legal concerns — price gouging is illegal in most US states — there are also reputational concerns: the last thing Amazon needs is to be called a price gouger and be the target of a public boycott campaign. Amazon sells millions of different products, it would make no sense to risk those revenue streams for the prospect of a small gain in selling masks.
Second, sellers other than Amazon increase prices considerably. For entrant sellers, the average price ratio can be as high as 8. Underneath this average there is considerable variation, including sellers setting prices 30 or 40 times higher than Amazon.

Third, there is a significant difference between the price patterns for “incumbent” and “entrant” sellers, which suggests that the reputation argument works not just for Amazon but also for continuing sellers, who have something to lose from being given the “price gouger” label.

Last but not least, notice that — not surprisingly — the sellers setting lower prices run out of stock more quickly. In other words, after mid February it is little consolation to say that Amazon sets a low price, for they are simply out of stock. In a certain sense, it’s as if they set an infinite price.

If you ask anyone on the street what they think about price gouging, their first reaction will likely be “it’s just not fair.” It’s not fair to charge more than $300 for something that just a few weeks ago cost less than $10. An economist might reply “how do you know the seller’s cost did not increase a lot too? More importantly, suppose that I only have one mask and that there are two interested buyers. The best way to decide whom to sell the mask to is whoever is willing to pay the most, and that’s what a high price does: separate those who need the mask from those who really need it.”

You can see the potential flaw in the argument. Under many circumstances, using willingness to pay as a measure of need may make sense. In an emergency situation, it might be a stretch. Joe is a billionaire and already has five masks at home. Strictly speaking he does not need an extra one, but who knows. And $300 is nothing to him. Across the street, a nurse who works with COVID-19 patients desperately needs a mask but simply cannot afford the $300 required to buy it on Amazon. It’s not fair!

There is, however, a much stronger economics argument in favor of allowing very high prices: incentives. Let us go back to the example at the beginning of the section: It’s October 2012 and Hurricane Sandy has just hit New Jersey. Gas stations have very low stocks and begin to ration by increasing gas prices. The governor declares a state of emergency and, with that, price gouging (“unconscionably” high prices) is deemed illegal. New Jersey Governor Chris Christie enacted a price ceiling and threatened a “zero-tolerance” approach
to violators. A big shift in the supply curve together with a price cap leads to an enormous excess demand. Car drivers line up for hours to fill up, which they may or may not succeed in doing.

Suppose instead that Christie allowed for prices to increase. Then clever entrepreneurs would think of ways of filling up large trucks in Philadelphia and drive them to New Jersey. With gas at $10 a gallon it pays to do so. With gas at $3 per gallon it does not. In sum, to the extent that the supply elasticity is high (e.g., supply is sensitive to price changes), the efficiency costs of a price ceiling can be significant.

As often is the case, there are trade-offs. There is something to be said for the efficiency role of the invisible hand of the marketplace (price). But there are many situations, including emergency situations, when markets are simply not competitive or efficiency is not the most important imperative (or both).
KEY CONCEPTS

- competitive markets
- homogeneous product
- price taker
- well-defined property rights
- perfect information
- market equilibrium
- excess supply
- excess demand
- market-clearing price
- law of supply and demand
- comparative statics
- producer surplus
- willingness to pay
- consumer surplus
- market value
- value in use
- First Welfare Theorem
- efficiency
- allocative efficiency
- deadweight loss
- productive efficiency
- X-inefficiency
- crowd out
- import tariff
- import quota
- lobbying
- special interests
- difference-in-differences
- price gouging
PART IV
MARKET FAILURE
In a competitive market, firms are small relative to the market. In fact, in Part III we considered the extreme case when firms are so small that the demand curve they face appears to be flat. In other words, from each firm’s point of view demand is extremely sensitive to price changes: the tiniest change in price would lead to the greatest change in output (from their own small-firm perspective).

In behavioral terms, we say that, in competitive markets, firms are price takers. This was the case, for example, with the small t-shirt factory considered in Chapter 6. Price-taking firms set output levels such that marginal cost is equal to price (cf Section 6.1). In other words, in competitive markets the resulting equilibrium price corresponds to the marginal cost of producing each firm’s \( q \)th unit.

However, it’s fair to say that in the real-world most firms are price makers, not price takers. Apple, to take a somewhat extreme example, is certainly not a price taker: Who decides the price of an iPhone 11? Largely, it’s Apple, not the market. Wireless companies such as Verizon may also have a say, but these are certainly not price takers. In this chapter, we consider the case when firms are sufficiently large so that their individual behavior has an impact on market outcomes. We begin by considering the sources and effects of market power, considering both the case of monopoly (one seller) and the case of oligopoly (a few sellers). We also report on recent trends in market power. The chapter concludes with a section on public policy directed at market power.
Whenever firms have power to set price — that is, whenever they are not price takers — we say there is market power. The degree of market power depends on industry structure (e.g., how many competing firms there are) as well as on the nature of demand (in particular, how sensitive demand is to price changes). In terms of industry structure, one extreme corresponds to monopoly, that is, the case when there is only one seller (and many buyers on the demand side). For example, Electricité de France is a monopoly supplier of electricity to French households. Another industry structure prone to market power is that of oligopoly, the case when there is a small number of competitors. Sometimes, there are a few, similar sized, competitors. For example, the art auction market is dominated by Christie’s and Sotheby’s. In other cases, there is a dominant firm and a bunch of smaller ones. For example, in the market for online searches “giant” Google competes against a series of smaller (sometimes niche) search engines.

How do monopolists become monopolists? One source of monopoly power is intellectual property. For example, in the 1980s Pfizer patented the medical drug atorvastatin, which it sold under the brand name Lipitor. The patent expired in 2011 (patents typically last for twenty years). Until then, Pfizer was a monopoly supplier of atorvastatin. A related source of an intellectual property right leading to monopoly is given by copyright. Disney, for example, has the exclusive right over the movie and the musical Lion King. No matter how good my singing and dancing abilities may be, I am not allowed to compete against Disney with my own Broadway production of Lion King (unless Disney gives me permission, which I doubt they will). Unlike patents, copyrights last for many decades.

But you don’t need government protection in order to become a monopolist: trade secrets also do the job. For example, Google’s search engine is not patented. Why don’t other firms copy Google’s algorithm? Because they don’t know how it works. Similarly, the legendary secret formula for Coca-Cola — regarded as the most closely...
SOURCES AND EFFECTS OF MARKET POWER

Patent protection is a common source of monopoly power.

Still another source of monopoly (or quasi-monopoly) power is given by network effects. This has become particularly important in recent decades, as the digital economy becomes an increasing slice of the modern economy. What are network effects? Many people use the Microsoft Windows operating system primarily because most other people use Windows as well. It’s not that I particularly like Windows (I don’t), but if most of the world is stuck on Windows then I might as well choose Windows. This provides Microsoft a degree of market power it would not have absent these network effects.

On the demand side, market power is determined by the degree of price sensitivity. As illustrated by Figure 5.11, if demand is very sensitive to price changes then the seller’s optimal price is relatively low. In this sense, it does not matter whether the seller is a monopolist or not. The fact that demand is so sensitive to price suggests that, de facto, the seller is not a monopolist: If demand is so sensitive to price changes, it must be that consumers have alternatives to the product sold by the putative monopolist.

OLIGOPOLY

Competitive markets (“many” sellers) and monopoly (one seller) are two extreme cases of industry structure. Most real-world markets fall somewhere in-between: more than one but fewer than “many” competitors. For all their (extreme) differences, monopolists and small, price-taking firms have one thing in common: neither needs to be
8.1 SOURCES AND EFFECTS OF MARKET POWER

FIGURE 8.1
Pricing game

Concerned with competitors. Monopolists have no competitors, and price takers are small enough that their actions have no effect on other price takers. By contrast, oligopolists must take into account their rivals’ actions. We thus enter into game theory territory (again).

BEST RESPONSES AND NASH EQUILIBRIUM

In Section 2.2 we introduced the first elements of game theory. To refresh your memory, we saw that a **game** is a model which combines four elements: **players, rules, strategies, and payoffs**. We also saw how you can represent simultaneous-moves two-player games as a matrix game: one player picks a row while another player picks a column. Finally, we presented a special type of game, the prisoner’s dilemma. This game has the feature that both players have a dominant strategy but the resulting outcome is the worse for both players. The climate change game played by China and the US was presented as an example.

Consider now the game represented in Figure 8.1. It describes the pricing game played by two competing firms. Each firm can set a high (5), a medium (4) or a low (3) price. Unlike the climate change game introduced in Section 2.2, no player has a dominant strategy in the pricing game in Figure 8.1 (check). As such, we cannot analyze (“solve”) the game in the way we analyzed the prisoner’s dilemma. Instead, we proceed by introducing a new concept: best responses.
8.1. SOURCES AND EFFECTS OF MARKET POWER

A player’s **best response** (or simply BR) is a mapping indicating their optimal choice for each possible choice by the rival player.

The term “response” can be a bit confusing, for we are not talking about players moving one after the other. Think of best response as my optimal choice if I believe or expect the other player to choose a certain action. Specifically, in the present pricing game Firm 1’s best response is as follows: if Firm 2 sets \( p = 5 \) (that is, if Firm 1 expects Firm 2 to set \( p = 5 \)), then Firm 1 should set \( p = 4 \). In fact, setting \( p = 4 \) gives Firm 1 a profit of 12, whereas \( p = 5 \) leads to 7.5 and \( p = 3 \) leads to 7. By the same token, if Firm 2 sets \( p = 4 \) or \( p = 3 \), then Firm 1’s best response is to set \( p = 3 \). Firm 2’s best response is similar.

Equipped with the best-response mappings, we are now ready to derive our prediction of the outcome of the play of the game. The most common way of doing this is to derive the equilibrium of the game. Specifically,

A **Nash equilibrium** (or simply NE) of a game is a combination of strategies (one for each player) such that no player can improve their payoff by unilaterally changing their strategy.

The concept of Nash equilibrium is closely related to the concept of best responses. By definition, a given combination of strategies \((a, b)\) forms a Nash equilibrium if and only if \(a\) is optimal for Player 1 given that Player 2 chooses \(b\); and \(b\) is optimal for Player 2 given that Player 1 chooses \(a\). Another way to state this is that \(a\) is Player 1’s best response to Player 2’s choice of \(b\) and \(b\) is Player 2’s best response to \(a\). We thus conclude that \((a, b)\) forms a NE if and only if both \(a\) and \(b\) belong to the players’ best responses.

In other words, we find the game’s Nash equilibrium (NE) by the “intersection” of the two firms’ best responses, that is, a combination of choices for Firm 1 and Firm 2 such that Firm 1 does its best given what Firm 2 does and vice-versa. In the present context, this corresponds to both firms setting a low price, i.e., \( p = 3 \).

So far so good: the NE of the pricing game suggests that, while the number of competitors is very small (only two), competition may lead them to set a low price, presumably a price close to their production cost. In other words, even though the number of competitors
is small, the outcome looks similar to that of a competitive market: not much market power. The problem is that when the number of players is small and their interaction frequent, then there is scope for collusion among competitors, that is, firm behavior that effectively leads them to set high prices ($p = 5$ in the present case).

From a game theory point of view, this presents a puzzle: We just saw that, if Firm 1 expects Firm 2 to set $p = 5$, then Firm 1 is better off by setting $p = 4$. So how can it be that firms collude by setting $p = 5$. In game theory jargon, how can $p = 5$ be part of an equilibrium? The answer is that the game real-world firms play is different from the game in Figure 8.1. This is true in many dimensions, but a particularly important one is that firms typically interact over many periods.

### REPEATED GAMES AND COLLUSION

A **repeated game** is game theory’s way of modeling on-going interaction between players. As the name suggests, a repeated game is simply a game obtained from repeating a one-shot game (like the one in Figure 8.1). For example, suppose that two gas stations set prices daily and that the matrix in Figure 8.1 reflects daily payoffs. Acknowledging that the two gas stations compete day after day, consider the following tacit agreement: set $p = 5$ as long as in the past both firms set $p = 5$, and set $p = 3$ otherwise. That is, if any firm “deviates” from the tacitly agreed-upon $p = 5$, then firms “plunge” into a “price war”, that is, set $p = 3$ thereafter.

In this proposed equilibrium of the repeated game each firm earns a profit of 7.5 in each period. Is this really an equilibrium? Wouldn’t firms have an incentive to undercut their rival? By doing so, a firm would earn a profit of 12 in the current period. However, beginning with next period, firms would switch to $p = 3$ for ever, leading to a profit of 3.5 per period. Effectively, firms compare the following two alternative profit streams:

(a) cooperate: 7.5, 7.5, 7.5, ...

(b) defect: 12, 3.5, 3.5, ...

Depending on how important the current period is with respect to the future, “cooperation” (that is, setting a high price) may be the best
response and (5,5) a NE. Specifically, if firms interact very frequently, then the short-term gain (12) is of little significance with respect to the long-term losses.

To conclude the analysis of the repeated game, notice that we assumed there are only two players. If instead there were many gas stations vying for market share, then the temptation to lower prices would likely be greater. The idea is that, if all gas stations set a high price, then each gets a small market share, but if one of them undercut the competition then it enjoys a big boost in terms of market share. In sum,

*If the number of firms is small and firms interact frequently then it may be an equilibrium for firms to set high prices.*

If high prices are an equilibrium and if firms are better off by setting high prices, then we should expect them to collude. There is only one problem: While firms are better off, consumers are worse off. Consumers would much rather if firms played the low price equilibrium. Moreover, as we saw in Chapter 7, the case can be made that, overall, society is worse off when firms collude by setting high prices. For these reasons, a formal agreement to set high prices is illegal (in fact, a crime in the US), as we will see in Section 8.2. However, this does not stop firms from forming secret cartels or colluding tacitly (i.e., without explicit communication). More on this in Section 8.2.

To conclude this section, I note that, although we reached the above repeated-game results in the context of collusion between competing oligopolists, the principle is sufficiently important to deserve special mention:

*Repeated interaction between players may lead to equilibrium outcomes (e.g., cooperation) that would not be equilibrium outcomes if the underlying game were played only once.*

I cannot emphasize enough how general and important this point is: A multitude of tacit agreements in society — including many social norms and what economists refer to as “relational contracts” — are enforced not so much by the rule of law as they are by the repeated interaction among society members. If I go to a coffee shop tomorrow, order breakfast and then walk out without paying — thus vio-
lating either the law or a social norm or both — it’s unlikely the coffee shop will sue me in court. However, it’s almost certainly the case that the coffee shop will not allow me in the their establishment again (a “punishment” strategy in the coffee-shop repeated game akin to setting a low price in the pricing game). More generally, we can think of our society’s institutions as a means to convert a prisoner’s dilemma (where all society members play their *myopic* dominant strategy and all receive a low payoff) into a cooperative equilibrium (where each member sacrifices their short-term gain for the sake of an overall societal gain). End of digression.

**OLIGOPOLY: PREEMPTION BY CAPACITY EXPANSION**

In the mid 1970s DuPont, one of the world’s largest chemical corporations, engaged in a curious strategy in the titanium dioxide industry: it build production capacity well in excess of its needs. As a result, in a relatively short period of time, DuPont’s market share increased from about 30 to more than 50 percent. Box 8.1 provides additional details.

What sense does it make for a firm to build more capacity than it needs? In order to understand the nature of DuPont’s strategic move, consider the following production capacity game. There is an incumbent firm which must decide whether to build a “normal” capacity level or a larger capacity level (which essentially amounts to excess capacity, that is, capacity that is not actually used). There is a potential competitor who must decide whether to enter and challenge the incumbent or rather stay out.

Consider first the case when the two players move simultaneously.

<table>
<thead>
<tr>
<th>Incumbent</th>
<th>Entant</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>low K</td>
<td>enter</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>high K</td>
<td>enter</td>
<td>0</td>
<td>-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

**FIGURE 8.2**
Production capacity game with simultaneous moves
Titanium dioxide (TiO\textsubscript{2}) is a white chemical pigment employed in the manufacture of paint, paper and other products to make them whiter or opaque. The primary raw material for the production of TiO\textsubscript{2} is either ilmenite ore or rutile ore.

By 1970, there were seven firms in the industry: a large firm, DuPont, and six smaller ones. During the 1960s, DuPont used mainly ilmenite, whereas its rivals used mainly rutile. In 1970, a sharp increase in the price of rutile ore created a significant cost advantage for DuPont with respect to its rivals: at 1968 ore prices, Dupont had a cost advantage of 22%; at 1972 prices, this advantage averaged 44%. Moreover, stricter environmental regulation meant that several of DuPont competitors would have to incur large costs in order to continue production.

DuPont found itself with a competitive advantage in several dimensions. First, its production process used a cheaper input than most of its rivals. Second, its production process complied better with environmental standards. Third, because of the cost advantage, the firm was in better financial shape, thus better positioned to expand capacity.

A Task Force was formed at DuPont to study how to turn these advantages to the firm’s greater benefit. The result was the strategy of expanding capacity at a pace sufficient to satisfy all of the growth in demand in the ensuing years. The idea was that by expanding rapidly, DuPont would discourage expansion (or entry) by rival firms. It was the Task Force’s conviction that deterrence of competitive expansion was necessary if DuPont was to establish a dominant position: according to the plan, DuPont’s market share would increase from 30% in 1972 to 56% in 1980 and perhaps 65% in 1985.

The game then corresponds to the matrix game depicted in Figure 8.2. The incumbent’s best response is to set low K regardless of what the entrant does (i.e., regardless of what the incumbent expects the entrant to do). In other words, low K is a dominant strategy for the incumbent. As for the entrant, the best response is to enter if the incumbent chooses low K and stay out if the incumbent chooses high K. Since low K is a dominant strategy for the incumbent, we
World demand, which had expanded at a whopping 7.7% per year from 1962 to 1972, barely changed from 1972 to 1982. Partly for this reason, partly as a result of Dupont’s strategy, various rival firms abandoned expansion plans or simply scrapped existing capacity. By 1985, five of the firms competing with DuPont in the domestic market had exited: three by acquisition, one by complete cessation of operations, and one by shutting down its US plants.

DuPont never reached the 65% target, but, starting from less than 30% its domestic market share surpassed the 50% mark by the early 1980s. Dupont’s motto may be “miracles of science.” However, its rise to dominance in the TiO$_2$ industry was more of a “miracle of strategy.”

Box 8.1: DuPont and the TiO$_2$ market (cont.)

conclude that (low K, enter) is the game’s NE. In fact, whenever one of the players has a dominant strategy, the easy way to find the NE is to find the other player’s best response to such dominant strategy.

SEQUELAL-MOVE GAMES

Suppose now that the game is played sequentially, with the incumbent moving first and, having observed the incumbent’s move, the small potential competitor deciding whether to enter and challenge the incumbent or rather to stay out. This order of moves seems consistent with the idea that DuPont was a larger firm, had more cash in hand, and was not under the pressure of having to retrofit its capacity (cf Box 8.1). As mentioned in Section 2.2, a game combines a set of players, rules, strategies and payoffs. So far we have considered games where the rule is very simple: both players make their choices simultaneously. This need not be interpreted literally: The relevant point is that each player makes their choice with no knowledge of the other player’s choice.

An alternative rule to simultaneous choice is sequential choice. The best way to analyze a game with sequential moves is to represent it as a tree game. This we do in Figure 8.3. The root node of the tree corresponds to the first choice made in the game, in this case by the incumbent. Each branch coming out of this node corresponds
to possible choices by the incumbent: low K or high K. Each possible choice by the incumbent leads to a node occupied by the second player, the entrant. As before, the entrant may choose to stay out or to enter. We follow the convention that the first number in the end nodes corresponds to the first player’s payoff, whereas the second number corresponds to the second player’s payoff.

Solving a tree game is in some way easier than solving a matrix game. Basically we follow the forward reasoning principle. First we put ourselves in the shoes of the entrant and answer the question, If the entrant were placed in this situation (i.e., on this node) what would he choose? To answer this question we simply compare the entrant’s payoff from entering and from staying out. If the incumbent chooses low K, then the entrant gets 0 if he stays out and 10 if he enters. It is therefore better off by entering. Similarly, if the incumbent chooses high K, then the entrant gets 0 if he stays out and -10 if he enters. He is therefore better off by staying out.

We can now put ourselves in the shoes of the incumbent. Anticipating the entrant’s behavior (enter if and only if the incumbent chooses low K), the incumbent is better off by setting high K. We thus conclude that the NE of the game corresponds to the strategies (high K, stay out) by the incumbent and the entrant, respectively. (For aficionados only: Strictly speaking, the look forward, reason backward procedure leads to what’s known as a subgame perfect NE. The set of subgame perfect NE is a subset of the set of NE.)

The comparison of the simultaneous and sequential move games leads to another important game-theory principle: the value of commitment.
By committing to a course of action a player may achieve a better outcome than by keeping its options open.

This may seem a bit contradictory: If you ask anyone on the street, I would expect most people to say that it’s always better to keep your options open. Not so, a game theorist would say: Committing to a course of action, thus reducing the set of options available, may be a better course of action to the extent that it induces other players to change their course of action. In the capacity game, by choosing high K, the entrant effectively “kills” potential competition. This comes at a cost, namely the cost of investing in capacity beyond what would be efficient (in terms of cost minimization). Ex-post — that is, having observed that the entrant stayed out — the incumbent might regret not having the option to choose low K: after all, a payoff of 40 is better than a payoff of 30. However, such regret would be misplaced: the reason the entrant decided to stay out in the first place was precisely that the incumbent committed to high K.

To summarize the past two subsections, we note that if the number of competitors is small (oligopoly), then there is scope for behavior that creates, maintains or increases positions of market power. This may happen through collusion, whereby oligopolists explicitly (secret cartel) or implicitly (tacit collusion) agree to set higher prices than independent competition would imply. Market power may also result from the behavior of dominant firms that discourages potential competition or kills existing competition. In other words, small numbers (of competitors) tends to lead to market power, and market power tends to lead to higher prices. In the next section, we look into the implications of market power and high prices.

**EFFECTS OF MARKET POWER**

Figure 8.4 illustrates the effects of market power. The top panel depicts the equilibrium of a competitive market, whereas the bottom panel depicts the outcome under market power. An important component of the First Welfare Theorem (cf Chapter 7) is that firms are price takers and set price equal to marginal cost. If, by contrast, firms
are price makers and set price above marginal cost — as we saw in the previous section — then the theorem fails to hold. Specifically, on the bottom panel of Figure 8.4 we see price is set to $p_2$, which is greater than $p_1$, the equilibrium price under competitive conditions. A first implication is that there are trades which fail to take place, that is, there are consumers whose willingness to pay is greater than cost, but no trade takes place because price is greater than marginal cost. All in all, this adds up to area $C$ on the bottom panel.

But there is more: As the bottom panel of Figure 8.4 suggests, and similarly to what we saw in Chapter 7, one of the effects of market power is that it reduces the competitive pressure to be cost efficient. This in turn implies that sellers (or the seller, if there is only one) have higher marginal costs. This is illustrated on the bottom panel of
Figure 8.4 by a higher marginal cost curve, which in turn eliminates an additional area of total surplus under competitive markets, area $D$. In sum, the switch from a competitive equilibrium to one with market power shrinks total surplus from $A + B$ on the top panel to a much smaller $A + B$ on the bottom panel.

In addition to a lower total surplus, we also note that its division into firm profits and consumer surplus becomes considerably more favorable to the seller. All in all, consumers are the greatest victims of market power. As can be seen, consumer surplus under market power (area $A$ on the bottom panel) is considerably lower than under competitive markets (area $A$ on the top panel).

**TRENDS IN MARKET POWER**

One of the most worrying features of the US economy during this century is the significant trend toward greater concentration and greater market power. By greater concentration we mean that in a given industry there are fewer sellers, so that even if we are not in a monopoly situation we are closer to a monopoly situation. How did we get to greater industry concentration? Partly, by means of mergers and acquisitions. For example, from 2008 to 2013, Delta Airlines merged with Northwest Airlines, United Airlines with Continental Airlines, and American Airlines with US Airways, overall lowering the number of major US-based airlines.

In parallel with this general increase in concentration, we also observe significant increases in price/cost ratios $\mu \equiv P/c$ (the price cost ratio $\mu$ is equal to the $1 + k$, where $k$ is the markup indicator introduced in Section 5.3). One study estimates that the average $\mu$ in the US hovered around 1.3 in the 1960s and 1970s but was as high as 1.6 by 2016.

It is hard to establish a causal relationship between the increases in concentration and the increase in prices. It is particularly difficult to do so with aggregate data. Careful industry-level analysis suggests that part of the increase in $\mu$ is due to the emergence of digital firms with high fixed cost and low marginal cost (e.g., Microsoft or Google). However, the secular increase in concentration and in prices is observed even if we restrict to particular industries.

As a specific example, consider housing construction. The top panel of Figure 8.5 documents the increase in concentration in the US
house construction industry: In 2006, six firms typically controlled a local construction market; ten years later, that number was reduced to 4. The bottom panel of Figure 8.5 shows that construction costs of single-family homes increased from less than $150,000 in 1998 to almost $300,000 in 2015. Some of this cost increase may be accounted for by an increase in average house size (which changed a bit in the past ten years), but most of it corresponds to more expensive construction.

A related trend is that, for a given number of competitors, we observe an increasing overlap in ownership. Specifically, many investment funds own shares in multiple competitors within the same industry. Table 8.1 lists the top 10 largest shareholders of the top 4
TABLE 8.1
Largest shareholders of the top 4 US airlines (source)

<table>
<thead>
<tr>
<th>Delta</th>
<th>(%)</th>
<th>Southwest</th>
<th>(%)</th>
<th>American</th>
<th>(%)</th>
<th>United</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlackRock</td>
<td>6.84</td>
<td>Berkshire</td>
<td>7.02</td>
<td>PRIMECAP</td>
<td>8.97</td>
<td>BlackRock</td>
<td>7.11</td>
</tr>
<tr>
<td>Vanguard</td>
<td>6.31</td>
<td>Vanguard</td>
<td>6.21</td>
<td>Berkshire</td>
<td>7.75</td>
<td>Vanguard</td>
<td>6.88</td>
</tr>
<tr>
<td>State Street</td>
<td>4.28</td>
<td>BlackRock</td>
<td>5.96</td>
<td>Vanguard</td>
<td>6.02</td>
<td>PRIMECAP</td>
<td>6.27</td>
</tr>
<tr>
<td>J.P. Morgan</td>
<td>3.79</td>
<td>Fidelity</td>
<td>5.53</td>
<td>BlackRock</td>
<td>5.82</td>
<td>PAR Capital</td>
<td>5.18</td>
</tr>
<tr>
<td>Lansdowne</td>
<td>3.60</td>
<td>State Street</td>
<td>3.76</td>
<td>State Street</td>
<td>3.71</td>
<td>State Street</td>
<td>3.45</td>
</tr>
<tr>
<td>PRIMECAP</td>
<td>2.85</td>
<td>J.P. Morgan</td>
<td>1.31</td>
<td>Fidelity</td>
<td>3.30</td>
<td>J.P. Morgan</td>
<td>3.35</td>
</tr>
<tr>
<td>AllianceBernstein</td>
<td>1.67</td>
<td>T. Rowe Price</td>
<td>1.26</td>
<td>Putnam</td>
<td>1.18</td>
<td>Altimeter</td>
<td>3.26</td>
</tr>
<tr>
<td>Fidelity</td>
<td>1.54</td>
<td>BNY Mellon</td>
<td>1.22</td>
<td>Morgan Stanley</td>
<td>1.17</td>
<td>T. Rowe Price</td>
<td>2.25</td>
</tr>
<tr>
<td>PAR Capital</td>
<td>1.52</td>
<td>Egerton Capital</td>
<td>1.10</td>
<td>Northern Trust</td>
<td>1.02</td>
<td>AQR</td>
<td>2.15</td>
</tr>
</tbody>
</table>

US airlines. It is remarkable how much of the ownership of airlines is concentrated on investment funds. It is also remarkable that several investment funds (those highlighted in color) own shares in all four major airlines. Several other ones own shares on two or three of the top four. Investment funds — and consultants writing on their behalf — claim that they are simple passive investors and that investing in multiple firms within a given industry is a way of reducing risk (and thus offer better results to their clients). However, there is some evidence that when, for some unrelated reason, institutional investors increase their ownership of multiple competitors, market prices increase and firm values increase. Moreover, there is some anecdotal evidence that shareholders have an influence on industry managers in the direction of increasing prices and decreasing output. For example, it has been reported that portfolio managers with ownership in multiple competitors in the oil and gas industry arranged meetings with industry executives with the purpose of pressing the latter to reduce output and thus increase profits (“pump less and profit more”). While there are multiple other examples of this sort, the actual effect of common ownership on market competition remains a contentious issue.
It seems clear that there is a correlation between greater concentration and higher prices. It is difficult to distinguish between correlation and causality. For this purpose, it helps to focus on a particular industry. In this section, we focus on health care, which represents an increasing share of the economy. In the US, health expenditures corresponds to about 17 percent of GDP. This is strikingly high, considering that in comparable European countries health expenditures are only about 10 percent of GDP. Why do Americans spend so much on health? One possible answer is that the quality of US healthcare is higher than in other equally developed countries. However, common outcome measures suggest that the difference in healthcare quality between the US and other developed countries is not that great, especially considering the gap in health expenditures.

There is no one single reason why Americans spend so much on healthcare, but one explanation on which economists generally agree is market power. The hospital industry provides a good example. A series of mergers have taken place since 2000. Many of these mergers do not include direct local competitors, which partly explains why they have been authorized by government agencies. Do they, however, contribute to higher prices? Prices have indeed increased, and recent research suggests that these price increases are probably related to the increased concentration of hospital ownership. From an empirical point of view, we are faced with the common problem of distinguishing correlation from causality. Could the correlation correspond to reverse causality? For example, some hospitals become better, are able to increase prices, and thus become more attractive targets for a merger, which in turn leads to the association of a merger with higher prices.

Typically, hospitals belong to hospital systems, and mergers correspond to bringing together different hospital systems. One possible strategy to avoid the above statistical problem is to focus on smaller hospitals within each system. The idea is that these hospitals, which are not the “crown jewel” of each system, are unlikely to be the motive for a merger between two different systems. We can thus estimate the causal effect on hospital service prices of merging different hospitals. Recent research following this strategy leads to the results summarized in Figure 8.6. The sample of hospitals is di-
US health expenditure corresponds to about 17 percent of GDP, whereas comparable European countries only spend about 10 percent of GDP.

vided into three groups: merged hospitals that are separated by 30 to 90 minutes of travel time, merged hospitals separated by more than 90 minutes of travel time, and non-merged hospitals. Notice that the analysis does not include hospitals which compete in local markets. Given our analysis so far, we would expect no effect on prices: after all, these hospitals are not competing in the same market, that is, are not competing for the same patients. However, the results suggest that merging hospital systems has an effect on prices even when the hospitals do not compete for the same patients. Specifically, the results — which correspond to acute-care hospital mergers over the period 1996-2012 — suggest that mergers “cause” a 10 to 15% price increase. The intuition is that larger hospital systems have greater bargaining power with respect to insurers (which act as intermediaries between hospitals and patients). In other words, what the research suggests is that the effective market may be more than the local market where hospitals compete to attract a certain pool of patients: the effective market may be at the level of hospitals and insurance companies negotiating reimbursement fees.

Hospitals are not the only source of market power in healthcare. Many other segments of the healthcare industry are highly concentrated. A partial list includes

- Pacemakers: 3 firms control 89% of the market
- PET Scanners: 3 firms control 82% of the market
- Medical devices: 4 firms control 77% of the market
- Orthopedic Products: 3 firms control 88% of the market
- Syringes and Needles: 2 firms control 69% of the market
Another important component of the health system is pharmaceutical drugs. Here too we observe significant concentration of market power. Moreover, we observe several instances of dominant firms wielding their power with respect to potential entry threats in ways similar to Dupont in the Titanium Dioxide industry. Specifically, recent research suggest that large pharma companies acquire innovative targets solely to discontinue the target’s innovation projects and preempt future competition, a practice referred to as killer acquisitions.

To recap, it is well known that US patients spend considerably more in healthcare than their European counterparts, even if the differences in quality of service are not very significant. There are surely many reasons for these differences, and market power is likely one of the more important ones.

8.2. ANTITRUST AND COMPETITION POLICY

Housing construction and healthcare — two industries considered in the previous section — are by no means the only sectors where seller concentration is high. Other industries with market power include online retail, where Amazon dominates; ride sharing, where Uber and Lyft dominate; the proprietary seed industry, where Monsanto and Dupont dominate; and online advertising, where Google and
Facebook dominate.

Nor is the problem of market power a new one. The first legislative effort to address the issue of firm dominance, Canada’s Competition Act, was passed in 1889. The US’ Sherman Act followed in 1890. Today, there exist many government agencies in North America, Europe and throughout world, with the single or the main purpose of limiting the negative effects of market power. In this section we summarize some of the dimensions of antitrust policy (sometimes referred to as competition policy).

**PRICE FIXING**

**Explicit collusion**, that is, an agreement between competitors to reduce competition (e.g., increase prices), is illegal. In fact, explicit collusion is a criminal offense in the US as well as in several other countries.

By way of example, consider the class-action suit brought in Canada in 2008 against several major chocolate makers. Lawyers representing chocolate consumers successfully argued that manufacturers conspired to increase prices. From court proceedings, it was revealed that discussions were held at trade shows and association events with the goal of restricting competition and push up consumer prices.

The total settlements from the case amounted to C$23.2 million (Canadian dollars), from C$3.2m paid by Mars to C$9m paid by Nestlé. Mars and Nestlé still face criminal charges from a separate case initiated by Canada’s Competition Bureau (as in the US, price-fixing is a crime in Canada). (On a personal note, and as a matter of full disclosure, I should mention that I indirectly benefited from this settlement: Since it was difficult to find the consumers who were harmed by high chocolate prices, the Canadian Competition Bureau destined the fines to help the development of competition policy scholarship, including your servant’s travel expenses to attend a conference in Canada.)

Unlike Mars and Nestlé, Hershey was left off the hook courtesy of the Canadian leniency program. The leniency program has the goal of incentivizing cartel participants to bring to antitrust authorities information pertaining to secret cartel agreements. Since these
agreements are secret, it would be difficult to find out about them without the help of insiders.

**Implicit collusion** — the situation when firms effectively set higher prices as part of an unspoken agreement — is not per se illegal, that is, it does not explicitly violate the law. In other words, there may be an implicit “agreement” between firms where there is no direct communication. That said, we should add that exactly what constitutes communication is still an open question.

**MERGER POLICY**

In North America, Europe, China, Japan, and pretty much all over the world, if two firms want to merge — especially if they are large and operate in the same industry — they require authorization from the relevant government agency. The idea is that a reduction in the number of competitors may reduce competition and thus harm consumers. **Merger policy** thus constitutes an important tool to address the sources and negative effects of market power.

For example, in 2007 Ryanair attempted to acquire Irish rival Aer Lingus. The European Commission (EC) blocked the acquisition on the grounds that it would imply a high risk of price increases (i.e., high airfares): the airlines overlapped on more than 30 routes from/to Ireland, thus the merger would imply reduced choice for many consumers. This decision was one of the Commission’s first prohibition decisions to be supported by extensive survey and quantitative data analysis to underpin the basic economic argument. Ryanair filed an application for annulment of the decision with the General Court of the European Union. However, in 2010 the Court ruled in the Commission’s favor. Undeterred by this and other defeats, in 2013 Ryanair reformulated its proposal, including a “remedies package” that, it claimed, clearly addressed all of the Commission’s objections. However, the EC again blocked the bid.

Both the US and the EU have specific guidelines regarding mergers. The idea is to balance the possible efficiency gains from a merger (e.g., **merger synergies**) against the threat of greater market power by the merged firms. One important step in this evaluation is the definition of the relevant market (a tricky business) as well as the measurement of the degree of concentration of market shares before and after the proposed merger takes place.
ABUSE OF DOMINANT POSITION

The DuPont example presented in the previous section shows how a firm with market power may cement its dominance by means of a preemptive strategy (capacity expansion in DuPont’s case). Other preemptive strategies include DuaneReade or Starbucks’s strategy of building a high density of stores such that there is little room for an entrant to compete.

Still in the realm of preemptive strategies, the artificial sweetener Nutrasweet provides an interesting example. On the eve of the expiry of Nutrasweet’s patent, Monsanto (the owner) signed long term contracts with PepsiCo and Coca-Cola, its two main customers. Since diet soda is the main use of artificial sweeteners, the long-term contracts effectively made it very difficult for potential competitors to enter.

In addition to Monsanto’s Nutrasweet, another example of contracts as a preemption strategy is given by Intel’s exclusive deals. Specifically, Intel offered its customers (computer manufacturers) a 15% discount with one condition, namely that they only purchase microprocessors from Intel. Understandably, this put AMD, Intel’s main competitor, in a difficult position.

If preemption does not work — i.e., if you cannot keep your rivals away — you can always try to push them out of the market. For example, when Spirit Airlines challenged Northwest Airlines in the Chicago-Detroit market, the incumbent dropped its fares from about $150 to about $50, in the process also increasing the number of flights. After a few weeks of such aggressive competition, Spirit Airlines decided to leave the market, upon which Northwest’s fares returned to $150+ values.

Similar to price-fixing and mergers, the above examples show how a dominant firm may use its muscle to maintain or increase its market power. In this context, the role of public policy is to limit the degree — and possible abuse — of this dominant position, which ultimately hurts consumers.

8.3. THE RISE OF THE TECH GIANTS

Unless you have been living under a rock for several decades, you will have noticed the emergence of several very large firms in the dig-
8.3. THE RISE OF THE TECH GIANTS

ital space, sometimes referred to as tech giants, or superstar firms, or some other term. The acronym GAFA (or GAFAM) has come to be used as a reference for the largest of these giants: Google, Amazon, Facebook, Apple (and possibly Microsoft, in which case we talk about GAFAM). There are many reasons why politicians and common citizens are concerned with the increasing size of the tech giants. One relates to issues of security, privacy and political power. We will return to these in Section 10.2. A second concern relates to market power. The concern is similar to what was discussed in previous sections of this chapter. However, the GAFA case is sufficiently important to warrant separate treatment.

Our story starts in the early 1990s. Before the Windows operating system became widely adopted, most personal computers (PCs) ran some version of the Disk Operating System (DOS), namely MS-DOS, DR DOS and IBM PC DOS. As the market share of DR DOS grew to levels that threatened Microsoft’s dominance, the software giant essentially imposed on computer manufacturers a per-processor-fee contract: instead of paying for each copy of MS-DOS, manufacturers were asked to pay Microsoft a fee for each computer sold (or each processor sold), regardless of whether it was equipped with MS-DOS or with any other operating system (such as DR DOS or IBM PC DOS). A per-processor fee placed computer manufacturers in a difficult spot. Suppose that initially both MS-DOS and DR DOS were priced at $50. Then a computer manufacturer with a mild preference for DR DOS would go with this version of DOS — and it appears there were a good number of such computer buyers. However, under the per-processor deal, the effective price of MS-DOS was zero. Even if the buyer had a preference for DR DOS, the $50 difference was much too high to justify not buying the MS version. Thus it came as no surprise that the market share of DR DOS gradually dwindled until the company went out of business.

In 1993, the US Department of Justice (DoJ) opened an investigation and considered Microsoft’s per-processor-fee deal to be anticompetitive. In 1995, Microsoft was forced to settle by signing a consent decree in which it agreed to stop the practice. Moreover, the software giant agreed not to tie other Microsoft products to the sale of Windows. By the time Microsoft signed the consent degree, its DOS rivals had left, so the agreement had no bearing on Microsoft’s dominance over the operating systems market.
In the meantime, technology was changing rapidly. Microsoft’s plain MS-DOS gradually gave way to the Windows operating system. Moreover, as the Internet expanded, the demand for web browsers increased. Netscape, one of the early entrants, dominated the browser market. Trailing Netscape by a few years, Microsoft introduced its own browser, the Internet Explorer (IE), which was bundled with Windows and thus offered at no extra charge.

In 1998, a major legal case against Microsoft was brought about by various states as well as the Department of Justice. Among other things, the plaintiffs claimed that Microsoft had violated the 1995 consent decree. The decree’s imprecise definition of tie-in become apparent: is IE an added feature of the Windows operating system, or is it a different product that Microsoft (illegally) bundled with Windows? Microsoft argued the former is the case; DOJ argued the latter is the case. Essentially, Microsoft lost the case. However, the final outcome, highly influenced by then recently elected President George Bush, was little more than a slap in the wrist. Several cases followed in the European Union and large fines were paid. However, by and large we might say that Microsoft successfully survived antitrust scrutiny over the past three decades.

The Microsoft case — Microsoft the company and the 1998 antitrust case — provide a good introduction to the high-tech giants for several reasons. First, in some ways it precedes the other giants chronologically. (This is not strictly true regarding Apple, but it is true regarding the “reborn” Apple.) Second, it exemplifies the power of network effects: People buy and use the Windows operating system largely because other people use the same system. Third, the 1998 case exemplifies the type of dominant behavior these giants are capable and prone to: leveraging market power in one segment (e.g., operating systems) to cement dominance in other segments (e.g., internet browsers).

As the Internet expanded and media (music, movies, etc) became decidedly digital (think mp3), other digital giants emerged: Amazon and Google in the 1990s, Facebook in the 2000s. Apple, the oldest of all companies, experienced a “rebirth” of sorts with its launch of the iPod and then the iPhone, so much so that the company became better known for its i-stuff then for the Mac line of products. As the “five knights of the digital apocalypse” grew in size we’ve observed a series of parallels: They command a near-monopoly position in their
own base segment. They leverage this power to largely dominate neighboring business segments. And they improve their technology largely by means of acquisitions.

At some level, there isn’t much new about this type of behavior. Antitrust started in the late 19th century to curb the power of giants such as Standard Oil, whose market power leverage strategies had much in common with what we observe in the 21st century. There are, however, some important differences. First, over the years antitrust has had the goal of getting consumers a good deal, and monopolies were seen as dangerous to the extent that they were able to charge exorbitant prices. But Google and Facebook, for example, don’t charge any user fees, which makes it difficult to find fault with their pricing. Second, while dominant firms of days past were wont to acquire their present and potential rivals, the vast majority of acquisitions by the tech giants correspond to companies you probably never heard about and probably would not have heard about had they not been acquired (Instagram and WhatsApp being two notable exceptions). Third, the way tech giants leverage their power is sometimes very difficult to observe or measure. For example, in order to understand the extent to which Google or Amazon favor their own offerings in their search results one would need to know and understand the way their search algorithms and designed and trained, which would require the knowledge of highly private information. Fourth, and perhaps most important, the key asset that many of these firms own is data, namely user-related data. How much data they have, how such data is used, how value is created from that data — much of this is unknown to regulators and the public alike.
To put it differently, although the basics of competition — or lack thereof — remain the same, the antitrust tools we’ve used for decades are proving difficult to apply to the high-tech giants. Consider the three parts of the previous section. First, collusion (price fixing or related). Clearly, this is not much of a problem with any of the GAFA (or GAFAM). In fact, prices are frequently zero, which makes it difficult to argue that prices are too high. In fact, they are likely high: The argument can be made that users should be compensated for the value that their data provides the platforms. However, it’s difficult to argue there is any violation of the Sherman Act’s prohibition of price fixing or related collusive practices.

Second, consider the application of merger policy. As mentioned earlier, the vast majority of acquisitions by the tech giants correspond to companies you probably never heard about and probably would not have heard about had they not been acquired. By traditional merger policy criteria, these mergers do not create market power. In fact, it might be argued that many of these acquisitions provide a significant spur for innovative efforts by startups, eager to be acquired by one of the giants.

We are left with the third component of antitrust: regulating possible abuses of dominant position. The evidence points to reasonably clear and repeated abuses of dominant position by the tech giants. While there are difficulties in making a case of illegal behavior and proposing an appropriate remedy, it is generally accepted that business as usual is not a sustainable outcome. As of writing this chapter, important efforts are under way both in the US and the EU to re-evaluate the way in which the tech giants can and should be regulated. Stay tuned.
KEY CONCEPTS

monopoly  oligopoly  dominant firm  network effects

game  players  rules  strategies  payoffs

dominant strategy  best response  Nash equilibrium

repeated game  tree game  forward reasoning

killer acquisitions  competition policy  collusion

leniency program  implicit collusion  merger policy

merger synergies  tech giants  superstar firms  GAFA
In Chapter 1 we saw how private property is one of the pillars of the market economy. Witness, for example, the effects of rural reform in China in the late 1970s and early 1980s, the so-called “green revolution”. In Chapter 2 we stressed the social nature of economics: A lot of what we discuss in economics involves transactions between individuals and/or firms, transactions which in turn may have repercussions for multiple third parties.

This chapter continues our survey of reasons why markets may not work as efficiently as discussed in Chapter 7. Specifically, we consider situations when an agent’s decisions imply costs or benefits to third parties — an “externality” — to such an extent that the incentives for individual agents are not aligned with the incentives for the collective of agents.

This may sound a little confusing, but hopefully by the end of the chapter it will make more sense. At this point, suffice it to say that Section 9.2 deals with the most important source of market failure of this type in the current economy: climate change. We all jointly own — or take care of — our planet. Everything I do that harms the planet harms me but also harms billions of other people. In this sense, property rights over the planet are not well defined, or not well enforced, or both. There is a certain parallel between Chinese farms before the 1970s (the fruit of my effort is shared by all) and the world economy in the 21st century (the damage of my CO₂ emissions is shared by all). So we might refer to this chapter’s market failure
as “poorly defined property rights” (to which we add market power, from the previous chapter, and information, from the next chapter).

9.1. INTERNALIZING EXTERNALITIES

This section corresponds to the core of the chapter. In it, we introduce the central concept of externalities. The section begins with an example that motivates the analysis: fisheries in the North Atlantic ocean.

THE TRAGEDY OF THE COMMONS

The Atlantic cod, a highly sought-after fish species, has roamed the coast of Newfoundland for centuries. In the 18th and 19th centuries, the Portuguese were known for line-fishing during long campaigns (up to three months), after which large quantities of dried cod were brought to Europe to feed a large population of poor and protein-needy peasants.

Fishing technologies improved remarkably during the 20th century. Particularly important was the introduction of dragnets, which can catch enormous quantities of cod. A significant number of large ships from Canada, Japan, Russia, etc, flocked to the northwest Atlantic in search of a profitable catch.

Figure 9.1 plots data on catch levels, quota levels, stock levels and the catch rate. Specifically, the vertical bars show the (estimated) stock level of cod. As can be seen, the relatively high levels in the late 1980s (between 800 and 900 thousand tones) dropped to essentially zero in the mid 1990s!

What happened in the early 1990s is a tale of the dangers of poorly regulated fisheries. The quota levels were set at a generous level. The figure shows catch quotas in green and catch levels in blue. The values of quota and catch should be read on the right scale. In the late 1980s and early 1990s, we observe values in the vicinity of 200 thousand tones.

When the stocks began to decline in the early 1990s, catch levels declined as well, but not quickly enough. As a result, a downward spiral took place, whereby lower and lower stocks were quickly deci-
mated. Only when stocks were very low compared to the early 1990s did catch levels drop to close to zero, but by then it was too late.

In the ensuing years, fishing authorities set more rigorous limits on cod fishing, in fact cod fishing was outlawed in Newfoundland for a number of years. Unfortunately, this did not stop several boats from fishing, leading to a failure to rebuild the stocks of cod. Finally, after 2005, with capture levels effectively at zero, we begin to observe a resurgence of the Atlantic cod.

The Newfoundland cod fishing episode is an instance of a more general problem: the so-called **tragedy of the commons**, a term coined by scientist Garrett Hardin. The term “commons”, and the concept it refers to, dates back to an 1883 essay by British economist William Lloyd. He claimed that common land used for grazing (known as a “common”) would inevitably be over-grazed. More generally, the idea is that individual choices, good as they may be from each individual’s point of view, may result in an inferior collective outcome whenever there is a **common resource**.

More generally, the term **externality** refers to the effect that an economic action or transaction might have on a third party that is not part of the action or transaction. So, for example, if I fly across the Atlantic I incur in a certain private cost (e.g., the airfare I need to pay) and I enjoy certain benefits (e.g., visit relatives). However, to the extent that the plane I travel on burns a lot of fuel, my economic
decision implies an increase in CO$_2$ emissions, which in turn implies costs to a lot of other people (the above-mentioned third parties to the economic transaction between me and the airline).

So far we only considered the tragedy of the commons as an instance of an externality. However, there are many other types of externalities, as we will see later in the section. Moreover, we normally think of an externality as a harm to a third party. However, in addition to negative externalities we may also have positive externalities. That said, we will continue the analysis by focusing on the negative externality resulting from common resource use: It’s easier to understand it and it’s quite important (note in particular the problem of climate change, the focus of Section 9.2).

**MARGINAL SOCIAL COST**

By now you are probably aware that the economics modus operandi is to put a number in everything. Suppose there is a negative externality, that is, an action by a specific agent which implies a harm to third parties. For the sake of concreteness, consider the air travel industry, as characterized by Figure 9.2. Let $q$ be the number of passengers flown and $p$ the level of airfares. The demand curve is given by $D$ and the supply curve by $S$. For simplicity, assume that the air travel industry is competitive (Chapter 8 suggests that it is not, at least not in the US, but for now we’ll make that assumption to simplify things).

We may dispute the precise numbers, but we all agree that burning fossil fuels contributes to climate change and related harmful effects. What does this mean in economics terms? Suppose that an additional passenger implies an increase in carbon emissions given by $e$ and let $c$ be the economic harm produced by emissions $e$. The product $e \times c$ then measures the carbon cost of an additional passenger. (In Section 9.2 we will discuss the issue of measuring this cost in practice.) In Figure 9.2, this cost is given by $MEC$, where $MEC$ stands for marginal external cost.

As we saw in Chapter 6, the market supply curve corresponds to the sellers’ marginal cost. We thus have two sources of cost from flying an additional passenger: the (private) marginal cost, which we denote by $MPC$ (or simply by $MC$, if there is no confusion); and the marginal external cost, which is simply the cost imposed on third
parties (in the present case, everyone else on planet Earth). Finally, by adding the private marginal cost and the marginal external cost we obtain the **marginal social cost**, which we denote by $MSC$.

As usual, the market equilibrium is given by the equality of supply and demand. This corresponds to point $(q^\circ, p^\circ)$ on the graph. If there were no externality, this equilibrium would also be an optimal outcome in terms of market efficiency: absent externalities, $(q^\circ, p^\circ)$ maximizes gains from trade, in other words, any value $q < q^\circ$ would be missing positive trades, and any value $q > q^\circ$ would include negative trades (that is, trades such that cost is greater than willingness to pay).

Once we factor in the externality, we notice that, at $q = q^\circ$, marginal social cost is greater than willingness to pay, that is, the $MSC$ curve lies above the demand curve (specifically, the value of $MSC$ is $p'$, which is greater than $p^\circ$). This implies that, from a social point of view, the $q^\circ$th trade should not have taken place. In fact, the social optimal output is given by the point where the demand curve (willingness to pay) equals the $MSC$ curve, that is, point $q^*$. Every trade corresponding to $q > q^*$ is inefficient, as the social cost is greater than the social value. We thus conclude that,

*If there is a negative externality, then the equilibrium output level is greater than the socially optimal output level.*
Finally, since all of the trades from $q^*$ to $q^\circ$ have a social cost greater than willingness to pay, it follows that $L$ measures the efficiency loss due to the over-production of the good subject to negative externalities.

As I mentioned earlier, externalities may be positive. For example, if I get a flu shot then I derive a (private) benefit, namely a lower probability of getting the flu. However, to the extent that the flu is contagious, my vaccine also accrues benefits to third parties.

The possibility of a positive externality is illustrated in Figure 9.3. This time, instead of a marginal external cost, an increase in $q$ (e.g., number of people taking a flu shot) leads to a benefit to third parties, namely the **marginal external benefit**, denoted by $MEB$ in Figure 9.3. Since the (inverse) demand curve measures the willingness to pay (private value), the total social value is given by the buyer’s willingness to pay plus the marginal external benefit. This sum corresponds to the **marginal social benefit** and is denoted by $MSB$ in Figure 9.3.

As before, the market equilibrium is given by point $(q^\circ, p^\circ)$. The social optimum, in turn, is given by the equality of marginal cost (supply curve, $S$) and marginal social benefit ($MSB$). This corresponds to point $(q^*, p^*)$. Unlike the case of a negative externality, where $q^* < q^\circ$, we now see that,
Box 9.1: External benefits from flu vaccination

Vaccination represents a typical example of a positive externality. However, it is not easy to estimate the magnitude of such externality, and such estimation is seldom done. One possible identification strategy is to compare different parts of the US with different rates of vaccination. However, this approach suffers from the common correlation-is-not-causality problem (that is, the issues of omitted variable bias, reverse causality and spurious correlation). For example, it could be that people in some states are very healthy and health conscious. This implies a cross-state correlation between the fraction of people who take a flu shot and how healthy people are, but such correlation would not necessarily correspond to a causal relation, the relation which would correspond to an externality.

A more promising identification strategy is to take into account the fact that there are many (hundreds?) of flu strains. As much as we try to predict the pesky virus, each winter brings a new slate and a few surprises. For the statistician, this provides the opportunity of an exogenous shock to the actual and effective degree to which people are vaccinated. For example, if many people took a flu shot but it turned out the relevant strains were strains not included in this year’s vaccine, then for all practical purposes it’s as if people had not been vaccinated.

Recent research based on this approach shows that the effects of flu vaccination on others are significant. In a given year, a 1% increase in the US flu vaccination rate saves about 795 lives, especially among individuals aged 75 and older. If we try to put a monetary value on this, it comes to a benefit of about $63 per vaccination. The 1% increase in the vaccination rate also saves 14.5 million work hours, which has the economic value of about $87 per vaccination.

For reference, a flu shot costs between $20 to $40.

If there is a positive externality, then the equilibrium output level is lower than the socially optimal output level.

Finally, similarly to the case of a negative externality, all of the trades from \( q^\circ \) to \( q^* \) have a willingness to pay greater than social cost. It fol-
PIGOU TAX

In markets with externalities, the First Welfare Theorem fails to hold: The market price is no longer the right guide for consumers and producers. In particular, if the externality is negative then the equilibrium output level is too high from an efficiency point of view. The “invisible hand” of the marketplace no longer has the “magic” power that it has absent externalities.

Is there any hope for the market? Enter economist Arthur C. Pigou, whose answer is yes, there is hope for the market system even when there are externalities, so long as we impose an output tax (if the externality is negative) or an output subsidy (if the externality is positive). Figure 9.4 shows how this is done. Basically, we choose a value of $t$ such that the new supply curve, given by $S^t$, crosses the demand curve at $q^*$. In other words, if we choose the right value of $t$, then the regulated market equilibrium (regulated by the Pigou tax $t$) is the same as in the socially optimal solution!

In other words, with Pigouvian taxation the “magic” of the price system is re-established, though with one caveat: Under the First Welfare Theorem, there is no information requirement for policy
9.1. INTERNALIZING EXTERNALITIES

makers: All they need to do is to let the market work its ways and expect an optimal $q$ and $p$ to emerge (as usual, by “optimal” we mean “efficiency maximizing”). By contrast, a Pigou-regulated market requires that we determine the value of $t$. As we will see in Section 9.2, this task is far from trivial.

A relative success story in applying Pigou’s ideas is taxation of gasoline consumption in Europe, where a Pigou tax has moved consumption levels closer to the social optimum. The numbers in Box 9.2 suggest that a tax on gasoline price may be an effective one.

OTHER SOLUTIONS

Once you start thinking about the concept of externality, you realize that the world is full of them. You also realize that there are different ways of dealing with it, some related to Pigou taxation, some different. This section surveys various instances of market failure due to externalities. I also recommend you look at the corresponding chapter in the exercise book, where additional real-world examples are considered.

**Congestion.** One externality that may not be obvious at first, but turns out to be quite important, is congestion. Actually, it’s an externality many of us have to deal with on a daily basis. If you fly a lot out of a busy airport like New York’s LaGuardia or London’s Heathrow, you know the drill: “Good morning folks, this is your captain speaking, we are now number 17 for takeoff.” Which means an extra half hour taxiing on the tarmac. Excessive congestion results from an externality: when an airline decides to schedule a flight during rush hour, it does not take into account the extra delay it imposes on all flights departing right after.

Car congestion in city centers can also be very annoying. Some cities have come up with the creative solution of charging cars to drive in the city center (London, Hong-Kong, Singapore). New York tried it, but the proposal was killed. Road pricing, as the policy is known, is effectively a form of Pigou taxation.

**Free riding.** Where can you find beer for 5 cents a bottle? My best answer: try a class reunion with 100 diners. Let me explain: When you and your 99 colleagues go out for a meal, chances are you will
In 2013, 1 liter of gasoline cost €1.53 at an average European Union (EU) pump, of which 58% corresponded to taxes and duties. In the United States, average price at the pump was $3.49 per gallon. Considering that one gallon equals 3.785 liters and that the yearly average exchange rate was 0.783 euros per dollar, this comes to about €0.72 per liter, about one half of the price in Europe. For US readers, the European price was about $7.40 per gallon, about twice the US price.

Not surprisingly, gasoline consumption (both per person and per car) is much lower in the EU than in the US. Arguably, the value of \( q \) is much closer to \( q^\ast \) in the EU than in the US (see Figure 9.4). This is partly because the price elasticity of the demand for gasoline is different from zero, but also because the demand for cars is very different in the EU than in the US.

Reacting to high gasoline prices and the demand for low-consumption cars, European cars are smaller and consume less gasoline than their American counterparts. In 2013, average fuel efficiency in the US was 32 miles per gallon, whereas the EU showed a whopping 45. For European readers, this corresponds to 5.2 liters per 100 kilometers (EU) against 7.6 liters per 100 kilometers (US).

Various US governments have made efforts to increase automobile fuel efficiency, for example enacting Corporate Average Fuel Efficiency (CAFE) standards, regulations that force automakers a minimum average fuel efficiency. However, the regulated firms have managed to keep the minimum standards at “reasonable” levels. Moreover, various loopholes make CAFE standards less effective than one might think: for example, gas-guzzling SUVs are treated as light trucks and thus excluded from the average, which effectively provides an “escape” for consumers and manufacturers who want to avoid regulation.
purely theoretical consideration, economists have actually estimated this effect: even for a party as small as 4, splitting the bill may lead to an increase in the total tab by as much as 40%.

In economics, this problem is referred to as the free-rider problem. The point is that an agent’s decisions (for example ordering an extra beer at a split-the-bill dinner) does not take into account the costs imposed on other agents (99% of that extra beer is effectively paid by them). As a result, the total quantity consumed at the dinner is likely to be higher than socially efficient: even if the marginal cost of a beer is, say, 30 cents, there will be a lot of beer consumed for which willingness to pay is lower than social cost.

If you think about it, the tragedy of the commons is really a free-riding problem. In fact, the tragedy of commons may be characterized by a concentration of benefits and a dilution of costs. What I mean is that if I fish in a fishery that is close to maximum extraction rate, then the benefit from the catch is concentrated (I get the fish), whereas the cost (increased risk of a total collapse of the fish stocks) is shared by everyone.

Let’s go back to the class reunion example. In addition to being a fun example, it also suggests that Pigou taxes may not be feasible, or indeed necessary, to solve the market failure caused by externalities. Suppose that one of your classmates comes to the reunion and insists in always ordering the most expensive items on the menu: Beluga caviar, Dom Perignon champagne, etc. He or she may get a lot for little money on this occasion, but chances are they will not be invited to the next reunion. In other words, our society includes a series of mechanisms and institutions that prevent people from constantly free riding. It’s not a market solution, but it works.

Community organization. The above point regarding social mechanisms and institutions warrants further discussion. In much of economics there is a tendency to contrast two extremes: the government, i.e., the case of centralized decision-making; and the market, where it’s each one for him or herself. This is clearly an exaggeration: There are multiple cases when local or regional organizations do a very good job at effectively preventing the free-riding problems of unregulated market competition.

More formally, the idea — attributed to economist Ronald Coase — is that, if negotiations costs are not very high, then we should ob-
serve an agreement between the relevant parties which effectively induces an efficient outcome. To put it differently, if direct negotiations between the relevant parties is relatively costless then the externalities problem is really not a problem!

In fact, research by economist Elinor Ostrom on the actual workings of commonly held resources (water, fisheries, pastures, etc), suggests that it’s much less of a tragedy than the “tragedy of the commons” would suggest. If communities are sufficiently small, then experience shows that they will find rules for a fairly efficient use of resources. In other words, there’s a lot going on between the extremes of a centralized government and the unregulated, decentralized market.

PUBLIC GOODS

Similarly to other professions, economists like to have their jargon. When it comes to characterizing goods and services, we make two important distinctions. First, goods can be rival goods or non-rival goods. An apple is a rival good: if I eat one you cannot eat it (that is, you cannot eat the same apple). Beethoven’s string quartets are a non-rival good: the fact that I listen to them on my earphones does not detract from your ability to do so on your own mp3 player.

A second relevant characteristic is whether consumers can or cannot be excluded from consumption. For example, if I issue a music EP with special copy-protection provisions, I can effectively exclude people from listening to my song (for example, I can choose to only allow paying fans to download my song). By contrast, if I go to Central Park’s Strawberry Fields on a Sunday afternoon and start playing my song, then I cannot exclude passersby from listening to it. In the former case, I say the good is excludable, whereas in the latter case the good is non-excludable.

With these two dimensions in mind, I can now create a 2x2 matrix (see Table 9.1) listing four types of goods. Earlier I mentioned that this chapter is about poorly defined property rights as a cause for market failure. In terms of Table 9.1, this corresponds to non-excludable goods, that is, goods for which I cannot prevent free riding. We’ve seen how, in this setting, an unregulated market may lead to an inefficient outcome. So far, most of the examples we considered correspond to the top right cell, common-pool resources. We
### 9.1. INTERNALIZING EXTERNALITIES

<table>
<thead>
<tr>
<th>Excludable</th>
<th>Non-excludable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rivalrous</strong></td>
<td><strong>Non-rivalrous</strong></td>
</tr>
<tr>
<td><strong>Private goods</strong>&lt;br&gt;food, clothing, cars, parking spaces</td>
<td><strong>Common-pool resources</strong>&lt;br&gt;fish stocks, timber, coal</td>
</tr>
<tr>
<td><strong>Club goods</strong>&lt;br&gt;cinemas, private parks, satellite television</td>
<td><strong>Public goods</strong>&lt;br&gt;free-to-air television, air, national defense</td>
</tr>
</tbody>
</table>

**TABLE 9.1**
Types of goods (source: Wikipedia)

Now consider the case on the bottom-right cell, **public goods**, that is, goods which are non-rivalrous and non-excludable. In addition to the list in Table 9.1, examples of public goods include neighborhood security, public health, public parks, art.

Public goods provide a particular challenge to the market economy. For the reasons considered earlier, the market equilibrium leads to under-provision of public goods. There are several possible solutions to this underprovision problem. One is direct provision by the government. For example, the US federal budget includes funds for the National Endowment for the Arts as well as the Public Broadcasting System (covering about 50% of the system’s budget). Another solution is private provision. Going back to the point made on p. 313, there’s a lot going on between the extremes of a centralized government and the unregulated, decentralized market. For example, New York’s Central Park Conservancy is a private, non-profit organization which takes care of the maintenance of the city’s most cherished (and used) park. The Conservancy receives no federal, state or city funds. It does receive multiple small and large donations, as well as many hours of volunteer work. Finally, another source of private provision of public goods is given by philanthropy. For example, the Gates Foundation has been instrumental in the development of a malaria vaccine, a clear case of a public good that has been under-provided for decades. We’ll return to the important issue of philanthropy in Section 12.3.

To conclude this section, we discuss a few cases of public goods in the corporate world. One interesting (and controversial) case is that of advertising commodities. This is a clear case of a public good...
It has been estimated that each dollar spent on advertising agricultural products like eggs, milk, beef, prunes or almonds yields $3 to $6 of additional revenue to producers. Not a bad return on investment. For example, the 1980s “California Raisins” campaign was credited with increasing sales by 10 percent. Before the ad campaign, raisins were “at best dull and boring,” states the California Raisin Board. After the campaign, people were no longer “ashamed to eat raisins.”

Profitable as they are, Commodity Promotion Programs — as these campaigns are called — are difficult to implement. Given that some producers are paying for such a campaign, other producers may have a strong incentive not to pay for it: there is nothing better than reaping the benefits without paying the costs.

In order to solve this free-riding problem, some of the programs are mandatory, that is, growers vote whether to start a marketing program; and if the vote succeeds, all growers are required to participate.

But mandatory programs create their own problems. Since the early 1990s, a series of U.S. producers have sued their respective boards claiming they cannot be forced to participate in such deals. Initially, the Courts ruled largely in the plaintiffs’ favor. But in 2005 the US Supreme Court ruled 6-to-3 that beef marketing programs did not violate the First Amendment (rights of free speech and association).

where the free-riding problems can be significant. As one of several thousand peanut producers in the US, I’m quite happy if an advertising campaign on the New York subway encourages consumers to eat more peanuts. I am particularly happy if the campaign is paid by someone else. As Box 9.3 shows, the potential gains from these campaigns are enormous, but you can see how the free-riding incentives are equally enormous. Trade associations may play an important role here, but this may imply difficult legal problems.

Another public-goods problem in the corporate world is given by corporate reputation. Take a franchise like McDonald’s. More than 80% of the restaurants are owned by franchisees, not by McDonald’s itself. This can be a source of great tension: If a customer has a bad
experience in my store on Broadway, they may decide never to return to McDonald’s, not just to my McDonald’s on Broadway. In this case, the solution to the free-riding problem is to have a very strict set of franchisee rules that all restaurants must follow (e.g., clean restrooms). Once again, there are many institutions between the extremes of a centralized government and a decentralized market.

Collective reputations go beyond corporations such as McDonald’s. Particularly important is the problem of country reputation. Each time a Chinese firm exports electronics to Europe, or a Chilean winery exports a carménère to Australia, or a Mexican farmer sells avocados in the US, the consumer experience affects their perception of Chinese electronics or Chilean wine or Mexican avocados, respectively. And this affects all of the exporters from that country, not just the firm that exported a particular product. Externalities never end!

9.2. CLIMATE CHANGE

Climate change is the ultimate public good problem faced by the modern economy. We all jointly own — or take care of — our planet. Everything we do to it will be, in one way or another, shared by all, not only all across the planet but also all across generations.

As mentioned in Chapter 1, there is a general consensus that human intervention — economic activity — is a central cause of climate change. Since about 1850 human activities have been releasing extra greenhouse gases (especially CO₂) into the air. The higher concentration of greenhouse gases in the atmosphere has slowly propelled a rise in average temperatures across the globe. Overall, the 2017 global average was 0.9 degree C (1.6 degrees F) higher than it had been between 1951 and 1980. There is also a general consensus that this change has implied — and will likely continue to imply — significant changes in climate patterns.

The economist’s perspective on the problem of climate change is not very different than that of other public goods or public bads: The equilibrium of an unregulated market economy leads to the under-provision of public goods and the over-provision of public bads. Left to its own devices, the unregulated world economy will produce too much CO₂.
9.2. CLIMATE CHANGE

EFFICIENCY AND EQUITY

Much of economics is about balancing equity and efficiency considerations, and climate change is no exception to the rule. One important consideration regarding efficiency is that the relation between economic activity and CO$_2$ emissions is not uniform across the world. Figure 9.5 shows the relation between level of economic activity and level of CO$_2$ emissions for a sample of large countries (GDP greater than 100 billion dollars). Given the large disparities in size across countries, the GDP (horizontal axis) is plotted on a logarithmic scale.

The blue line in Figure 9.5 depicts the estimated average relation between economic activity and CO$_2$ emissions. Specifically, let $\gamma$ be the coefficient indicating the level of CO$_2$ emissions per dollar of economic activity. On average and in 2019 this coefficient comes to about .3753 metric tons per $1,000 (or .3753 kilograms per dollar). However, this average hides a big disparity across countries. For example, the value of $\gamma$ in the European Union is .2694, whereas the value of $\gamma$ in China is a whopping 2.2964. We should also add that these are average values for each country. As we saw in Chapter 5, marginal costs are typically increasing, which implies that the marginal cost (in terms of CO$_2$ emissions) is likely higher than the average cost measured by $\gamma$.

The large cross-country disparities imply an important point: If we want to reduce the level of emissions at the lowest economic cost
possible, then it might be better to do so by reducing the level of economic activity in China rather than in Germany. To understand why this is what economic efficiency dictates, consider Figure 9.6, which depicts the cost, in terms of CO$_2$ emissions, of a given level of (carbon emitting) economic activity. Notice that the cost curve is convex. This reflects the fact that, at lower levels of economic activity, we can choose the ones that are the “cleanest”. If we insist on reaching yet higher levels, then we gradually need to have recourse to “dirtier” technologies. This is similar to the idea of decreasing marginal product introduced in Chapter 3.

Suppose that Germany is initially at point $g_0$ and China at $c_0$, that is, China has a higher level of carbon-emitting activities. Now consider an increase in China’s activity level from $c_0$ to $c_1$ and a decrease in Germany’s activity level from $g_0$ to $g_1$. In terms of economic value created, China’s increase is equivalent to Germany’s decrease. However, the increase in emissions created by China with the shift from $c_0$ to $c_1$ (the difference $\Delta c$) is considerably greater than the decrease in emissions achieved by Germany with the shift from $g_0$ to $g_1$ (the difference $\Delta g$). If the goal is to reduce the level of CO$_2$ emission on a per-dollar of economic activity basis, then we are going in the wrong direction: It’s China that should be reducing its emissions, not Germany. However, recently the German government announced plans to spend €40 billion (about $44 billion) over four years to help the
country cut its CO₂ emissions. The move is expected to reduce the global rise in temperature by 0.00018°C in a hundred years, a considerably small gain for such a large cost. In the meantime, China continues to build about one coal-fired power plant a week!

This leads to the equity issue. Although it is more efficient to reduce emissions in China, the argument can be made that, for decades, Germany has experienced considerable economic growth largely based on CO₂-emitting activities. Why cannot China now enjoy the same type of economic growth? The argument is even more cogent when we consider CO₂-creating activities in developing countries that have still not reached China’s level of economic development. We must also recall that Figure 9.5 measures the distribution of production across the world, not the distribution of consumption. Much of the CO₂-emitting production in China corresponds to consumption elsewhere, including in particular the US. Were we to draw a similar graph with consumption instead of production, US consumers would likely come out as prominent contributors to CO₂ emissions. Finally, to further compound the cross-country equity dilemma, we must recall the simple fact that the countries that stand to lose the most from climate change — in terms of rising sea levels, forced migration, etc — are typically developing countries.

As important as cross-country differences may be, the big elephant in the room, when it comes to major asymmetries, is the differential impact on current and future generations. Future generations have the most to gain or lose from today’s climate policy. However, future generations do not have a voice or a vote (except possibly for the very next generation). Repeatedly, we hear individuals, governments, and other organizations pledge to be good stewards and hand on a well-kept planet to the next generations. But do we really give future generations the same weight as ours? We will return to this issue below.

THE MARGINAL SOCIAL COST OF CO₂

Climate change poses a particularly difficult public bads problem because of the sheer severity of the problem as well as its truly global nature. That said, the solution proposed by most economists is not that different from the solution proposed to most other externality problems: Most economists argue for a Pigou tax, or, to use the com-
mon term in the present context, a carbon tax. Implementing such a tax requires an estimate of the marginal social cost of the externality in question. We now turn to this critical problem in the context of climate change.

In climate change lingo, the marginal social cost of CO$_2$ emissions is normally referred to as the **social cost of carbon** (SCC). Typically, it is expressed as the dollar value of the total damages from emitting one ton of carbon dioxide into the atmosphere. Over the past decades, economists and scientists have developed models to generate such estimates. These are typically dynamic models of the world economy based on equations which relate the level of economic activity to the rate of CO$_2$ emissions, as well as equations which describe the cost of coping with varying climate conditions. The models then estimate how a change in the level of CO$_2$ emissions in year $t$ reflects on the overall evolution of the world economy. Adding up all of the costs across time and space finally leads to an estimate of SCC, which basically corresponds to the marginal social cost introduced in Section 9.1.

The reports issued by the **Intergovernmental Panel on Climate Change** (IPCC), the world’s authority on climate change issues, typically include a variety of scenarios: there is no certainty in climate change science. Similarly, there is no certainty in economic modeling of the future impacts of climate change. The best that current economic research can do is to consider different values for key parameter values and estimate the sensitivity of results to changes in these parameters. In addition to parameter values, we must also account for differences in how to model climate and the economy.

Overall, this process leads to a significant degree of variation in the estimate of SCC. Figure 9.7 provides estimates based on three different studies, codenamed DICE, FUND, and WITCH. The process underlying this figure consists of running a given model with multiple combinations of parameter values and deriving the distribution of resulting values of SCC (a process known as Monte Carlo simulation). In other words, each model does not produce one value of SCC, rather a distribution of possible values. One way of representing these results is to plot the **cumulative distribution function**. This is a function stating, for each value $x$ of SCC, the probability that the true value of SCC is lower than $x$. For example, the bullet point at (28,75) signifies that, according to the DICE model, there is a 75%
The DICE model is the oldest and, arguably, the best known economic model of climate change. As the figure shows, the model’s prediction of the value of SCC is far from precise. The average is given by $21.6 and the standard deviation by $13.8. Although the graph only shows values of SCC from 0 to $100, the model admits the possibility of SCC being greater than $100, though with small probability.

In addition to variation of estimates by a given model, we also notice considerable variation from model to model. This results from different modeling assumptions, e.g. assumptions regarding the evolution of technology (i.e., CO\textsubscript{2} emissions per dollar of economic activity).

Finally, as mentioned earlier, one big issue regarding the impact of climate change is the relative weight given to current and to future generations. In terms of estimating SCC, this is reflected in the discount rate used in calculations. Suppose that I give you the option of receiving a dollar today or $1 + r$ dollars a year from today. Different people might ask for different values of $r$. For example, if I say “I will give up a dollar today if you give me a dollar and 10 cents a year from now,” then you would say my discount rate is 10%.

Now, suppose that a certain climate policy leads to an estimated cost of 1 billion dollars today and an estimated benefit of 2 billion
dollars fifty years from now. Is it worthwhile? Well, that depends on the relative weight we give to our descendants half a century from now. If a dollar now is the same as a dollar in fifty years, then the net benefit is positive and we should go ahead with the policy. However, if the weight given to people fifty years from now is less than 50% of the weight given to today’s costs and benefits, then the answer is negative, that is, it is not worth incurring a cost of 1 billion dollars now.

Unlike other model parameters, which reflect matters of science or economics, the discount rate is largely an ethical choice: how much weight should we give to future generations? This matters a lot because the value of SCC is highly sensitive to changes in the discount rate. For example, the value of SCC in the US ranges from $10 at a 5% discount rate to $50 at 2.5% discount rate. For reference purposes, in a major survey of 197 economists, the average long-term discount rate was 2.25%. This would put us closer to the $50 figure than to the $10 figure. However, there are those who think that a 3% rate into the distant future is unethical. For example, it implies that the weight given to costs and benefits 100 years from now is 1/20th of the weight given to current costs and benefits. This does not seem entirely fair, some might say. If we choose a lower — and declining — discount rate, then the estimate of the current cost of CO$_2$ can be as high as $400. The Stern Review, considered by many as one of the more credible economic analysis of climate change, uses a discount rate of 1.8% and estimates a SCC of $85 per metric ton. By contrast, the Heritage Foundation, a conservative thinktank, calls for a 7% rate, which would bring the SCC to a one-digit dollar value.

**UNKNOWN UNKNOWNS**

The economics approach to measuring the cost of climate change, and in particular the cost of CO$_2$ emissions, is subject to a number of criticisms. In addition to the problem of discounting, one of the greatest limitations of models such as DICE is that they do not account for so-called **unknow unknowns**. At a famous February 12, 2002 news briefing, US Secretary of Defense Donald Rumsfeld mentioned that

As we know, there are known knowns; there are things we know we know. We also know there are known un-
knowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don’t know we don’t know.

Rumsfeld was referring to the US involvement in Iraq, but the idea applies to climate change as well. The idea is that the economic/climate science models, as accurate as they may be with respect to what is known, may miss out relevant implications of climate change.

A related criticism, coming from climatologists and economists alike, is that there are tipping elements in the earth system. For example, some scientists have argued that global warming beyond 2°C might lead to an irreversible melting of the Greenland ice sheet. This and other tipping-like possibilities imply that climate change may lead to very large effects — indeed, catastrophic effects — even if these occur with small probability. The argument is then that, given the nature of the climate change “lottery”, one cannot safely conduct the cost-benefit analysis that economists are used to.

One possible response to this criticism is that risk and uncertainty are an integral part of life, and we must make decisions based on the possibility (better still, the probability) of a catastrophic outcome. As a brief but relevant digression, consider another risk which we are constantly subject to: being hit by an asteroid. This is not a theoretical possibility: A meteor exploded over the Bering Sea between Russia and Alaska at the end of 2018. It happened with no warning whatsoever. Had it hit a populated area, the effects would have been catastrophic: This particular asteroid was relatively small, but it had the energy of about 10 Hiroshima-like atomic bombs.

In any given year, the chance of a decent-sized asteroid hitting the earth is somewhere around 1 in 500. The probability that humans are hit by an asteroid is even smaller. The effect, however, is as high as the probability is low. It turns out that a catastrophe of this nature is largely avoidable: We have the technology to detect and deflect these asteroids.

It would cost us about $500 million to create the appropriate asteroid monitoring system. The B612 Foundation, led by former astronaut Ed Lu, has not been able to raise more than a small fraction of the necessary funds. This suggests that we are able to live with the risk of being hit by an asteroid. Clearly, to have an asteroid wipe out the entire Northeast of the United States is not the same as large parts
9.2. CLIMATE CHANGE

of planet Earth becoming unlivable (the essence of the climate change threat). However, the point is that balancing costs and benefits in a world of risk and uncertainty is part of life. And in this context models such as DICE play an important role. (On this matter, recently deceased economist Martin Weitzman represented an important dissenting opinion.)

CLIMATE STRATEGY

There is a consensus for the need to reduce the density of greenhouse gases in the atmosphere. What is not as clear is how to best achieve this goal. The following equation, depicting the dynamics of greenhouse gases, helps address the issue of climate strategy:

\[ C_1 = C_0 + E \times \gamma - R \]  

(9.1)

where \( C_0 \) and \( C_1 \) denotes the level of carbon emissions (tons) currently and in the future (respectively), \( E \) the level of economic activity (dollars), and \( R \) the level of direct carbon reduction efforts. Finally, \( \gamma \) is the coefficient introduced on p. 317, that is, the coefficient indicating the level of CO\(_2\) emissions per dollar of economic activity.

The above equation is helpful because it indicates three types of policies aimed at achieving a lower value of \( C_1 \), starting from a given \( C_0 \): reducing the value of \( E \), reducing the value of \( \gamma \), or increasing the value of \( R \). We next consider each of these.

The argument for the first strategy is that economic activity has been responsible for the increase in atmospheric greenhouse gases. Therefore, we should reduce the level of economic activity so as to counteract the negative effects of the past two centuries. For example, many climate activists argue that we should urgently reduce transportation services based on fossil fuels (air travel, car travel, ocean shipping, etc).

A very different approach is to reduce the value of \( \gamma \), that is, to find forms of economic activity that are more efficient in terms of CO\(_2\) emissions. In the field of transportation, significant progress has been achieved in the past decades (transportation represents an important slice of total CO\(_2\) emissions). We already mentioned the case of fuel efficiency in the auto industry. Impressive as the gains have been — especially among European car manufacturers — the
improvements in air travel have been even more impressive. In 2019, for the first time ever the fuel cost of air travel, on a per-mile basis, is lower than that of car travel. As to ocean shipping, Mersk, one of the world’s largest shipping companies, has been working on switching to hydrogen-fueled ships, a switch that would imply a considerable lower value of $\gamma$.

Another area of great promise in terms of lowering the overall value of $\gamma$ is renewable energy, including solar and wind. Many skeptics of the promise of renewable energy have been surprised by the very rapid drop in production costs. For example, China is currently producing solar panels at a small fraction of the cost of one or two decades ago. Notwithstanding these great achievements, renewable energy production is still subject to the so-called **intermittency problem**, the fact that one cannot control weather at will: it’s windy when it’s windy and the sun shines when it does. Since electrical power cannot be stored in a cost efficient way — at least not yet — renewable energy sources suffer from the limitation that the moment of production may not coincide with the moment of demand.

Finally, going back to Equation 9.1, we have the term $R$, efforts to reduce the level of CO$_2$ on the atmosphere. It may be said that technology had a lot to do with the increase in carbon emissions since about 1850. To a great extent, technology is also one of our best bets to reduce the level of carbon emissions going forward. In other words, technology is both part of the problem and part of the solution. Among others, atmospheric CO$_2$ reduction may be achieved by

- reforestation
• capturing and sequestering CO₂ from point sources
• direct capture of CO₂ from air

The feasibility and efficiency of each alternative is still open to debate (see one view and a different view). As of 2019, the most efficient path, on a carbon per dollar basis, is reforestation. However, this may well change in the future. In fact, this is one of the points that economists repeatedly make: The advantage of the price system is precisely to provide the incentives and the information to guide investors in the right path. If we are able to put the right price on carbon, then (a) investors will have incentives to find ways to reduce the level of carbon from the atmosphere; and (b) the price level will lead investors to choose each alternative to the extent that it is efficient.

UNINTENDED CONSEQUENCES

The economy is a large “network” of interconnected markets. What may seem a good idea in isolation (e.g., in a given market) may lead to negative effects in other parts of the economy (e.g., other markets) which more than outweigh the positive effects. Climate change and climate-related policies provide many instances of such unintended consequences.

One example of this is the growing movement known as “flight shame”, which has been popularized by well-meaning climate activists and is gaining momentum around the world. Its premise is that flying is bad for the climate (it contributes 2.5% to global carbon emissions). Therefore, if you care about the planet you should simply avoid air travel and encourage others not to fly and perhaps simply ban flights.

But would stalling air travel really save the planet? The problem is that the tourism industry, in particular eco-tourism, depends heavily on air travel. In some destination countries, nature-based tourism is a top foreign exchange earner, and tourist inflow is one of the greatest incentives for reforestation. Finally, by most estimates the positive CO₂ impact of reforestation dwarfs the negative CO₂ impact of air travel. In sum, fly shaming may actually lead to the opposite effect of what’s initially intended to have.

Another example is given by what economists call leakage, when partial regulation of a product results in increased consumption of
unregulated goods. In 2015 Portugal banned plastic bags from supermarkets. This was very effective in reducing one type of plastic consumption. But households need plastic in order to dispose of some types of garbage. Since supermarket plastic bags were recycled as garbage bags, the ban implied that households started to purchase garbage bags (which moreover were imported). The trade-offs are complicated: Supermarket bags are thin and run the risk of getting out of the waste system and into the oceans. Garbage bags, on the other hand, are heavier and require more energy and CO$_2$ emissions to produce. The bottom line is that a positive effect (lower consumption of one type of bags) was counteracted — and possibly outweighed — by a negative effect.

THE POLITICS OF CLIMATE CHANGE

So far we have focused on the role of scientists and economists in the climate debate. However, these are by no means the only players in the game. Scientists and economists do not make decisions; politicians and voters do. The problem with politicians is that they are subject to all sorts of influences, in a way that is not necessarily representative of the general consensus. For example, the Trump administration rolled back a number of climate policy measures of the Obama administration. One important piece of this reversal is the Trump administration’s estimate that the current value of SSC is around $1 to $7. This is a considerable drop from the value of about $50 that had been set by the Obama administration, which in turn was closer to the mean estimate from economic models.

The problem with voters is manifold. First, voters are not always given the right explanation regarding carbon taxes. The very word “tax” is a loaded term: it suggests that voters are poorer by the amount of the tax, which is not true considering all of the effects ensuing from correcting an externality. In this regard, one success story is given by British Columbia’s (BC) pioneering carbon tax, first enacted in 2008, where the tax hike was clearly linked to a cut in income tax rates. Analysts argue that the system has been effective in reducing fuel use, with no apparent adverse impact on the province’s economy. Specifically, since 2008 BC’s fuel consumption has fallen by 17.4% per capita (while in the rest of Canada it slightly increased). BC’s GDP, in turn, kept pace with the rest of Canada’s
over the same time period. Moreover, carbon tax revenues enabled BC to set Canada’s lowest income tax rates. Overall, taxpayers may even have benefitted in net terms.

In the US state of Washington, however, Initiative 732 (I-732), largely modeled after the BC tax, was rejected 59.3% to 40.7% when it appeared on the November 2016 ballot. One reason for this defeat might be the free-riding problem alluded to earlier in the chapter: why should Washington state residents pay a carbon tax if all other 49 US states do not pay a carbon tax?

Finally, we have climate activists, who, by any reasonable impact measure, play a crucial role in the process. Differently from scientists and economists, activists tend to view the problems and the solutions more radically. For example, US Representative Ocasio-Cortez went on record stating that “the world is gonna end in 12 years if we don’t address climate change.” Greta Thunberg, arguably the leading activist of our time, addressed the World Economic Forum with an emphatic statement:

Adults keep saying: “We owe it to the young people to give them hope.” But I don’t want your hope. I don’t want you to be hopeful. I want you to panic. I want you to feel the fear I feel every day. And then I want you to act. I want you to act as you would in a crisis. I want you to act as if our house is on fire. Because it is.

Economists’ typical reaction to this stance — here illustrated by Thomas Schelling — is that

Exaggerating the threat won’t help. When people find out that you are doing that — and they will at some point — you lose credibility and end up further behind than when you started.

This is by no means the first time we face “apocalyptic” views regarding the future of the economy and the world. In 1968 Paul Erlich declared that “the battle to feed all of humanity is over” and predicted that “sometime in the next 15 years, the end will come,” by which he meant “an utter breakdown of the capacity of the planet to support humanity.” Also in 1968, a report by MIT’s Jay Forester, commissioned by the so-called Club of Rome, projected imminent col-
lapse for the world economy unless economic growth was halted immediately! All of the above said, one must also admit — as economist Daron Acemoglu does — that

The young activists leading school strikes and mass protests around the world have been highly effective in sounding the alarm about climate change. ... The world — particularly the United States — needs a wake-up call.
KEY CONCEPTS

tragedy of the commons  common resource  externality
negative externalities  positive externalities
marginal external cost  marginal social cost
marginal external benefit  marginal social benefit
free-rider problem  rival goods  non-rival goods
excludable  non-excludable  public goods
social cost of carbon  cumulative distribution function
discount rate  unknow unknowns  intermittency problem
Competitive markets presume that (a) there are many players both on the supply side and on the demand side; (b) property rights are well defined; and (c) all agents are well informed about the relevant information required to transact. In Chapter 8 we discussed what happens when the first assumption fails to hold, that is, when there is market power. In Chapter 9 we discussed what happens when property rights are not well defined. In this chapter, we conclude Part IV of the book by looking at the third important source of market failure, namely imperfect information. The chapter is divided into two parts. In Section 10.1, we look at markets when one side — either the seller or the buyer — is better informed that the other, what economists refer to as asymmetric information. We deal with two typical situations, adverse selection and moral hazard. The second part of the chapter, Section 10.2, deals with the role that public policy might have in protecting consumer interests when the latter are poorly informed.

10.1. ASYMMETRIC INFORMATION

Health insurance is a very unusual market. Buyers are very different from each other. Some people are very healthy and hardly need any health care beside preventive care. Other people are very sick, perhaps chronically sick, and require constant (and expensive) health
care. Relatively healthy individuals don’t perceive the need of having a lot of health insurance, and if insurance costs more than a small amount they may actually prefer not to buy insurance (if they can avoid it). The chronically ill, by contrast, badly need health insurance, not only because of the greater uncertainty regarding future expenses, but also because of the high level of health care they require at all time.

Because of this heterogeneity among insurance buyers, health insurance companies are faced with a pricing dilemma: a low price attracts many patients, healthy and unhealthy — but a low price leads to small margins. A high price, in turn, runs the risk of attracting only high-risk buyers, the ones who require large health expenditures — and a high cost leads to small margins too.

What should the insurance company do? Specifically, suppose that there are two types of insurance buyers: low-health-risk and high-health-risk buyers. Low-health-risk buyers are individuals in good health, unlikely to require medical services. High-health-risk buyers, by contrast, are individuals with a medical condition who are very likely to require expensive medical treatment. High-risk buyers constitute 10 percent of the population.

Suppose that health insurance is worth $20,000 a year for a high-risk buyer and $3,000 for a low-risk buyer. In other words, a high-risk buyer is willing to pay up to $20,000 for insurance, whereas a low-risk buyer is only willing to pay up to $3,000. This reflects the assumption that a low-risk buyer is less likely to be sick and, if she is, is less likely to require serious treatment. The differences between patient types are also reflected in insurer average (or expected) costs, which we assume are $30k for a high-risk buyer and $1k for a low-risk buyer. The idea is that a high-risk buyer most likely will use health services repeatedly, whereas a low-risk buyers will hardly show up at the doctor’s office.

Suppose we were to take a random sample from the population and ask their willingness to pay for insurance. In 90 percent of the cases we would be asking a healthy person, in which case willingness to pay is $3,000. In 10 percent of the cases we would be asking a sick person, in which case willingness to pay is $20,000. It follows that, on average, that is, considering a random individual taken from the population, willingness to pay is given by $10\% \times 20,000 + 90\% \times 3,000 = 4,700$. Moreover, on average, the
cost of providing insurance is given by \(10\% \times 30,000 + 90\% \times 1,000 = 4,500\). Since average willingness to pay is greater than average cost, one might argue that there are gains from trade, that is, society as a whole is better off if the insurance company offers insurance services to all individuals.

Specifically, suppose that buyers do not know their risk type. (I’m aware that this is not very realistic assumption, I’m only making it for expositional purposes.) Suppose moreover that a buyer’s willingness to pay is the (weighted) average of the high-risk and the low-risk buyer’s willingness to pay: \(10\% \times 20,000 + 90\% \times 3,000 = 4,700\). Then, by setting an insurance premium of, say, 4,650, the insurance company makes an expected profit of $150 per buyer and the buyer gets an expected consumer surplus of $4,700 - $4,650 = $50. In other words, everyone is better off with insurance than without.

In reality, however, buyers have knowledge of their risk level. Maybe not perfect knowledge, but at least some knowledge. For simplicity, suppose that the insurance buyer does know her risk level whereas, as before, the insurance company only knows that 10% individuals are high-health-cost buyers (that is, does not know whether a particular individual is a high- or a low-cost type). In reality, this is not an entirely realistic assumption: the insurance company is likely to know something about the buyer’s health status, and the buyer does not know her health status with certainty. What’s important for our purpose is that the buyer has a better idea of her health condition than the insurance company, which seems realistic.

Suppose that the insurance company sets the insurance premium $4,650 as before. The problem is that, since insurance buyers know their risk level, only high-risk individuals accept the offer of buying insurance for $4,650. By contrast, low-risk individuals have a willingness to pay of $3,000, which is well below the price of $4,650. It follows that, conditional on an insurance offer being accepted, the cost of providing health services is given by $30,000, which is considerably higher than the price of $4,650. To understand this, notice that, while the insurance company does not know the health condition of each person in the population, the insurance company knows that, if its offer is accepted, the buyer is a high-risk patient for sure and the cost of offering her health services amounts to $30,000. We conclude that, by selling at $4,650 an insurance company risks losing money: the cost per customer, $30,000, is considerably greater than
Surgeons during an operation.
Healthcare markets are arguably the most important instance of adverse selection.

the revenue per customer, $4,650.

In order to avoid losing money, the insurance company would need to set a price at least equal to $30,000. However, such a high price would drive away even high health-risk, high willingness-to-pay buyers. Bottom line: there is no equilibrium where health services are bought and sold in the market! This example illustrates the problem of adverse selection.

If the uninformed side of the market makes an offer which is then accepted or rejected by an informed side of the market, then there may be situations when the market shrinks or completely collapses even though there are gains from trade. In other words, information asymmetry may lead to market failure.

In the above example, the uninformed side of the market is the seller of health insurance, whereas the informed side of the market is the patient. Unfortunately, there are many other real-world situations featuring this sort of information asymmetry and this sort of negative implications (cf next section). Fortunately, one of them led to a famous funny quote by comedian Groucho Marx. While there is some dispute about the exact quote and its circumstances, it appears that New York’s famous Friar’s club admitted Groucho Marx as a member. Since Groucho did not participate in the club’s activities, he sent in his resignation letter. The president inquired why Groucho was resigning, to which he replied, “Because I don’t want to belong to any club that would have me as a member!” Can you see how this is a case of adverse selection?
SOLVING THE ADVERSE SELECTION PROBLEM

In Chapter 8 we saw how public policy (antitrust) plays an important role in addressing market failures resulting from market power. In Chapter 9 we saw how Pigouvian taxes (and other mechanisms) play an important role in addressing market failures resulting from externalities. What about adverse selection — are there policy instruments to address this type of market failure?

Let us go back to the problem of health insurance. One first solution is to impose a health mandate. As the simple high-risk-low-risk example in the previous section suggests, the market would exist and create gains from trade if all patients purchased health insurance. This implies that one possible solution to the adverse selection problem is to enforce purchase. This induces a pooling equilibrium, that is, an equilibrium where low and high risks are pooled.

The Massachusetts health reform, which was signed into law in April 2006, included a health insurance mandate as one of its key components (non-poor residents either purchase a health insurance plan that meets minimum coverage criteria or pay a penalty). The Massachusetts experiment would later became a model for the national reform popularly known as Obamacare.

Did these policies make a difference in terms of alleviating the adverse selection problem? The example presented earlier is just that — an example. Is adverse selection really important in real-world health markets? One possible test for the presence of adverse selection is the following: as the number of insured individuals increases, we should observe a decrease in the average cost of the insured. This would be consistent with the observation that the added insured individual was a lower-risk individual than the previously insured individuals.

The evidence from Massachusetts is consistent with this interpretation: The 2006 reform increased coverage by 26.5 additional percentage points, and the growth in coverage was associated with a reduction in the average cost of the insured by 8.7 percent. Overall, the reform is estimated to have increased total surplus by about 5 percent, of which about 4 percent corresponds to alleviating the adverse selection problem.

A more radical way to solve the adverse selection problem in health markets is to follow a universal healthcare system, as many
countries in Europe and throughout the world do. Although there is a lot of variation across countries, it is not uncommon for the state to offer directly health care services through state owned health care facilities and affiliated doctors. This solution essentially solves the adverse selection problem because by design all patients are included in the same pool.

A related policy, in terms of its pooling effects, is that of a single payer system, in which the government pays for health care for every citizen, even as health care services are provided by private hospitals and doctors. In the US, the Medicare program offers health insurance to residents over age 65 and people with disabilities. One proposal often discussed, “Medicare for All”, would extend these benefits to all of the population. In sum,

*Solutions to the adverse selection problem in healthcare include insurance mandates, government-offered insurance, and government-offered health services.*

Although the discussion in this section is couched in terms of health markets, the basic ideas are more widely applicable. For example, one of the reasons why car insurance is mandatory is that it pools all types together, thus mitigating the adverse selection problem. In reality we do not have complete pooling, in the sense that different drivers are subject to different rates, depending on observable characteristics (e.g., age), but the pooling effect described above plays a role.

**OTHER EXAMPLES OF ADVERSE SELECTION**

There is a reason why people focus on healthcare when discussing the adverse selection problem: Healthcare represents a huge chunk of GDP (in the US, about 17%, in Europe about 10%). However, there are many other instances when adverse selection plays a role, resulting in some form of market failure.

The original economics research on adverse selection took used car sales as a motivating example. Suppose I’m in the market for a 2014 Toyota Camry. I know that there are good used cars and bad used cars. I have a hard time finding out whether a particular 2014 Toyota Camry is in good shape. Equally important, I know that the
10.1. ASYMMETRIC INFORMATION

Quintin Soloviev

American Airlines AAirpass program is a classical example of the perils of adverse selection and moral hazard.

seller knows whether that particular 2014 Toyota Camry is in good shape (maybe not perfectly, but certainly better than I do). Consulting the blue book, I see the average price is about $15,000. However, depending on the car’s condition its value can be as high as $20,000 and as low as $10,000. Now, here’s the essence of the adverse selection problem in this context: If I offer to pay $14,000 (a little less than average value), then the owners of the best 2014 Toyota Camry’s (those worth about $20,000) are unlikely to accept the offer. This implies that, conditionally on the offer being accepted, the average value is lower than $15,000, maybe even lower than my $14,000 bid.

In 1981, American Airlines launched a membership-based discount program called AAirpass. At one point the program offered lifetime unlimited travel on American Airlines and unlimited access to Admirals Club locations. Some of these passes remain valid today. The lifetime passes cost $250,000. Not surprisingly, they attracted very frequent flyers. In 2007, an AA internal report concluded that two specific passengers were costing the airline more than $1 million annually (for example, one of them made 500 trips to London in 10 years). Their passes were eventually cancelled due to fraudulent behavior (attempting to sell a companion ticket). Although these were rather extreme cases, overall the pass was a major flop on account of the adverse selection problem.

A more recent example of the “adverse selection tragedy” is provided by the MoviePass program. This is a subscription-based movie ticketing service founded in 2011. Subscribers were issued a branded prepaid debit card. Using the MoviePass mobile app, users checked-in at a supported cinema, and selected a film and showtime occurring
within the next 30 minutes. The card was then automatically loaded with the amount of money needed to purchase a single ticket from the cinema for that film. Although Movie Pass’s business model was never completely clear, one important element was to use or sell the user information obtained from users. The program was reported having 2 million subscribers by February 2018 and to be supported at 91% of US cinemas.

The adverse-selection nature of MoviePass is that it is disproportionately likely to attract heavy movie goers, so that the cost per subscriber is greater than what it would be if MoviePass attracted a representative sample from the population.

Over the years there were many changes in rules, including reducing the “unlimited” plan to only offer three no-cost tickets per-month. Despite all of these adjustments, the program remained essentially unviable, largely due to the adverse selection problem. On September 13, 2019, MoviePass announced the service would shut down the following day.

**MORAL HAZARD**

I read the other day that “there are two types of people in this world — people who have suffered from a cracked iPhone screen and those who live in fear of the inevitable.” If this is so, why don’t we all buy insurance against a cracked screen? As it happens, there are insurance products in the market. However, they are typically very expensive. For example, protectyourbubble.com sells insurance for £8 a month, which comes to £96 a year, or about $125 a year. That’s a lot of money! In fact, $125 is not very different from the annual amortization of a phone. Oh, and did I mention there is also a $125 deductible?

Why is insurance so expensive? Partly because of adverse selection: If I have a history of smartphone screen cracking, I am more likely to buy insurance. Insurance companies are aware of the relation between buyer types and likelihood to purchase, and so price accordingly.

However, important as the adverse selection problem is, insurance contracts such as screen cracking are subject to another huge problem, one which frequently leads to market failure: **moral hazard**. The issue is that the likelihood of a cracked screen depends a
Moral hazard is the main reason why it is so difficult to find affordable insurance against a shattered screen.

lot on how careful the owner is. And the incentives for the owner to be careful are considerably lower if her phone is insured against a cracked screen.

How can moral hazard lead to market failure? Suppose that the probability of a cracked screen is small if the owner takes good care of it (e.g., places a protective cover on the phone) and high if the owner does not take good care of the phone. Suppose that there are gains from trade in the former case but not in the latter. In other words, if the owner takes good care of the phone then her willingness to pay for insurance is greater than the cost of offering insurance. However, to the extent that the insured phone owner is likely to be careless — and the insurer is aware of it — then there exists no price such that both seller and buyer are willing to trade.

Why is the section on moral hazard included in the chapter on information? Because the moral-hazard market failure results from the insurer’s inability to observe the insured party’s effort.

The inability of one party in a transaction to observe the other party’s actions (moral hazard) may reduce the realized gains from trade.

If the insurer could easily monitor its customer’s actions, then it could offer a policy such that you would only be paid if you had been careful not to crack the screen. In other words, we are faced with another case of market failure due to information asymmetry: The user knows whether she is taking good care of her phone, but the insurance company does not.
INCENTIVES

In Section 2.3 we mentioned incentives as one of the central themes of microeconomics. The problem of moral hazard brings the issue of incentives to the fore. The operative expression is, “it’s not my money.” If people talk or think or act in this manner, then there is a moral hazard problem. One example that touches many of my readers is that of course textbooks. The choice of a textbook is made by the instructor but the purchase is made by the student. Suppose there are two candidate textbooks, one selling for $30 and the other for $300. Even if the latter is just a little better than the former (e.g., it comes with better teaching materials), an unscrupulous instructor is likely to assign the expensive book. It’s not my money.

In Chapter 8 we mentioned the fact that healthcare is considerably more expensive in the US than in comparable European countries. One possible explanation — one that I proposed in Chapter 8 — is market power. Another explanation is incentives: Frequently, healthcare providers are reimbursed based on the fee-for-service model, that is, they are paid for every appointment, test, procedure, etc. This creates an incentive to multiply the quantity of healthcare even when the increment in terms of quality is not that significant (do we really need that MRI at this moment?). Since the cost is passed on to the insurance company, the healthcare provider has little to worry about. It’s not my money.

One can argue that the adverse selection examples considered in the previous section are also examples of moral hazard (frequently, adverse selection and moral hazard go hand-in-hand). For example, if I purchased the AAirPass then my decision to fly will be determined by many factors but not by the cost of flying (since the price I need to pay is zero). This is similar to the phone user who is not careful about cracking the screen. It’s not my money.

At a deeper and more fundamental level, the issue of moral hazard came to the fore in the aftermath of the 2008 financial crisis. In the fall of 2008, Congress authorized then-Treasury Secretary Henry Paulson to spend $700 billion to rescue the rapidly-failing financial system. The Troubled Asset Relief Program (TARP) was set in place. Paulson used the money to invest in banks, auto companies and the insurance giant AIG. By many measures, the program was a success. However, at a Congressional hearing, Neil Barofsky, the special in-
spector general for TARP, argued that

The good financial news should not distract from the careful and necessary assessment of TARP’s considerable, non-financial costs, [in particular] the increased moral hazard and potentially disastrous consequences associated with the continued existence of financial institutions that are ‘too big to fail.’

The idea is that being “too big to fail” diminishes your incentives to avoid risks, in the expectation that losses will be “socialized” through TARP-like government bailouts. However, considering how rare these events have been (in US history) one can argue that the benefits of TARP — avoiding the failure of key corporations, with all of the “domino” effects that this might imply — exceeded its costs, including its moral hazard costs.

THE PRINCIPAL-AGENT PROBLEM

American Express has a problem: its AmEx card is not accepted at all stores. In the past, this was partly due to AmEx charging higher rates from merchants than other cards. Since then, rates have lowered, but the reluctance of retailers to accept AmEx remains. Faced with this difficulty, American Express hires consultants whose job is to persuade stores to install the AmEx equipment and accept the AmEx card. The job is done when finally that magic decal is placed on the store’s window (“We accept AmEx cards”).

How should American Express reward its consultants for this task: Pay consultants by the hour? Pay consultants by the number of stores they “convert” to AmEx? One thing is clear: American Express does not have the ability to carefully monitor its consultants’ actions: The stores are dispersed throughout the country, and so the job of following the consultants would be as difficult as American Express itself doing the consultant’s job.

The above example illustrates one of the most canonical cases of moral hazard, a case known as the principal-agent problem. A principal (American Express in the above example) hires an agent (the consultant in the above example) to perform a certain task. What makes this a challenging problem is that the outcome depends on the
agent’s effort as well as on other factors, and moreover the agent’s actions are not directly observable by the principal. Specifically, the challenge is how to best design a reward system so as to get as close as possible to the perfect-information outcome, that is, the outcome when the principal can easily observe the agent’s actions. Before, we considered two options: pay the consultant by the hour or pay the consultant by the number of stores “converted”. The advantage of paying by the hour is that it provides the consultant with secure compensation — you know what you can count on — which in turn makes it easier to find consultants willing to take the job. The advantage of paying strictly based on success is that it provides the agent with strong incentives: it is no longer someone else’s money, it is now my money.

More generally, economists say that paying based on success is an example of a high-powered incentive scheme, whereas paying a fixed wage is an example of a low-powered incentive scheme. Since a high-powered incentive scheme involves both costs and benefits, there is no general optimal scheme: as often is the case in economics, it depends. Specifically, it depends on how closely the outcome tracks the effort put in by the agent. If the outcome is very closely related to effort, then the risk associated with pay for performance is low, and thus the benefits from a high-powered incentive scheme dominate. By contrast, if there is a lot of noise in the outcome measure, then pay for performance leads to a very volatile compensation level, and the relative benefits of a fixed wage dominate.

10.2. CONSUMER PROTECTION

In the previous sections we examined two cases of asymmetric information: the cases when a seller does not know a buyer’s type (e.g., is the patient a health risk?) and the case when the seller is unable to observe a buyer’s action (e.g., will the insured owner take care of the product?). In this section we consider the all-too-common situation when a buyer is poorly informed about the seller or the conditions offered by the seller.

Have you ever read the entirety of a user agreement before entering a website? I didn’t think so. I don’t know anyone who has: we all click on the “agree” button without giving it much thought.
larly, many sales contracts (e.g., a credit card contract) include pages and pages of small print which we could read but do not actually read. This may be a problem — in fact, may lead to market failure — because it allows sellers to impose terms that are legal (strictly speaking) but harmful to the buyer. In this case, the market failure is that a transaction takes place (perceived willingness to pay is greater than price) even though it should not have taken place (actual willingness to pay is lower than cost).

In the United States, many insurance plans require that the patient use hospitals and services from their insurance network. However, with some frequency patient treatment includes out of network services. As a result, when the patient is asked to pay the bill, in addition to the normal co-pay he or she may be asked to pay the balance of the bill that is not covered by the insurance plan. This is often referred to as balance billing or surprise billing, a reference to the fact that many patients — especially in emergency situations — are unaware that some services are out-of-network services. In recent years, several states in the US have enacted legislation which limits the extent of this type of surprise billing. In other countries (e.g., Germany, Japan) balance billing is simply prohibited.

A similar phenomenon occurs in the wireless industry, where consumers may fall victim to the so-called bill shock. One common reason (in the US) used to be exceeding the monthly allotment of voice minutes, text, or data consumption. To remedy this problem, in 2013 the US Federal Communication Commission successfully persuaded cell phone companies to inform users when they approach and exceed their voice, text, or data allowances.

In many instances, consumers purchase a bundle of products and/or services. For example, if I buy a Keurig coffee machine I will likely also buy Keurig coffee capsules as well. As of January 2020, I can buy a machine (from the Keurig website) for as little as $64.99. Coffee pods go for about 50 cents each. If I drink one per day, after a year I already spent north of $180 on coffee pods. In other words, expenditure with add-ons is considerably greater than with the initial purchase. If consumers are perfectly informed about all prices, including add-on prices, then this is not a problem: each consumer will compare his or her willingness to pay to the price they are asked to pay, and then make the right decision. However, many consumers are not aware of the total cost they will need to incur when
they buy a coffee machine or a printer or when they open a bank account. This can lead consumers to underestimate the actual effective price they have to pay and, as such, lead to a market failure: a transaction takes place (willingness to pay is greater than perceived price) when it should not take place (willingness to pay is lower than actual price).

To conclude this section, a related area of public policy is the enforcement of **truth in advertising**. To quote from the FTC site

> When consumers see or hear an advertisement, whether it’s on the Internet, radio or television, or anywhere else, federal law says that ad must be truthful, not misleading, and, when appropriate, backed by scientific evidence.

In this context, much of the agency’s role is to warn companies when there is suspicion that they may be violating the FTC Act and that they can face serious legal consequences if they do not immediately stop.

**SEARCH GOODS AND EXPERIENCE GOODS**

In some cases, the quality of a product can be found out before purchase. For example, if you are shopping for a coat, you try it on to see if it fits, you look at it to see if you like it. You may be in for a surprise later on, but by and large what you see is what you get. Economists refer to these goods as **search goods**. In other cases, however, you are only able to judge the quality of what you bought after you bought it and experienced it. For example if a new falafel food truck parks in your neighborhood you won’t be able to know how good it is until you try it. Economists refer to these goods as **experience goods**. There is an even more extreme case, what economists call **credence goods**. This is the case when, even after experiencing a service, the buyer has difficulty ascertaining its quality. Suppose your doctor says you need a double-bypass surgery and suppose you actually go through it. Even if you survive the operation, how can you be sure everything was well done? How do you know the decision to operate was the right one in the first place? (Suggestion: ask for a second opinion.)

Experience goods (and credence goods) lead to a potential market failure. If the consumer can only discover the quality of her purchase
after she paid for it, then the incentives for sellers to provide a quality product are lower than in a search good. In the latter case, if the seller offers a low-quality product then consumers — who are able to determine quality before purchase — avoid purchasing the low-quality product. In a world of experience goods, however, sellers may be tempted to offer low quality products in the hope that consumers will purchase them before they find out they made a mistake.

Fortunately, there are market mechanisms to prevent this opportunistic behavior by sellers. In particular, if consumers purchase a product or service repeatedly, then they can “punish” bad sellers by not buying from them again. As the saying goes, “Fool me once, shame on you, fool me twice, shame on me.” Box 10.1 provides an interesting application of this idea to the issue of restaurant quality. Because of the role played by repeat purchases, you’d expect quality to be greater in restaurants frequented by repeat customers than in restaurants frequented mostly by tourists (ever heard the expression “tourist trap”?).

In addition to repeat purchases, there may be other reputational mechanisms which give sellers the right incentives. Think, for instance, of professional or crowd-sourced reviews. Would you go to a restaurant with a 1-star rating on Yelp? With the increase in online sales and online product reviews, this mechanism of seller reputation has become increasingly important. It also has its issues (e.g., fake reviews), but it’s certainly better than no information at all.

Many market transactions involve considerable uncertainty for consumers. Seller reputation partly helps keeping seller opportunism in check, but public policy also plays an important role in consumer protection.

NUDGES

(Reader advisory: the ideas discussed in this section are highly controversial.) In the previous two chapters we saw two reasons why governments should regulate markets: market power (normally on the seller side) and poorly defined property rights (a.k.a. externalities). In Section 10.1, we added one more reason: information asymmetry (one side of the market is better informed than the other, as
Restaurant hygiene cards are now a common feature in the US. As you walk into an establishment, you will likely find its grade (A, B, C) displayed by the entrance. (If it’s a C, consider going elsewhere).

Mandatory hygiene cards, which were first introduced in California in 1998, provide a way to test the relation between consumer information and market outcomes. First, research shows that making restaurant grades available to consumers led to a 20% decrease in foodborne hospitalizations. This suggests that better information puts pressure on sellers to provide better service.

It is also interesting to note that the impact of hygiene cards was not uniform across restaurants: Restaurants located in tourist areas typically had worse grades before the introduction of scorecards and showed a greater improvement in hygiene following the introduction of score cards.

The figure below shows the best and the worst restaurants in Santa Monica according to hygiene card grades. Note, for example, that the Venice boardwalk — a typical tourist area — includes few of the best restaurants but many of the worst ones.

![Map of best and worst restaurants in Santa Monica](image)

This pattern is consistent with the idea that restaurants in tourist areas have less market incentives to provide good service, since most customers are not repeat customers. In this case, market regulation provides a good substitute for seller reputation.

in the adverse selection or moral hazard problems). Section 10.2 considers a related possibility: consumers are poorly informed about the seller’s conditions (product quality, add-on prices, etc).

In some cases, public policy goes beyond that: It addresses not the individuals’ lack of information but the individuals’ poor judgement
or inertia or inability to do what they want to do. Right away, you can see how controversial this is. Libertarians — about whom we write more in Chapter 12 — are firmly set against most of the policies included in this section: The individual, they argue, is the supreme judge of their own preferences and, absent any significant side-effects on third parties (externalities), there should be no public policy with the goal of directly or indirectly influencing their behavior.

Reality is more complicated than this. Is it right for the government to mandate the use of safety belts or bike helmets? “It’s my life,” you might say. But in case an accident happens there is a good chance you will impose a significant burden on the health system (which is paid by all), a burden that would be smaller (on average) if you’d buckled up when driving or wore a helmet when riding. In this sense, we might think of safety regulation not so much as a form of “paternalism” but rather as a way to correct for an externality.

A recent doctrine in favor of government intervention comes under the name of nudge theory. Its main premise is to encourage people to make decisions that are in their broad self-interest. Economists Richard Thaler and Cass Sunstein write that

> By knowing how people think, we can make it easier for them to choose what is best for them, their families and society.

For example, in 2012 the UK government changed the way private pension fund contributions work. Workers became automatically placed into a firm’s scheme, with contributions deducted from their pay packet, unless they formally requested to be exempted. However, the new regulation did not change the set of options available to each worker, it simply changed the default choice. The policy change had an enormous effect: Active membership in private sector pension schemes jumped from 2.7 million in 2012 to 7.7 million in 2016.

A similar example is given by organ donation regulations. Spain operates an opt-out system: All citizens are automatically registered for organ donation in case of death by accident unless they choose to state otherwise. By contrast, in the UK donors have to opt in in order to become organ donors. The Spanish system is one of the reasons why Spain is a world leader in organ donation.

Nudge theory is not free from criticism. Some claim that
Studies estimate that Facebook creates a consumer surplus of about $10 billion per month in the US only. However, these estimates, based on compensation required to de-activate Facebook, may denote the level of addiction to Facebook more than the value it creates.

Government-by-nudging amounts to a kind of technocracy, which assumes that experts will know which choices are in the interests of ordinary people better than those people know themselves. This may be true under some circumstances, but it will not be true all of the time, or even most of the time, if there are no good opportunities for those ordinary people to voice their preferences.

What about addiction? Is it right to enact policies that discourage people from products and services that are addictive? Government policy regarding substance abuse — for example, the US recent fight against the opioid crisis — suggests a positive answer. But what about Facebook? In Chapter 7 we presented estimates of the demand for Facebook and the value created by the social network. These estimates are based on asking people how much they would require to de-activate their account. By these calculations, Facebook creates billions and billions of dollars of value each month. But what if the reluctance to break from social networking is a reflection of addiction rather than value creation? For example, it has been estimated that approximately 50 percent of 18–24 year-olds visit Facebook as soon as they wake up. This suggests that the idea of social networking in general, and Facebook in particular, may share some of the addictive features present in other activities (such as substance abuse).

Continuing in the direction of greater controversy, what about public policy regarding access to information? When and how should public policy regulate the information individuals have access to? Is it right to block consumers from accessing fake news? If so, who defines what constitutes fake news and how is that classi-
fication made? You can see how these types of policies are ripe for criticism, and not just from die-hard libertarians. We will return to these issues in Section 12.2 when we discuss the government provision of goods and services as a form of “interpreting” and pursuing the interests of ordinary people.
KEY CONCEPTS

adverse selection  health mandate  pooling equilibrium

ObamaCare  universal healthcare  single payer system

moral hazard  incentives  principal-agent problem

principal  agent  high-powered incentive scheme

low-powered incentive scheme  surprise billing

bill shock  add-on pricing  truth in advertising

search goods  experience goods  credence goods

nudge
PART V
SOCIAL JUSTICE
As mentioned in Chapter 7, competitive markets perform quite well in terms of efficiency, that is, in terms of generating value from trade. However, as we then mentioned, and as many have repeated in recent years, “market forces and capitalism by themselves aren’t sufficient to ensure the common good.” This chapter is based on the observation that the unregulated market economy may result in outcomes that are not equitable — no matter how efficient they are. In Section 11.1, we deal with inequality — in particular inequality in the distribution of income —, whereas Section 11.2 focuses on discrimination and exclusion.

11.1. MEASURING AND EXPLAINING INEQUALITY

In Chapter 1, we put forward the idea that the capitalist revolution, by having taken place selectively in a few countries, led to an increase in inequality across nations. As we will see next, this is only one of many dimensions of the inequality phenomenon.

TYPES AND MEASURES OF INEQUALITY

There are marked differences in income levels across the world. One way of measuring these is to derive the worldwide income distribution. The blue lines on the top panel of Figure 11.1 do this for the
11.1. MEASURING AND EXPLAINING INEQUALITY

years 1988 (dashed blue line) and 2008 (solid blue line). Specifically, these lines correspond to the income probability density function in each of the years. The term probability density function may sound a bit intimidating for the reader less familiar with statistics jargon, but the idea is simple: Suppose that you create a histogram by dividing income levels into a series of bins: from 0 to 100, from 100 to 200, from 200 to 300, etc. For each bin, you measure the percentage of people whose income falls into that bin and draw a column the height of which measures the percentage value. Now suppose you make the bins thinner and thinner. Eventually, instead of a series of columns you have something like a continuous line that connects the tops of the bins. That’s essentially the probability density function.

The world income probability density function ranges from about $100 to about $100,000, meaning that most people’s income levels fall within that range. Of course there are people who earn more than $100,000, and unfortunately there are also people who earn less than $100 a year. But most fall within the 100-100k range. This range is so wide that we measure income on a logarithmic scale: each extra tick corresponds to multiplying income by 10, not adding 10 (as we would on a normal scale).

Notice the dashed blue line is bimodal, which means it has two humps: one at about $1,000 — or a little less than that — and another one at about $10,000. Bimodal distributions like this one reflect a high degree of inequality, with two relatively “separate” groups. Naturally, this separation is not perfect, but one might say that, by 1988, the world was broadly divided into those earning less than $5,000 a year and those earning more than that.

By contrast, the 2008 world distribution looks rather more unimodal, that is, closer to the typical bell-shaped distribution we find in many other settings. In this sense, we might say that, worldwide, the two-decade evolution from 1988 to 2008 has been in the direction of lower inequality. The main reason for this evolution is given by the other pair of densities plotted in the top panel of Figure 11.1 (in red), both representing China. As can be seen, the twenty years from 1988 to 2008 saw a remarkable shift in the density to the right, that is, in the direction of higher income levels. The peak of the China 1988 density — the mode, in statistical jargon — was well below $1,000 in 1988, whereas by 2008 it was already above $1,000. In plain English, we might say that, in the past few decades, economic growth
11.1. MEASURING AND EXPLAINING INEQUALITY

in China has expanded the world’s middle class to an extent that the world income distribution now looks more like one group rather than two different groups of people (that is, the income distribution went from bimodal to unimodal).

This is all very well, but there are also reasons to believe world inequality is still considerable — and, in some sense, greater than it was a few decades ago. The bottom panel of Figure 11.1 plots the income density functions for two groups of countries: Sub-Saharan African countries and mature economies (essentially Europe, North America and a few other countries). The two distributions hardly overlap, de-
noting a high level of cross-country-group inequality. Moreover, the movements from 1988 to 2008 have been in the direction of increasing inequality, the opposite of what the top panel seems suggests. The difference between the two perspectives corresponds, essentially, to the “China effect”: If we take China into consideration, then we are inclined to say the world is more equal than it was a few decades ago. If we exclude China and compare the wealthier and poorer countries, then we are inclined to say the world is more unequal than it was a few decades ago.

As important as cross-country inequality levels are, most of the time the problem of income inequality is cast in terms of within-country inequality. The social protests of recent years — for example, Occupy Wall Street in 2011 or the Yellow Vest Paris movement of 2018 — were more concerned about social conflicts in the US and France than about inequality with respect to other countries.

Within-country inequality is frequently represented by a Lorenz curve and measured by the associated Gini coefficient. Figure 11.2 plots the Lorenz curves of a variety of countries. Take Brazil, the curve in green. First we compute the percentage of total income that is accounted for by the bottom 10 percent of the income distribution. In Brazil, this corresponds to a very low value, about 1.2 percent. We therefore plot the point (.1,.012), denoting that the bottom 10 percent population receive 1.2 percent of the total income of Brazil. Next we compute the percentage of total income that is accounted for by the bottom 20 percent of the income distribution. In Brazil, this is given by 3.6 percent, thus leading to the point (.2,.036). We continue this process until we finally get to (1,1): by definition, the bottom 100 percent of the population (everyone) must account for 100 percent of income (all of it). Connecting all of these points, we get Brazil’s Lorenz curve.

In the limit when every citizen receives the same income level, the Lorenz curve coincides to the main diagonal, that is, the line going from (0,0) to (1,1). This is the only Lorenz curve such that the bottom x percent of the population receives exactly x percent of total income. All of the Lorenz curves in Figure 11.2 fall below the main diagonal, meaning income distribution is to some extent unequal. For example, notice that Brazil’s Lorenz curve includes the point (.9,.6). This implies that the bottom 90 percent account for 60 percent of total income, which in turn implies that the top 10 percent account for 40
percent of total income. This is a high number! In fact, income distribution in Brazil is rather unequal. This we can see by the fact that Brazil’s Lorenz curve is very low, that is, very far from the main diagonal.

From the Lorenz curve we derive the Gini coefficient, a summary measure of inequality in a given economy. The Gini coefficient corresponds to the area between the Lorenz curve and the main diagonal (multiplied by 2). In the limit of complete equality, the Lorenz curve corresponds to the main diagonal, and thus the Gini coefficient is equal to zero. In the opposite case, when one person only out of millions gets all of the income, the Lorenz curve would coincide with the horizontal axis and then jump to 1 at 1. The area between the Lorenz curve and the main diagonal would be \( \frac{1}{2} \) (the area of a triangle with sides equal to 1). Multiplying by 2, we get a Gini coefficient of 1. We thus conclude that the Gini coefficient varies between 0 (no inequality) and 1 (maximum inequality).

One disadvantage of the Gini coefficient is that it loses some information with respect to the Lorenz curve. For example, as can be seen from Figure 11.2, the Lorenz curves for Sweden and India cross: the bottom 10 percent of incomes in India correspond to a higher percentage of total Indian income than the bottom 10 percent of incomes in Sweden represent of Sweden’s total income. However, the top 10 percent of incomes in India correspond to more of India’s income.
than the top 10 percent of Sweden. In those cases, we cannot make unambiguous comparisons unless we introduce additional inequality measurement criteria.

One advantage of the Gini coefficient is that it’s just one number. This allows us better to compare a large number of countries. For example, Figure 11.3 plots, for each country, the value of per-capita GDP and the country’s Gini coefficient. Is there a relation between a country’s per-capita GDP and the country’s level of inequality? Figure 11.3 suggests that there may be a weak negative relation — wealthier countries have lower Gini coefficients — but the relation is weak. Moreover, there are outliers such as the US, with a considerably higher Gini coefficient than other countries with similar levels of per-capita GDP.

One feature of the Gini coefficient is that it gives the same weight to all parts of the Lorenz curve. However, judging by the social protests and more generally by the media coverage, people tend to be more concerned with income and wealth concentration at the very top. The Occupy Wall Street movement, for example, was very much focused on the top 1 percent of the income distribution. In fact, one of its main slogans was precisely “we are the 99%.”

Figure 11.4 shows the evolution of the top 1 percent share in some selected countries. It illustrates a much-talked-about stylized fact: that the level of inequality, which had declined in many countries
since the middle of the 20th century, has been increasing since about 1980. This happens to coincide with the shift to a more free-market economic policy approach in the US, the UK, and several other countries (cf Section 2.4). This correlation suggests one possible cause for the increase in inequality, but the question is more complex than that.

Although much of the debate on inequality centers on income inequality, there are many other dimensions of disparity to consider. Figure 11.5 shows the concentration on the top 1 percent of three eco-
11.1. MEASURING AND EXPLAINING INEQUALITY

Economic measures: expenditure, income, and wealth. Two remarkable facts stand out: First, there is considerable variation across countries, with the US standing out as a clear outlier in terms of wealth concentration: the top 1 percent concentrate more than 35 percent of the country’s wealth! Second, the concentration of wealth is greater than the concentration of income, which in turn is greater than the concentration of expenditure. One possible explanation for the difference between concentration of income and concentration of income is that the latter results from income variability over time rather than real inequality. Suppose for simplicity that everyone lives for three periods and that income is very high during period 2 but very low during periods 1 and 3. If I measure the income distribution in a given year, I will observe a lot of variation, but much of it results from life-cycle variation rather than genuine inequality. Continuing with the example, each individual is aware of the variation in their income throughout their lifetime. Accordingly, they save and draw on savings so as to keep a more constant expenditure stream. In other words, each individual’s expenditure pattern represents a smoothing of the uncertain lifetime income stream. This implies that any measure of the expenditure distribution in a given year will show greater uniformity than the income distribution in a given period. Research shows that part of the increase in income inequality may be explained by increased income volatility, a trend that is shown in measures of income inequality but less so in measures of expenditures inequality.

SOURCES OF INEQUALITY

The relation between income and wealth in Figure 11.5 suggests a neo-Marxist explanation for the widening gap we’ve observed in the past few decades: Very wealthy individuals consume a small fraction of their income. What they don’t consume accrues to their wealth. Moreover, their wealth does not simply increase as a result of non-spent income — it also grows as a return on investment. Meanwhile, for the poor, who essentially consume what they earn, the only hope for improvement is that the expansion of output will lead to an expansion in their wage income. As Piketty puts it in Capital in the Twenty-First century, “once constituted, capital reproduces itself faster than output increases. The past devours the future.” (A not-too-subtle reference to Das Kapital and the Communist Mani-
festo’s “the past dominates the present”). Numerically speaking, the economic condition of the wealthy grows at the rate of interest $r$, whereas the economic condition of the poor grows at the economy’s growth rate, $g$. And historically $r$ has been considerably higher than $g$. A beautiful theory of an ugly reality.

One criticism of this theory is that the concentration of capital income among the very wealthy is not that high. Figure 11.6 shows the split of income sources for various US income brackets. The figure shows that labor income represents almost 80 percent of the income of the bottom 25 percent, but less than 50 percent of the income of the top 10 percent. More than half of the income earned by the top 10 percent results from business income, dividends, capital gains and other non-labor sources. At first, this seems to provide credence to the capital vs labor theory of widening gaps. However, it’s important to understand what underlies the “business” component of income. As a recent study shows (pithily titled Capitalists in the Twenty-first Century), much of the business income of the top 10 percent corresponds to S corporations (as opposed to the more commonly known C corporations). These S corporations essentially work as tax shelters for highly skilled professionals: doctors, lawyers, engineers, accountants, consultants. The idea is that a large fraction of what is classified as business income in the statistics (and in Figure 11.6) is actually labor income (even if it is not wage income). Once this correction is taken into account, we conclude that the labor income frac-
11.1. MEASURING AND EXPLAINING INEQUALITY

FIGURE 11.7
K and L: perfect substitutes (top) and perfect complements (bottom)

...tion among the wealthy is not that different from the poor. In fact, the fraction of income accounted by labor — about two thirds — is fairly constant across income levels.

An alternative source of increased income inequality is therefore the widening gap between the labor income of the rich and the labor income of the poor. Specifically, economists refer to the phenomenon of skilled-biased technical change. The idea is that developments in robotics and artificial intelligence (AI) have disproportionately increased the demand for high-skilled workers, thus increasing the wage gap between high-skill workers and low-skill workers.

In order to understand the concept of skill-biased technical progress, it helps to go back to Section 5.2’s analysis of the firm’s in-
11.1. MEASURING AND EXPLAINING INEQUALITY

put mix. Figure 11.7 largely reproduces Figure 5.8 but casts the problem in terms closer to the problem at hand. The top panel depicts the effect of technical progress on the cost of robots, which we assume are close substitutes for low-skilled labor. In fact, in the graph we assume they are perfect substitutes, so that the isoquants $q_1$ and $q_2$ are straight lines. While the perfect substitutes assumption does not hold strictly, this extreme case serves the purpose of illustrating the main effects. Initially robots are expensive (price $r_1$), so that the optimal input mix (point $m_1$) is to offer $L_1$ jobs and employ no robots. Then the cost of a robot falls to $r_2 < r_1$. Assuming — for simplicity — that the firm continues to operate on the same input budget, the budget line pivots upward around $m_1$. The new optimal input mix is $m_2$, that is, the firm employs $K_2$ robots and no labor. In practice, however, we would expect the isoquants to be curved somewhat. Moreover, we would expect $w$ to adjust. All in all, we would expect both a drop in $w$ and in $L$.

Consider now the bottom panel, which depicts the effect of technical progress on the cost of AI systems, which we assume are close complements of high-skilled labor. In fact, in the graph we assume they are perfect complements, so that the isoquants $q_1$ and $q_2$ are perfectly L-shaped lines. While the perfect complements assumption does not hold strictly, this extreme case serves the purpose of illustrating the main effects. Initially AI systems are expensive (price $r_1$), so that the optimal input mix (point $m_1$) is to employ $L_1$ units of labor and $K_1$ AI systems. Then the cost of an AI system falls to $r_2 < r_1$. Assuming — for simplicity — that the firm continues to operate on the same input budget, the budget line pivots upward around $m_1$. The new optimal input mix is $m_2$, that is, the firm employs $K_2$ AI systems and $L_2$ units of labor, where $L_2 > L_1$. In practice, however, we would expect the isoquants to be curved somewhat. Moreover, we would expect $w$ to adjust. All in all, we would expect both an increase in $w$ and in $L$.

In sum, taking into account the heterogeneity of capital (robots are not the same thing as AI systems) and the heterogeneity of labor skills, we observe that the drop in the cost of capital has depressed the demand for low-skilled jobs and increased the demand for high-skilled jobs.

Figure 11.8 shows the evolution of the relative quantity and price of skilled vs unskilled labor in the US. Specifically, on the vertical axis
we measure the ratio of the wage paid to a college graduate and the wage paid to a high-school graduate. Sometimes the same information is also presented as the percent difference rather than the ratio and referred to as the college premium. (We will return to this in Chapter 13 when talking about college as an engine of economic mobility.) On the horizontal axis we measure the ratio of the number of jobs held by people who attended college (including those who did not graduate, as well as those who attained post-graduate degrees) and the number of jobs held by people with no college experience. For example, in 1980 for each job held by a person with no college education there were 0.76 jobs held by people with some college education; and those with a college degree earned 39.9 percent more than those with a high-school diploma (i.e., the earnings ratio was 1.399). By 2015, the ratio of the number of jobs increased to 2.00, whereas the college premium increased to 72 percent. Overall, we observe a significant increase in the supply of skills as well as an increase in the college premium. However, the latter has slowed down somewhat in recent years.

Figure 11.9 helps understand the evolution of the labor market according to the comparative statics framework introduced in Chapter 7. Equilibrium $E_1$ represents the US economy at the turn of the cen-
tury. Since then, we’ve observed a shift in demand for skilled workers (e.g., workers with a bachelor’s degree). There has also been an increase in the supply of such workers (as one can check from college enrollment statistics). However the increase in demand dominates in terms of effects on price, here defined as the relative wage of college graduates with respect to non-college graduates (high school degree or less). In other words, the college premium results from a “race” between technology, which pushes the demand for skilled labor to the right, and schooling, which pushes the supply of skilled labor to the right. In the past few decades the demand shift effect has dominated the supply shift effect.

To summarize the discussion so far, we have seen that rising income inequalities may be partly attributed to ownership of physical and financial capital, but likely more important are differences in human capital, namely in qualification levels. This makes access to advanced education all the more important as a means for moving up the income ladder. We will return to this in Section 13.3. It also raises the question of how much of the skill-bias of technical progress is a matter of “fate” or rather a result of economic incentives? We will return to this in Section 12.3.

Going back to Figure 11.6: Even though, as we argued before, most of the business income of the top 10 percent corresponds to labor income (highly qualified labor), it still remains the case that it
is earned through corporations. Moreover, Figure 11.6 shows quite clearly that the top percentile receives a considerable part of their income in the form of dividends and capital gains. In fact, dividends and capital gains are almost exclusively concentrated in the top 10 percent of the income distribution. To be precise, I must add that this is not exactly true: a good chunk of retirement funds include stock holdings, and these presumably extend to some of the lower income brackets. That said, it seems fair to say that an increase in firm profits benefits higher income brackets disproportionately. In this sense, the increasing market power trend reported in Chapter 8 is likely to have contributed to increased inequality.

But there’s more: The market power wielded by corporations is manifested not only in higher prices charged to the consumer — as reported in Chapter 8 — but also in the form of lower wages paid to workers. A significant trend in the US has been the weakening power of labor and the parallel increase in monopsony power wielded by corporations. Figure 11.10 shows the employment-weighted average of the Herfindahl-Hirschman Index (HHI) of employment by firms computed at the county-three-digit industry-year level. (The HHI is defined as the sum of the squares of all market shares. As the figure shows, it has steadily increased since the 1970s.

In addition to greater concentration on the employer side, there is also evidence of collusive behavior by employers. For example, in 2010 the US Department of Justice initiated an investigation into “no poach” agreements among high-tech companies (Adobe, Apple, Google, Intel, eBay) and animation companies (Pixar, Lucasfilm). The allegation was that these firms would agree not to “cold call” other companies’ highly skilled employees. The DOJ investigation resulted in various civil settlements on behalf of employees whose wages were likely lower as a result of the agreement among employers.

On the other side of the labor market, we observe that, due to legal and political changes over the past 40 years, only about one in 10 American workers now belongs to a union. For example, when the city of Seattle tried to facilitate collective bargaining for Uber drivers, the US Chamber of Commerce — with the cooperation of the Department of Justice and the Federal Trade Commission — sued the city based on the Sherman Antitrust Act (cf Section 8.2) and compelled the city to back down. The decline in union power is relevant
11.1. MEASURING AND EXPLAINING INEQUALITY

Since the Industrial Revolution, we have repeatedly observed the pattern of increased automation, whereby capital substitutes for labor. What is the evidence regarding the recent wave of industrial robots? One pattern that seems common across all adopters is that production workers get laid off, tech workers get hired, productivity increases, and output expands.

Data from France suggests that only a small fraction of manufacturing firms adopt automation, typically the larger firms. The share of production workers in employment declines, but output increases by so much that total employment actually increases. So far, so good, but this increase in employment is at the expense of competitors who do not adopt automation and experience a sharp decline in output and employment. Overall, robot adoption in a given sector is associated with lower employment.

In terms of wages, data from Denmark suggests that average real wage increases by .8 percent but wages in manufacturing drop by 6 percent. Welfare losses are particularly concentrated among older workers, who have fewer options to switch into the tech sector. Overall, industrial robots can account for a quarter of the fall in the employment share of production workers and 8 percent of the rise in the employment share of tech workers since 1990.

Box 11.1: Robots, jobs and wages

Since the Industrial Revolution, we have repeatedly observed the pattern of increased automation, whereby capital substitutes for labor. What is the evidence regarding the recent wave of industrial robots? One pattern that seems common across all adopters is that production workers get laid off, tech workers get hired, productivity increases, and output expands.

Data from France suggests that only a small fraction of manufacturing firms adopt automation, typically the larger firms. The share of production workers in employment declines, but output increases by so much that total employment actually increases. So far, so good, but this increase in employment is at the expense of competitors who do not adopt automation and experience a sharp decline in output and employment. Overall, robot adoption in a given sector is associated with lower employment.

In terms of wages, data from Denmark suggests that average real wage increases by .8 percent but wages in manufacturing drop by 6 percent. Welfare losses are particularly concentrated among older workers, who have fewer options to switch into the tech sector. Overall, industrial robots can account for a quarter of the fall in the employment share of production workers and 8 percent of the rise in the employment share of tech workers since 1990.

not just for union members but also for non-union members due to the so-called threat of unionization: When unions are powerful, non-union corporations tend to treat workers better so as to avoid inducing the work force to unionize. A related labor market trend is that, at the US national level, more than a third of US workers now have part- or full-time arrangements related to the gig economy. Finally, at the bottom of the wage distribution one important factor has been the drop in minimum wage in real terms (that is, in most states minimum wage has increased less than inflation). All in all, in terms of income inequality, this is a double-whammy: not only are wages lower, which harms the lower-income brackets, but firms profits increase, which benefits the higher-income brackets.

To conclude this section, we should mention the role played by so-called superstar effects. These effects are quite pervasive, as I will
argue in the next paragraphs, but they are particularly evident in the entertainment industries. Soccer fans of a certain age (football fans, if you live outside of the US) will appreciate the following comparison. Up until the 1980s the reach of cable and satellite television was very limited (and there was no Internet). For this reason, national football leagues (in England, Spain, Italy, etc) essentially catered to a local market. The main revenue source was gate receipts, with merchandizing and TV rights representing a very small fraction. In this context, the revenue ratio between, say, the English and the Portuguese leagues was about a factor of two or three, reflecting the higher purchasing power of English fans and the larger average size of their stadia. The advent of cable TV and globalization had a dramatic effect: Now everyone in the world can be a fan of Liverpool or Manchester United and follow their games live. Put yourself in the shoes of a soccer fan in Shanghai. Once you have access to TV and the Internet, you can access any league you like. If you are to follow a non-local league, you might as well follow the best. Granted, not everyone agrees on what the best is (England? Spain?), but most would agree that a small set of leagues is more competitive than the rest. Finally, this implies that the gap between the top leagues and the rest widens considerably with respect to the era of national leagues with national reach. Currently, the earnings of the English league are approximately 15 times bigger than those of the Portuguese league (up
11.1. MEASURING AND EXPLAINING INEQUALITY

This is not to say that the Portuguese league is not good or competitive. In fact, it is one of Europe’s top 10 leagues. However, in the global digital era the difference between the top spots and the near-the-top spots is much greater than it used to be. In other words, globalization and digitization together have created strong superstar effects, which in turn result in increasing inequality among earners.

Another way of measuring these superstar effects is by estimating the ratio between top earners today and top earners in the past. Even correcting for overall growth of the sector, top talent earns disproportionately more. This can be found in various segments of the sports and entertainment world, as illustrated in Figure 11.11. For example, in 1953 Humphrey Bogart was paid $300,000 for his role in *Sabrina*. It was the most he was ever paid for a role. Based on simple calculations I would expect that, in today’s money and considering the growth of the movie industry, this would correspond to about $7 million today. This is considerably less than an actor like Tom Cruise charges for a movie role. John Grisham, the leading writer in the late 20th century, sold about 60 million copies in one decade. Fast forward to the 21st century and J K Rowling sold more than 600 million copies in one decade. Similar comparisons could be made between The Beatles and Madonna (rock music), Pele and Ronaldo (soccer). In fact, the same superstar phenomena is also found in professions such as management (contrast GM’s Alfred Sloan and GE’s Jack Welch) or
11.1. MEASURING AND EXPLAINING INEQUALITY

architecture (contrast Le Corbusier and Frank Ghery).

A recent study brings these ideas to the data. First, it classifies industries along two dimensions. First, how skill-intensive they are, as measured by the percentage of workers with higher education. Second, how tradable the industries output is, as measured by the ratio (exports + imports)/output. Examples of sectors that are skill intensive but not tradable include education and health. Examples of sectors that are tradable but not skill-intensive include transportation, warehousing, wholesale trade. Examples of sectors that are neither skill intensive nor tradable include accommodation and food services, retail, administrative support, waste services. Finally, the important quadrant is that of skilled tradable services (STS), corresponding to sectors that are skill intensive and tradable. Examples include information, finance, insurance, management of companies, professional services. These are the sectors where superstar phenomena are more likely to take place. In fact, the evidence shows that, from 1980 to 2015, wages in STS increased by a factor close to 5, whereas in other sectors the factor was between 3 and 4.

Income inequality is caused by the concentration of capital, including in particular human capital. Other important factors include skill-biased technological change, market power, and superstar effects.

The superstar effect on income inequality is also felt at the level of firms. A recent study suggests that as much as two-thirds of the rise in wage inequality occurred due to a rise in the dispersion of average earnings between firms, as opposed to within firms. In other words, it’s not that the top and bottom salaries within firms are increasing, rather that some firms are paying much higher salaries than other firms. Why? Largely because superstar firms disproportionately attract higher-skilled workers. These firms pay more so as to recruit top talent or to add workers with sought-after skills. This concentration of talent further fuels the companies’ momentum, turning them into large giants with rapidly rising productivity. Meanwhile, employees in less-successful companies continue to be poorly paid and their companies fall further behind.

Another study suggests that industry sales are increasingly concentrated among a small number of firms. These superstar firms are so much more productive than the average firm that, while they pay
We have seen a lot of anti-inequality social protests in recent years, from Occupy Wall Street to Paris’ Yellow Vests to Santiago’s Metro protests. In California, some of this animus is targeted at specific “superstar” firms such as Apple or Google. In December 2013, for example, as Google employees boarded the bus that would take them on their daily commute, a group of protesters showed their anger against the tech giant’s privileged employees. The pamphlet they were distributed read:

In case you’re wondering why this happened, we’ll be extremely clear. The people outside your Google bus serve you coffee, watch your kids, have sex with you for money, make you food, and are being driven out of their neighborhoods. While you guys live fat as hogs with your free 24/7 buffets, everyone else is scraping the bottom of their wallets, barely existing in this expensive world that you and your chums have helped create.

One of the protesters summarized the crowd’s mood when they unfurled a banner bearing the words, “F—Off, Google.” One protester even smashed the whole of the bus’ rear window.

Better, the share of labor is lower than average. In other words, Facebook pays higher-than-average wages, but the share of Facebook’s sales that corresponds to labor is lower than the national average, so much so that Facebook’s generous wages don’t contribute much to the aggregate labor share. But this then implies that a concentration of economic activity in this small set of superstar firms has the effect of lowering the overall labor share. In sum, superstar firms, though they treat their own workers better than average, may be part of the trend whereby the average worker gets a smaller cut of the overall pie. Box 11.2 suggests that this phenomenon has not gone unnoticed.

Understanding the sources of increased inequality is an important step toward finding the best remedies. For example, a wealth tax — which we will discuss in Section 12.3 — has been proposed as a means to counteract the increasing concentration of wealth (and...
income). The perception that capital innovation has been biased against low-skilled labor has led to calls for a “robot tax”, something we also discuss in Section 12.3. To the extent that market power is a central force toward inequality, the discussion on antitrust presented in Section 8.2 is particularly relevant.

**11.2. DISCRIMINATION**

Discrimination is the act of making distinctions between human beings based on the groups, classes, or other categories to which they belong or are perceived to belong. Examples include age, disability, national origin, race/color, religion, sex. Discrimination is a wide-ranging phenomenon, and we do not have the space — or the ability — to cover all of its forms. Instead, in this section we restrict primarily to the economic roots and the economic effects of discrimination. Discrimination is one of causes of economic inequality — including, but not restricted to, income inequality. For example, on average women in the US earn 12 percent less than men, and the percentage of African-Americans not working is twice as large as that of white Americans.

As mentioned in Chapter 1, economists distinguish between statistical discrimination and taste discrimination. An example of statistical discrimination is provided by car insurance rates. A 16-year-old woman pays an average 6-month premium of $3,378, whereas a 16-year-old man pays a higher $3,897. The fact that women are charged less does not reflect any animus against men, simply the fact that, statistically speaking, 16-year-old men are more prone to accidents than 16-year-old women. Similarly, the fact that a 24-year-old man pays on average $1,381 — substantially less than his younger counterpart — does not necessarily reflect age discrimination in the sense that the expression normally has. By contrast, the differential way that racial minorities are treated in the labor market frequently goes beyond statistical outcome expectations, rather reflecting a distaste for hiring minorities. For the remainder of this section we will be primarily concerned with the non-statistical part of discrimination (gender, age, race, or otherwise based).
THE GENDER PAY GAP

One of the most important instances of discrimination is gender discrimination; and one of the most commonly documented forms of gender discrimination is the gender pay gap: women earn less than men. A simple way of measuring the gender gap is to compute the ratio between the average wage earned by women and the average wage earned by men. Figure 11.12 plots, for each detailed occupation, median earnings of men (horizontal axis) and women (vertical axis). Each dot in the figure corresponds to a particular occupation. For example, in the “credit counselors and loan officers” sector men earn on average $1,443 per week, whereas women earn on average $948 per week. As can be seen, almost all points are located below the main diagonal, the line such that $y = x$, that is, the equal-pay line. On average, the relation between men’s and women’s salary is given by the red line, the slope of which is .811. This means that in the US and on average, for every dollar that a man earns a woman earns $100 - 81.1 = 18.9$ cents less. This is the easy part of the analysis: the more difficult part is to decompose those 18.9 cents into its various components.

Claudia Goldin, one of the world’s leading experts on the matter, argues that this number “answers a particular question but it doesn’t say that men and women are doing the same thing. It doesn’t say that
they’re working the same amount of time, the same hours during the
day, or the same days of the week.” In other words, one part of the
wage gap is due to the fact that women self-select into different types
of jobs (including in particular lower-pay jobs); and that, even within
the same job, they work different hours or choose a different career
path.

Specifically, the US evidence shows that recent graduates start
at about the same salary level, with salary gaps arising later dur-
ding one’s career. Moreover, only about 15 percent of the emerging
pay gap is due to self-selection into types of jobs (e.g., more women
choose to be grammar school teachers, and these jobs pay less than
average). In other words, most of the gender pay gap (85 percent) oc-
curs within professions and as a career proceeds. One reason for this
age-increasing pay gap seems to be that the pay-worktime relation
is not linear: employees are paid more than proportionately for their
willingness to work long hours (and be flexible about it), and men are
more likely than women to choose such deals. One may be tempted
to say that this is just another instance of self-selection (“women earn
less because they choose to do so”). However, as we will see below,
even this self-selection effect may correspond to a form of discrimi-
nation.

IDENTIFYING DISCRIMINATION

Similarly to climate change and other “hot” topics, there are those
who deny that taste discrimination plays an important role, claiming
that the differences we observe are either due to statistical discrimi-
nation or result from selection and aggregation biases. For this rea-
son, the work of various economists has been instrumental in identi-
fying actual bias in the way women and minorities are treated. One
particularly interesting example refers to American symphony or-
chestras. By as late as 1970, about 10 percent of musicians in each
of the US “big five” orchestras were women. (The “big five” are the
Boston Symphony Orchestra, the Chicago Symphony Orchestra, the
Cleveland Symphony Orchestra, the New York Philharmonic, and
the Philadelphia Orchestra.) Economists Claudia Goldin and Cecilia
Rouse state that the extent of bias against female musicians was con-
siderable and likely responsible for such small percentages.
Many of the most renowned conductors have, at one time or another, asserted that female musicians are not the equal of male musicians, claiming that “women have smaller techniques than men,” “are more temperamental and more likely to demand special attention or treatment,” and that “the more women [in an orchestra], the poorer the sound.”

In July 1969, at the height of the civil rights movement, two black musicians — a double bassist and a cellist — accused the New York Philharmonic of racial discrimination. Although the musicians lost their case, various symphony orchestras gradually introduced blind auditions as a means to avoid bias in hiring. Since the introduction of blind auditions was gradual — that is, different orchestras did it at different times — the historical data provides a strategy for estimating sex-bias in hiring new musicians: Did the percentage of female musicians increase when blind auditions were introduced? By 1970, about 10 percent of orchestra members were female. By the mid-1990s, that percentage was up to 35 percent. Goldin and Rouse estimate that about 30 percent of this gain was due to blind auditions. This vindicates the claim that discrimination against women was present when the candidate could be identified by the jury. Many other similar studies provide compelling evidence that some form of discrimination has been present, and continues to be present, in various industries.
11.2. DISCRIMINATION

IMPLICIT DISCRIMINATION

Many social psychologists believe that an individual’s attitudes occur in both implicit and explicit modes. In other words, people can think, feel, and behave in ways that oppose their explicitly expressed views. One application of this dichotomy between explicit and implicit attitudes regards discrimination, namely racial discrimination. Specifically, we may refer to implicit discrimination as the unconscious mental association between a target (such as an African-American) and a given attribute. The evidence suggests that these associations exist and moreover are correlated with actual behavior patterns. In one much-discussed experiment, economists Marianne Bertrand and Sendhil Mullainathan sent a bunch of resumes to potential employers. The experiment was designed such that a given list of qualifications was present both in a resume with a white-sounding name (e.g., Emily or Greg) and in a resume with an African-American-sounding name (e.g., Lakisha or Jamal). Having sent a large number of randomly selected resumes to potential employers, the experiment tests the possible association between a name, a perception of race, and the weight given to objective qualifications listed on the resume.

The results are remarkable: White names receive 50 percent more callbacks for interviews. Moreover, this difference is relatively uniform across occupation, industry, and employer size. The one dimension on which the difference is not constant is qualification level: a higher-quality resume widens even more the gap between whites and African-Americans. What makes these results particularly remarkable is that the experiment was carried out in Boston and Chicago, two cities where the population and employers frequently consider themselves free from racial prejudice. In fact, the racial gap is the same if we restrict to employers who explicitly state that they are “Equal Opportunity Employers.” (In Chicago, the racial gap is slightly smaller when employers are located in more African-American neighborhoods.)

As we saw in Chapter 2, the basic economics model assumes rational, logical agents who take actions in the pursuit of individual welfare maximization. There are several reasons why this is a simplifying and limited set of assumptions. One is given by implicit attitudes, attitudes that we are not conscious of.
In the previous section we saw how economics has much to learn from social psychology when it comes to the issue of discrimination. Sociology, another related field of social research, can also be quite relevant. Unlike economics, which focuses on the preferences and choices of individual agents (households, firms, etc), sociology places a greater weight on structural and institutional factors. Consider again the gender pay gap. As we saw earlier, much of the gap can be accounted for by different career paths followed by men and women. In this sense, creating an egalitarian workplace might require reducing the cost of offering workers flexibility in their schedules. In other words, it’s not just that there is or might be an animus against women in the workplace but rather that the “system” is set up in a way that primarily addresses the needs and preferences of men.

More generally, the idea is that, even if no individual person pursues discriminatory actions, the system is such that women and/or minorities are at a considerable disadvantage. Consider the following story which, as they say in Hollywood, is inspired by true events. A recent male economics PhD graduated from a good but not top school. His work was of high quality, but not having graduated from the very top schools makes it more difficult to get the attention of potential employers, especially leading economics departments. So our recent graduate decided to attend an economics conference where a number of leading economists were presenting. One day, after the formal sessions had taken place, our recent graduate heads for the bar and finds that Professor X, a famous faculty member at University Y, is sitting by himself drinking a beer and watching a baseball game on TV. Our recent graduate decides to risk it: he introduces himself to Professor X and asks if they can watch the game together. Professor X is happy to have some company. Now, if you have watched baseball before you know that it’s not a super intense watching experience: There are plenty of dull moments, which our recent graduate took advantage of to talk about his research. The rest is history: Professor X recommends that University Y look into our recent graduate’s work, who eventually was hired by Y.

What’s the moral of the story? The moral is that a woman would have found it more difficult to follow the same path as our recent
graduate: walking into a sports bar and introducing yourself to a man sitting by himself drinking beer is a far less common and socially accepted course of action for women than for men. It is possible that, were Professor X presented with equally good work by a recent female graduate, he would recommend that she too be hired. However, one must admit that certain institutions, customs, etc, lead to a “system” that effectively discriminates against women.

More generally, diversity traps may imply that women and minorities have difficult access to a given profession because of a bad chicken-and-egg equilibrium. For example, the number of female faculty in economics is remarkably small, considering the number of women who study economics at the undergraduate and graduate level. One explanation is lack of mentoring and role models. Female faculty find it difficult to be the only one or one of the very few faculty members in a given department, and end up moving away from this career path. In the end, there are very few women in academic economics because there are very few women in academic economics — the essence of the chicken-and-egg equilibrium. A similar argument — in some ways, a stronger argument — might be made with respect to racial minorities, for whom mentoring and the sense of belonging may be particularly important.

**AFFIRMATIVE ACTION**

The policy of affirmative action can be seen as a strategy to move away from “diversity traps”. The idea is that by bringing in a critical mass of minority students, it becomes considerably easier to successfully attract additional ones, for there will be enough mentors and role models to make a minority member feel accepted and have a sense of belonging. In this sense, affirmative action would be a temporary measure designed to break an undesirable chicken-and-egg equilibrium.

However, this is not the only — or even the main — argument in favor of affirmative action. A more common argument is that diversity is a value by itself. For example, going back to symphony orchestras, the argument has recently been made that blind auditions should be scrapped and replaced by a recruitment system that takes into account race, gender and other factors, so that eventually ensembles reflect the communities they serve. In other words, the system
(blind auditions) that has played such an important role in fighting gender discrimination (see previous section) may be an obstacle to racial diversity.

Ultimately, the issue for symphony orchestras, colleges, and other organizations is determining the objective function. If the goal is a single performance measure — e.g., average SAT score of the college entering class — then affirmative action is clearly detrimental to the college’s objective. But there is no reason to believe indicators such as the average SAT score should be the sole or even the main objective in selecting an entering class.

Even if the above considerations are understood, affirmative action remains a controversial system, as the recent Harvard admissions case shows. In 2014, a class-action suit against Harvard University claimed that the college discriminates against Asian-American applicants in its undergraduate admissions process. The claim was rejected by the court in 2019, but you can see the arguments on each side: On one hand, minority students have a harder time adapting to an environment where mentors and role models are rare. On the other hand, favoring such applicants must come at the expense of some other group or groups.

BELIEFS ABOUT BELIEFS

Earlier on we discussed the gender pay gap. In countries like Saudi Arabia the bigger problem is not so much the gender pay gap but rather the gender gap in access to the labor market. Less than 15 percent of the Saudi female population aged fifteen and above were employed in 2017. Moreover, only a small fraction of these 15 percent
worked outside of the home (around 4 percent of the female population). Is this a case of discrimination? If so, what are its roots?

Through the custom of guardianship, husbands typically have the final word on their wives’ labor supply decisions in Saudi Arabia. A recent study finds that the vast majority of young married men in Saudi Arabia privately support women working outside the home (WWOH). However, they substantially underestimate the level of support for WWOH by other similar men, even men from the same social setting (for example, neighbors). In some way, this is similar to the chicken-and-egg equilibrium mentioned before, with the difference that in the present case it’s all about beliefs: We’re stuck in the no-WWOH equilibrium because men believe that other men disapprove of WWOH — even though they don’t actually disapprove of WWOH. The situation where most people privately hold an opinion but incorrectly believe that most other people hold the contrary opinion is known as pluralistic ignorance.

Consider the following (actual) experiment. A group of 500 Saudi men from Riyadh are asked for their opinion about WWOH. 87 percent are in favor, but most underestimate the percentage of other men in favor. Now split the 500 sample in two and disclose the results to one half only. It turns out that even this simple correction of Saudi men’s beliefs has an important real effect: A few months after the initial survey, the subjects are asked to sign their wives up for a job matching mobile application specializing in the Saudi female labor market. (They are incentivized with a gift card.) The percentage of men who accept the deal is substantially higher if they were previously disclosed the survey results (36 vs 23 percent). In fact, the dif-
ference is particularly significant among those who had underestimated support for WWOH by a greater margin. In sum, the results provide significant support for the pluralistic ignorance narrative.

Women working outside the home is by no means the only instance of pluralistic ignorance. In 1968, most white Americans substantially overestimated the support for racial segregation among other whites. In other words, white Americans thought white Americans were more racist than they actually were. More generally, these examples show the importance of social norms as a form of institutionalized discrimination. Finally, this all begs the question: How can inferior social norms and pluralistic ignorance be an equilibrium, and so persist over time? One reason is that individuals are reluctant to reveal their private views for fear of social sanction. This is one of the reasons why freedom of speech and the ability to openly debate these social issues is so important. It is my hope that this section contributes to this ongoing debate.
<table>
<thead>
<tr>
<th>Key Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorenz curve</td>
</tr>
<tr>
<td>Gini coefficient</td>
</tr>
<tr>
<td>skilled-biased technical change</td>
</tr>
<tr>
<td>college premium</td>
</tr>
<tr>
<td>human capital</td>
</tr>
<tr>
<td>market power</td>
</tr>
<tr>
<td>monopsony</td>
</tr>
<tr>
<td>superstar effect</td>
</tr>
<tr>
<td>skilled tradable services</td>
</tr>
<tr>
<td>superstar firms</td>
</tr>
<tr>
<td>implicit discrimination</td>
</tr>
<tr>
<td>diversity traps</td>
</tr>
<tr>
<td>affirmative action</td>
</tr>
<tr>
<td>pluralistic ignorance</td>
</tr>
<tr>
<td>social norms</td>
</tr>
</tbody>
</table>
In much of this book we follow the economics approach of assuming that economic agents are primarily concerned with their own interests. As mentioned in Chapter 2, this is not a normative statement, rather a positive statement: the world would be a much better place if we cared more for each other! Fortunately, even as a positive statement the selfish behavior paradigm is more a point of reference than a reality. In this chapter, we consider individual (Section 12.1) and social (Section 12.2) attitudes towards others: fairness, income distribution, provision of goods, etc. The chapter concludes with Section 12.3, devoted to the economics of taxation, one of the main tools of economic solidarity.

12.1. FAIRNESS

As we saw in Chapter 2, the basic model of economics — the *homo economicus* model — assumes that agents are rational, individual maximizers (of their own interests). As 19th century economist Francis Edgeworth put it,

> The first principle of economics is that every agent is actuated only by self-interest.

But a few decades earlier, Adam Smith — arguably the founder of economics as a discipline — thought of human behavior as a complex
mix of rational, selfish motivation, on the one hand, and social virtue on the other hand. He wrote that

How selfish soever man may be supposed, there are evidently some principles in his nature which interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it.

Who was right, Smith or Edgeworth? In this section, we attempt to answer this question, or at least come closer to answering this question.

THE ULTIMATUM GAME

One approach to understanding human behavior is to organize a laboratory experiment. One of the most popular such experiments is the so-called ultimatum game. Here’s how it works: Player 1 is given $100 with a condition: she must offer to share it with Player 2. Specifically, Player 1 offers Player 2 a deal, $s$ for Player 2 and $100 - s$ for Player 1. Player 2 then decides whether to accept or to reject Player 1’s offer. If he accepts, then Player 1 gets $100 - s$ dollars and Player 2 $s$ dollars. If however he rejects Player 1’s offer, then both players get zero.

Figure 12.1 illustrates this game. (Since players clearly move sequentially, it is better to represent the game in tree form.) As usual, we look forward and reason backward in order to “solve” the model, that is, in order to predict the players’ choices. First, put yourself in the shoes of Player 2. Any $s$ value that Player 1 offers is better than nothing. Therefore, Player 2’s optimal strategy as a rational agent is to accept any positive offer. (If $s = 0$, Player 2 is indifferent between accepting and rejecting.)
Given the above, consider now Player 1’s strategy. Believing Player 2 to be rational, Player 1 expects that any positive offer will be accepted by Player 2. Therefore, Player 1’s optimal strategy is to offer the smallest value $s$ possible. (If there is a strictly positive probability that $s = 0$ will be rejected by Player 2, then Player 1 does not want to offer $s = 0$, rather a very small $s$.) In sum, the Nash Equilibrium of the game corresponds to the prediction that Player 1 offers $s \approx 0$ and Player 2 accepts such offer.

Evidence from laboratory experiments is at odds with the predictions from game theory and the conventional rational behavior assumption. While the many studies vary in terms of their precise predictions, the values in Figure 12.2 provide a representative sam-
ple of the experimental data. It corresponds to a laboratory experiment with Emory University undergraduate students. The top panel shows the frequency with which a proposal is accepted by Player 2. The simple theory of own-payoff maximization would predict a 100% probability of acceptance (except for an offer of zero). However, the observed probability of acceptance is well below 100%, especially for offers close to zero.

These data suggests that Player 2 derives utility from financial gain but also from “punishing” Player 1 for what Player 2 perceives as an unfair proposal. Experiment after experiment, this is one of the more common departures from the simple model of own monetary payoff maximization: economic agents derive utility from what they perceive is a fair outcome and are willing to sacrifice monetary payoff for the sake of imposing a lower payoff on players perceived not to act in a fair manner.

The bottom panel of Figure 12.2 shows the frequency with which Player 1 makes different offer levels. The Nash Equilibrium derived above would predict that Player 1 offers $s$ as small as possible, so we might say that the *homo economicus* model would essentially predict an offer of 0 with probability 100 percent. The observed frequency of offers differs from this: an offer of zero is only observed with 21 percent probability. However, considering that Player 2 does not behave according to his own payoff maximization we must consider that even a rational, own-payoff maximizing Player 1 should play differently than prescribed by the game’s Nash Equilibrium. The “Ultimatum game” exercise in the companion book of exercises is highly recommended as a complement to this section. The exercise suggests that, given Player 2’s behavioral pattern, Player 1 is better off by offering 30 to Player 2. In other words, $s = 30$ maximizes Player 1’s expected payoff if Player 1 expects Player 2 to behave according to the probabilities shown in the top panel of Figure 12.2. In this sense, the surprising pattern in the bottom panel of Figure 12.2 is that Player 1 is more “greedy” than a rational player should be (and sometimes, though less frequently, more generous than a selfish, own-monetary-payoff maximizer should do).
12.1. FAIRNESS

THE FAVOR-EXCHANGE GAME

The experimental literature is filled with studies addressing non-standard behavior by economic agents, that is, behavior which departs from own monetary payoff maximization. Concepts such as altruism, kindness, fairness, reciprocity, etc, are brought to bear on the results. One limitation of this literature is that, frequently, only one of the above narratives is presented. A better way to test theories is to run a “horse race” that allows multiple explanations to explain the data.

Consider the following intertemporal favor-exchange game. Each period, a pair of numbers is randomly generated. These numbers correspond to a potential payoff for Players 1 and 2. The sum of the two payoffs is always positive, but one of the values may be negative. For example, one player stands to gain $8 while the other stands to lose $3. The two players then simultaneously decide whether or not to accept the proposal. If any of the players vetoes the proposal, then both get zero during the current period. Otherwise, they receive the indicated payoff.

If the game were played only once, then the standard economic model would predict that a payoff pair such as \((8, -3)\) would not pass the veto test. However, if the game is repeated, then there may be Nash Equilibrium such that a player accepts losing 3 during the current period with a view at earning a higher payoff in the future. Independently of the game being repeated or not, a player may also accept losing 3 simply because he or she cares for the other player’s payoff and understands that a loss of 3 is less important than a gain of 8.

The data shows that in the indefinite repetition of the above game players exchange “favors” frequently (i.e., accept to lose money apparently for the other player’s sake). Figure 12.3 plots the various observations split into acceptances and rejections. Whenever both players receive a positive payoff, typically both players accept the proposal. This can be seen on the positive quadrant of the left panel of Figure 12.3 compared to the positive quadrant of the right panel of Figure 12.3.

Comparing the second and fourth quadrants in the left and right panels of Figure 12.3, we see that, whenever one of the players gets a negative payoff, the proposal is typically rejected (higher density on
the right panel). However, we observe a significant numbers of proposals yielding a negative payoff which are nevertheless accepted.

Another interesting feature of the data is that there are cases when one of the players rejects a proposal even though both players stand to receive a positive payoff. In other words, the first quadrant in the right panel (rejections) includes some observations.

There are various theories compatible with these observations. One is **altruism**: Player 1 accepts a negative payoff for the good of Player 2’s positive (and higher) payoff. A second narrative is what we might refer to as **intrinsic reciprocity** (related to the concept of fairness mentioned in the context of the ultimatum game). The idea is that if Player 2 was good to Player 1 in the past (in the sense that Player 2 accepted proposals that yielded Player 2 a negative payoff), then Player 1 derives satisfaction from being kind to Player 2. Finally, a third possible explanation is that players, selfish as they are, offer to help other players as part of a Nash Equilibrium of the repeated game they play. In other words, the favor exchange is a case of **instrumental reciprocity** (ultimately a form of self-interested behavior).

Careful analysis of the data suggests that instrumental reciprocity (a.k.a. forward-looking reciprocity) explains the lion’s share of variation in the data. From a testing point of view, an important source of differentiation between the three explanations is a treatment where
subjects are told, at the beginning of the last period, that this is indeed the last period. What we then observe is a significant decline in the acceptance of negative payoffs, an observation which is consistent with the instrumental reciprocity explanation but not with the intrinsic reciprocity or the altruism explanations.

In sum, the ultimatum and the favor-exchange games suggest that

Real-world economic agents value fairness, to the point that they are willing to sacrifice monetary value for the sake of fairness.

The experiments also show that the model of rational behavior has much to say about real-world economic agents: not only many of their actions are motivated by their own payoff maximization but also they are strategically chosen considering the others’ preferences for fairness. In other words, human behavior is a complex combination of selfish and unselfish behavior, very much along the lines of Adam Smith’s quote. We are a social species, we do not live as isolated individuals. We care for each other, from local neighbors to fellow human beings who live far apart. In the next section, we consider how societies are organized so as to address these feelings of solidarity, in particular in the economic sphere.

12.2. POLITICAL ECONOMY

Whenever we talk about caring about each other at a societal level, we talk about politics. This section is about politics, namely how economists think about politics and political doctrines.

THE PARETO FRONTIER

In Chapters 2 and 3 we introduced the concept of feasible set. In the context of agent choice, a feasible set corresponds to the set of possible individual outcomes (for example, Alexei’s set of possible combinations of leisure and course grade). The concept of feasible set is very broad. In particular, we can think of a society’s feasible set as the set of possible welfare combinations of individual welfare. Consider a society with two individuals, A and B. Suppose we can measure A and B’s welfare by means of some index, for example, the
total surplus each receives in equilibrium. Figure 12.4 illustrates this idea. In it, we plot A’s welfare on the horizontal axis and B’s welfare on the vertical axis. The shaded area shows the set of attainable combinations of A’s welfare and B’s welfare.

The frontier of this feasible set is called society’s Pareto frontier. The idea is that points like $E_1$ and $E_2$ are Pareto optimal outcomes or simply Pareto optima. This means that it would be impossible to improve A’s welfare without reducing B’s welfare, and vice-versa. A point like $E_3$, by contrast, is not Pareto optimal: by moving to $E_1$ or $E_2$, for example, we can make both A and B strictly better off.

A Pareto optimal allocation is an allocation such that we cannot increase on agent’s welfare without reducing another agent’s welfare.

In Chapter 7 we introduced the First Welfare Theorem. It states that in a competitive market the equilibrium levels of output and price correspond to the maximum total surplus. Figure 12.4 allows us to restate the theorem as follows.

In an economy where markets are competitive, the equilibrium distribution of welfare levels corresponds to a point along the frontier of the welfare feasible set, that is, a Pareto optimal point.

So, in terms of Figure 12.4, points $E_1$, $E_2$ and $E_4$ would correspond to equilibrium points of the (competitive) economy. Does this mean
that there exists more than one equilibrium? Yes: In Chapter 7, we implicitly assumed an initial set of endowments. In particular, we assumed that $A$ and $B$ were endowed with certain levels of capital, labor ability, cash, etc. These endowments lead to a series of supply and demand curves, which in turn lead to a series of equilibrium values of $p$ and $q$, which in turn lead to surplus levels for $A$ and $B$.

Suppose that we are initially at competitive equilibrium $E_1$. Suppose that we take some of $B$’s initial endowment and give it to $A$, and then let the equilibrium be “replayed”. Such reallocation of initial endowments would imply a different equilibrium from the initial one. To the extent that we transferred endowments from $B$ to $A$, we should observe a movement along the Pareto frontier in the SE direction — point $E_2$, for example. The new equilibrium is as efficient as the first one, that is, both $E_1$ and $E_2$ maximize the gains from trade (First Welfare Theorem), which can also be interpreted as stating that, at either $E_1$ or $E_2$, one cannot make $A$ better off without making $B$ worse off (Pareto optimality).

For a while now we’ve been referring to the First Welfare Theorem. You may have wondered — is there a second one? Yes! If, in terms of the feasible set, the First Welfare Theorem states that a competitive economy equilibrium corresponds to a point on the frontier, the Second Welfare Theorem states that

Any competitive economy equilibrium, that is, any point along the Pareto frontier, may be attained by an appropriate shift in initial endowments.

Similar to the First Welfare Theorem, a few notes are in order. First, the issue is not whether these theorems are right or not. They are formal statements that follow logically from a set of assumptions. In this sense they are right. The important question is whether they are relevant. As discussed in Part III, the argument can be made that competitive markets are the exception, not the norm. In this sense, it would appear that the First Welfare Theorem should be considered as a reference point, not an actual description of economic reality. Is the Second Welfare Theorem relevant? To this question we turn next.
POTENTIAL AND ACTUAL PARETO MOVES

One of the first examples — perhaps the first example — of gains from trade is Adam Smith’s analysis of wine and wool exchange. Smith observed a fairly obvious fact: Scotland produces excellent wool but horrible wine, and France, by contrast, produces lousy wool and excellent wine, as you surely tasted (the wine, that is). Consider an initial situation where no exports between France and Scotland are allowed. We might plot the initial situation in the Europe feasible set as point $E_3$. (For the sake of illustration suppose that Europe comprises France and Scotland only.) Note point $E_3$ is not Pareto optimal, it lies strictly below the Pareto frontier. Why? As Smith pointed out, it’s a pity that the French use French wool and the Scots drink Scottish wine. By an appropriate exchange, we might reach a new outcome so that everyone drinks French wine and wears Scottish wool. And everyone will be happier!

Everyone? Not necessarily so. There is this fellow in Scotland who used to produce wine — or what he called wine — and is now out of work (it’s hard to compete with French wine). In terms of Figure 12.4, we might say that opening the country to trade implied a movement from $E_3$ to $E_4$. This movement implies a huge gain for $A$ but a loss — smaller, but still a loss — for $B$. If markets are competitive, you might argue that we could and should make the move from $E_3$ to $E_4$ and, at the same time, distribute resources from $A$ to $B$ so that we end up — as per the Second Welfare Theorem — in a point like $E_2$, where both $A$ and $B$ are strictly better off.

The above example helps illustrate an important difference. A movement from $E_3$ to $E_2$ is called a strict Pareto move: it’s a move such that everyone is better off. By contrast, a movement from $E_3$ to $E_4$ is a potential Pareto move: conditional on a subsequent transfer from $A$ to $B$, the final outcome is also $E_2$, in which case both $A$ and $B$ are strictly better off. However, the latter move is based on a big “if”, namely the movement from $E_4$ to $E_2$. The distinction between a potential and an actual Pareto move is quite important form a political economy point of view.

The distinction between potential and actual Pareto moves is very central in the globalization debate. Pro-trade economists argue that Adam Smith’s principles apply as well to France-Scotland trade in the 18th century as they do to US-China trade in the 20th century.
Admittedly, there are Americans who stand to lose from lowering import tariffs to Chinese imports. Admittedly, opening up to trade leads us from $E_3$ to $E_4$, but surely we will be able to compensate $B$ such that all Americans are better off. The counter-argument, however, is that compensation rarely if ever takes place, partly because it requires public funds which are hard to come by. We will return to this below.

**ECONOMIC AND POLITICAL THOUGHT**

In Chapter 7 I noted that the Fundamental Theorem has played a very important role in political thought for the past two centuries or so. I should now add that the same is true, though to a lesser extent, about the Second Theorem of Welfare Economics. Together, these two results form the basis for the following general principle of political economy: To the extent that markets are close to competitive, market equilibrium leads to efficiency. Efficiency per se does not mean that the equilibrium is fair. For example, in terms of Figure 12.4 one might argue that $E_1$ is not a fair outcome, for $B$’s welfare is considerably greater than $A$’s. This is where the Second Theorem kicks in. The idea is to redistribute initial endowments — and cash is usually the easiest endowment to distribute — and let markets do the rest, namely, let markets lead us to a point on the Pareto frontier (this time one closer to a balanced outcome). In other words, the political thought based on the the above two theorems corresponds to the idea that we should create **equality of opportunities** — as measured by initial endowments — rather than **equality of outcomes** — which may be very different if $A$ and $B$ have different tastes.
This is by no means a point of general agreement. At the risk of slightly simplifying the various schools of thought in terms of political economy, I think there are four different camps one might consider.

1. **Libertarians.** Respect private property, respect markets, respect individual freedom (so long it does not harm the freedom of others). Concern for others is an individual attribute, not a social one.

2. **Market-based, socially concerned.** Use the welfare theorems as a guide. Government policy should be focused on redistributing endowments, letting markets do the rest.

3. **Regulated markets, socially concerned.** Most markets are not competitive (think Facebook, Apple, etc). As such, the first theorem fails frequently. This calls for public policy focused on regulating markets, also as a form of distribution.

4. **Socialists.** Both of the theorems of welfare economics fail frequently. This calls for public policy focused on market regulation and distribution policies largely based on public provision of goods.

If I had to characterize the economics profession, I would say the vast majority fall under 2) and 3), with a recent slight shift from 2) to 3). In other words, one might divide the spectrum of political and economic thought along two dimensions. The first one is the extent to which the state should push for redistribution. This ranges from zero, for an extreme libertarian, to maximally (equality), for an extreme socialist. The second dimension refers to the instrument for distribution. At one end (extreme market based) the state enforces monetary transfers (initial endowments) between its citizens, letting markets do the rest. At the opposite end (extreme socialism), the state directly offers goods and services to its citizens, thus largely replacing the market in that role.

Many (most?) of the hot-button economic issues in the current political climate may be characterized as positions within the above spectra. They are frequently correlated, in the sense that if you favor little distribution you are likely also to favor reliance on markets; and
if you favor massive distribution you are likely also to favor social provision of goods. However, the two dimensions are not perfectly correlated. In fact, one may argue that, compared to an average citizen, a typical economist tends to favor more distribution and more reliance on markets and market mechanisms.

Economics research (cf Section 12.1) suggests that there is “demand” for reducing inequality levels (that is, there is a desire for fairness even if at the cost of one’s own share). However, the debate about the precise desirable level of distribution is largely a matter of political debate. By contrast, the debate about the extent of state provision of goods and services is largely — though not entirely — an economics question. It is, to a great extent, a question of alternative means of achieving a given end (a fair distribution of well-being). This we discuss in the next section.

INCOME VS IN-KIND TRANSFERS

Should distribution take place primarily by means of money transfers or primarily be means of in-kind transfers, that is, by direct provision of goods and services? Market-oriented economists insist on the contrast between opportunities and outcomes. We are all different. For example, some people prefer a large house, whereas other people put more weight on their house’s location. Therefore, even if A and B’s surplus levels (an economic measure of well-being) are identical, these surplus levels may correspond to very different bundles (for example, A lives in a large house in New Jersey, whereas B lives in a tiny apartment in Manhattan). One size does not fit all. This view, which emphasizes differences in tastes, tends to favor income distribution rather than in-kind transfers. It’s not the state’s business to get into business.

The opposite view also has its merits. First, as mentioned above, many markets are not competitive, and the market equilibrium may lead to undesirable outcomes. One specific but very important example is given by health markets, as we saw in Chapter 10. This can be fixed by market regulation (e.g., legislation such as the US Affordable Care Act) or simply by making the state the main supplier (e.g., health services offered by state-owned facilities with state-employed personnel, as is the case in several European countries).
Second, as mentioned earlier, not all consumers are fully rational at all times. In this context, direct supply of products and services may be an effective way of nudging consumers in the direction society considers desirable. However, you can see how controversial this line of argument can be, with more libertarian-leaning individuals warning against the dangers of “paternalism”, that is, a society that turns into a “big brother” or “nanny state”.

The above list of points in favor of in-kind transfers is longer than the pro-market doctrine of “given them cash and let the market work.” However, notwithstanding the many imperfections of many markets, the argument for separating resource distribution from resource allocation — the argument that the two welfare theorems correspond to two different societal tasks — is a strong argument and one that has had a deep influence in economic thought.

The welfare theorems suggest that cash transfers are preferable to in-kind transfers. However, market imperfection and limited consumer rationality may favor in-kind transfers.

Pro-market economist are frequently accused of being heartless. In some cases, the defendant is guilty as charged. In other cases, however, the accusation is based on a confusion (or conflation) of the roles of efficiency and fairness. When pro-market economists decry the effect that Hugo Chavez and Diego Maduro had on Venezuela since the beginning of the century they are not critical of the leaders’ concern for the poor, rather they are critical of the policies followed in trying to help the poor (unsuccessfully, many argue). There are ends and there are means. Most of the disagreement in matters of political economy stems from the latter, not the former.

EXAMPLES

Consider the case of schooling. Some people favor the system of public schooling whereby access to primary and secondary education is directly provided by the state (in the US, by each individual state). An alternative view is that education might be financed by the state but supplied by private entities which follow the public-school principle of accepting students without charging tuition. This
corresponds to the system of charter schools. Still another alternative view is that education might be financed by the state in the form of school vouchers which can be redeemed at privately-run schools. Still another alternative (less popular) view is simply for the state to let the market provide education services. (This is the case in many instances of specialized and advanced education, less so in the case of primary and secondary education.) All in all, the above possibilities span a vast range of solutions with greater emphasis on income transfers, direct provision of services, or none of the above. Can you see the mapping from these to the classification included on p. 393?

Consider now the case of pension systems. Many economists argue that retirement age should be made as flexible as possible. Some people are very eager to retire and put their newly-acquired free time to good use. Other people, by contrast, enjoy their work (and the payment that comes with it!). Let each decide what’s best for them: different strokes for different folks, as the saying goes. However, research suggests that about a third of US households headed by individuals older than 55 have no savings. In other words, a large (and growing) percentage of Americans are not prepared for retirement. For this reason, Oregon, Illinois and California have launched initiatives to create retirement savings accounts for residents whose employers do not offer company-sponsored programs. Basically, the government program takes money out of people’s paychecks (5 percent) and places it into a retirement account with a few basic investing options. This type of policy seems to have voter support. For example, a poll conducted in Oregon for the AARP found that 82 percent of state residents like the idea of an automatic savings plan, i.e., “big brother” taking care of your retirement plan.

One final, somewhat unusual, example, comes from Baldwin, Florida. For a variety of reasons, in 2018 the only grocery store in town shut down. This is a common problem with very small towns which do not have enough scale to justify the opening of a store. However, the case can be made that the consumer surplus created by a store is sufficient to justify (from a social point of view) a store opening. The solution was for the city hall itself to open its own grocery store!
UNIVERSAL BASIC INCOME

A specific application of the principle described above — “give them money and let markets do the rest” — is supplied by the Universal Basic Income (UBI) policy. The idea is to provide every citizen with a certain amount of income (and to finance this transfer with taxes that are paid primarily by people with higher income or wealth or both). In defining UBI it’s important to state what it is not. In particular, UBI differs from means-tested support, a welfare policy whereby income transfers are targeted to those who need them the most. During the 2020 presidential campaign, candidate Andrew Yang made UBI part of his platform. His proposal was to provide each American adult $1,000 per month (what he called the “Freedom Dividend”). This would be financed with a value-added tax of 10% (half of the typical rate in European Union countries). A tax expert claimed that a more realistic plan would require reducing the Freedom Dividend to $750 per month and raising the VAT to 22 percent, but even then you can see how the idea may attract pro-market economists with social concerns, that is, economists who think cash transfers are the best way of moving along the Pareto frontier.

Opponents of the UBI system point to its very high cost: why raise distortionary taxes in order to send Bill Gates a $1,000 check? A very different but equally important criticism relates to the emotional effect of cash transfers such as the UBI. Economist Robert Shiller explains:

We need to find a way to insure people against the risks of the global market without in any way demeaning them. ... When the government spends tax money on universal public education and health care, it does not strike many as redistribution, because the services are offered to everyone, and accepting them appears more patriotic than abject.

In this sense, UBI is clearly better than means-tested income support: having to pass a means test creates a social stigma that UBI does not, or does so to a lesser degree. According to Shiller and others, direct service provision is an even better means. In conclusion, not all economists agree when it comes to the means to achieve a given end.
Regardless of whether distribution takes the form of income transfers or direct provision of goods and services, the state needs resources to finance its activity. The main source of these funds is taxation. In this section, we cover some basic principles of the economics of taxation. For most of the section, we analyze, with the help of a supply-and-demand framework, the effects of a sales tax. However, the principles developed below apply to other forms of taxation as well.
12.3. TAXATION

TAX INCIDENCE

We begin with one of the most important principles in the economics of taxation: Ultimately, it makes no difference which party pays a tax on a given transaction.

*Consumer surplus and producer surplus from a given tax \( t \) is invariant with respect to the actual party responsible for paying the tax.*

In economics jargon, we say that the *incidence* of a given tax does not depend on who actually pays it. Figure 12.5 illustrates this idea. For the same set of supply and demand curves, the two panels consider two possibilities: in the top panel, a tax \( t \) paid by sellers (who therefore require a higher price for each unit sold); and in the bottom panel a tax \( t \) paid by buyers (who therefore are willing to pay a lower price for each unit sold). As can be seen, the prices effectively paid by buyers and sellers are the same in both cases, and so is tax revenue.

This first principle reflects an important idea in economic analysis: When estimating the impact of any policy, we must take into account how economic agents (e.g., buyers and sellers) will adapt to the policy in question. At some level, one might think that a buyer prefers that the tax be paid by the seller, not the buyer herself. However, once the seller is asked to pay the tax he will increase price accordingly. At the end of the day, the buyer will be need to spend the same amount, regardless of who is actually paying the tax. This is true beyond retail markets. For example, the US federal government collects payroll taxes from every job at the rate of 12.4%. Half of this amount is paid by the employer, half by the employee. Suppose the government were to decide that, from now on, all 12.4% are paid by the employer. Eventually this would be translated into a lower value paid to the employee and the net take-home would be the same.

There may be additional considerations that break the indifference between seller and buyer as the actual tax payer. Consider, for example, the case of a gasoline tax. It would be rather cumbersome to collect the tax from the millions and millions of buyers who fill up at the pump. For this reason, it makes a lot of sense for the government to collect the tax from sellers, not from buyers.
Finally, note that the above indifference result only applies to collecting from the buyer or from the seller of a given transaction. It does make a difference whether the government taxes a transaction in a market X or in a market Y. In fact, this is one of the main issues in the economics of taxation, determining which transactions the government should tax.

EFFICIENCY LOSS

As we saw in Chapter 7, the equilibrium of competitive markets maximizes efficiency, that is, maximizes gains from trade. To the extent that a tax on a market transaction has an effect on equilibrium price and output level, we conclude that such a tax implies a loss of efficiency. For example, in Figure 7.16 we measured the deadweight loss implied by a competitive market distortion. Taxes make up one of the main sources of such distortion.

\begin{quote}
Taxes imply a loss in allocative efficiency (distortionary effect of taxes).
\end{quote}

The top panel of Figure 12.6 illustrates this point. Initially (no tax), equilibrium is given by price \( p_1 \) and output \( q_1 \). Suppose a tax \( t \) is imposed, to be paid by the seller. This implies a shift in the supply schedule and a new equilibrium where price paid by consumers is \( p_2 \) and the net price kept by the seller is \( p_2 - t \). Given the price increase, output is now given by \( q_2 \). This implies that all trades from \( q_2 \) to \( q_1 \) fail to take place as a result of the tax. These are trades such that willingness to pay, given by the (inverse) demand curve, is greater than marginal cost, given by the (inverse) supply curve. The sum total of the value of trades lost is given by the area of the shaded triangle. This is the deadweight loss due to the tax.

One common measure of distortion implied by a tax is deadweight loss per dollar raised in taxes. The deadweight loss is the loss from missing trades: it’s the sum of seller and buyer surpluses (gains from trade) that are lost due to a market distortion. In other words, it’s a dollar value. The revenues raised through taxes are also a dollar value. We can thus compare the dollar revenue raised with the dollars lost due to the distortionary tax. This leads us a rather unfortunate principle:
A tax’s deadweight loss increases at an increasing rate. In other words, the deadweight loss per revenue dollar is increasing in tax rate $t$.

Let us rephrase this important principle in a more formal way. The value of a tax’s deadweight loss is given by the area of the shaded triangle on the top panel of Figure 12.6. This area is given by

$$\frac{1}{2} t (q_2 - q_1)$$

The difference $q_2 - q_1$ is greater the greater $t$ is. In fact, it varies approximately at the rate of $t$. This in turn implies that deadweight loss is given by

$$L \approx \alpha t^2$$
where $\alpha$ is a constant. In words, the deadweight loss varies approximately at the rate of $t^2$. That is, not only is deadweight loss greater the greater $t$ is, but the rate at which $L$ increases is itself increasing in $t$. This in turn implies that deadweight loss per unit of revenues raised is increasing in $t$. Figure 12.6 illustrates this by considering two different levels of $t$ (top and bottom panels). Specifically, the bottom panel considers a higher tax rate than the top panel. In each panel, the deadweight loss is given by the area of the triangle from $q_2$ to $q_1$. Tax revenue, in turn, is given by the area of the rectangle $t q_2$. As can be seen, the ratio between the area of the triangle (deadweight loss) and the area of the rectangle (tax revenue) is greater in the bottom panel than in the top panel. In the limit, as $t$ becomes very large, tax revenue goes to zero, and the ratio converges to infinity.

One important implication of this principle is that it may be better to spread taxation over many goods (e.g., single VAT rate) rather than targeting only a few. That said, one should add that there may be reasons why some goods are targeted by taxation. This is the case, for example, with Pigou taxes (e.g., carbon tax), that is, goods which are taxed due to negative externalities (cf Chapter 9).

Another important application of this principle is to the policy of opening up to trade (that is, reducing import tariffs), a policy we discussed in Sections 7.3 and 11.1. An import tariff is effectively a tax: its efficiency cost increases more than proportionately. Conversely, lowering an import tariff from a high starting value has a huge efficiency benefit. However, as we get down to low tariff values the additional gain is very small. Now, opening to trade has an important domestic cost: it hurts the workers in the sectors that are now subject to foreign competition. Compensating these workers for their loss requires government funds, and raising government funds, which implies an efficiency cost that is increasing as we saw above. So, to conclude, as we open up to trade the marginal benefit from tariff reduction gets lower and lower, whereas the marginal cost from compensating losers gets higher and higher.

**ELASTICITY AND THE EFFECT OF TAXES**

There are many reasons why you would want to tax certain transactions and not others. For example, in Chapter 9 we look extensively at the role played by Pigou taxes in counteracting negative external-
Carbon taxes, in particular, have been proposed by economists as a central tool to address the climate change crisis. Another important consideration is to minimize the distortionary cost of taxation. In the previous section we saw that, everything else constant, we might want to “spread the wealth”, that is, tax everything at the same rate. However, everything is not constant. In particular, different transactions correspond to different supply and demand curves. In this context, one important principle is that

\[ \text{The deadweight loss of a tax is greater the greater the price elasticity of demand.} \]

This is illustrated by the two panels in Figure 12.7. Both panels depict
the same supply curve and the same tax (to be paid by the seller). The difference between the two panels is that the demand curve is steeper (more inelastic) in the top panel than in the bottom panel. As can be seen, the area of the deadweight loss triangle is greater in the bottom panel, that is, the more elastic demand is, the greater the deadweight loss.

This principle suggests that, if minimizing deadweight loss per dollar of tax revenue is the goal, then it is better to tax goods with inelastic demand. Note, however, that there may be other reasons to tax goods with inelastic demand. For example, so-called sin taxes (e.g., taxes on cigarettes, alcohol, gambling) are targeted at goods with typically very low demand elasticities. However, the reason why these goods are targeted by a tax is the goal of “nudging” consumers away from those products, products which happen to have inelastic demand.

Demand elasticity matters for the size of a tax’s deadweight loss. It also matters in terms of incidence.

*The incidence of a tax on consumers is greater the lower their demand elasticity.*

Again, the two panels in Figure 12.7 illustrate this contrast. When demand is more elastic (bottom panel), most of the impact of a tax $t$ consists of a lower net price received by the seller. In other words, consumers are relatively unaffected by the tax. This is intuitive: If demand is elastic, then consumers have options, alternative goods they can purchase instead of the good in question. As such, they are relatively immune to a tax. Obviously, the situation would be different if all goods — including the alternatives — were taxed.

One implication of the above principle is that, if the government wants to protect consumers, then it should primarily tax elastic demands. However, this may come at a high cost in terms of deadweight loss. In fact, as we saw earlier, the more elastic demand is, the greater the deadweight loss. (By now, you may have realized that there is always a “but” in economics.)

Finally, although Figure 12.7 refers to changes in demand elasticity, the principles also apply to supply elasticity. For example, if supply is very inelastic (i.e., the level of supply is not sensitive to price changes), then (a) the deadweight loss of a sales tax is rela-
tively smaller and (b) the incidence of the tax will fall primarily on the seller.

**INCOME TAXATION**

So far, the discussion on the effects of taxation has been cushioned in terms of a sales tax. As mentioned earlier, the principles apply equally well to other types of taxes. For example, if the demand for labor is very elastic, then a tax on labor is likely to lower the net wage received by workers.

Notwithstanding the general nature of the principles on taxation, in this section we focus specifically on taxes on inputs, such as labor or capital. These are particularly important transactions in the economy and as such deserve a specific focus. In this context, one important principle is that

*Taxing revenues from production inputs (capital, labor) may create a disincentive for input supply.*

Figure 12.8 illustrates this idea. Specifically, it considers a worker’s choice between income and leisure, similar to what we considered in Section 4.2. Initially, there is no tax. The worker’s feasible set is limited by a straight line extending from (24, 0) to (0, 24 w), that is, from 24 hours of leisure and no income to no leisure and an income...
of 24 \( w \). Given the worker’s preferences (indifference curves), the optimal choice is given by point \( E_1 \), corresponding to \( l_1 \) hours of leisure and, therefore, \( 24 - l_1 \) hours of work. This corresponds to a consumption level \( c_1 \), which in turn is given by \( (24 - l_1) \cdot w \).

Suppose now that a tax \( t \) is levied on labor income. This implies a rotation of the budget line around (24,0). Specifically, if the worker chooses zero hours of leisure (and 24 hours of work) then (net) income is given by \( 24w(1 - t) \). Given the worker’s preferences, the new optimal point is given by \( E_2 \), corresponding to \( l_2 \) hours of leisure and, therefore, \( 24 - l_2 \) hours of work. This corresponds to a consumption level \( c_2 \), which in turn is given by \( (24 - l_2) \cdot w \).

Putting it all together, we conclude that the imposition of an income tax implies a decrease in labor supply from \( 24 - l_1 \) to \( 24 - l_2 \). This is the essence of the problem of taxing income: the greater the tax rate, the lower the incentives for agents to engage in the economic activity underlying income generation, which in the present example corresponds to work.

Some notes are in order. First, as discussed in Section 4.2, the effect of a price change on quantity may be positive or negative depending on the size of the income effect. Looking at the specific example of labor supply, the idea is that an increase in wage may imply an increase or a decrease in labor supply. In fact, empirical evidence suggests that, for low levels of \( w \), an increase in \( w \) leads to greater supply, whereas, for high levels of \( w \) an increase in \( w \) leads to lower supply (because the income effect of a higher \( w \) dominates the substitution effect).

Second, the relation between income taxes and input supply is a contentious issue. At a now-famous 1974 dinner, economist Arthur Laffer, arguing against President Gerald Ford’s tax increase, reportedly sketched a graph on a napkin illustrating that, for high levels of tax rates, an increase in tax rate leads to lower revenue. Although the
idea had been known for a long time, the term **Laffer curve** caught up. The idea is that the effect illustrated in Figure 12.8 can be so strong that the decline in input supply (e.g., labor supply) more than compensates for the increase in tax rate, to the point that tax revenues are lower.

The Laffer curve is certainly relevant for very high levels of the tax rate. In particular, if the tax rate is 100%, then the incentives for labor supply are limited, and it's likely that the value created is lower than it would be were tax rates lower than 100 percent. Consider, for example, a dentist who must decide how many days she wants to work each week. If the income tax rate is, say, 90%, then for each $100 she earns she only gets to keep $10. You can see how this would lead her to work less (for example, to work Monday to Wednesday instead of Monday to Friday).

What does the data suggest regarding the effect of income tax rates on input supply? Before answering this question, a very brief history of income taxation in the US is in order. Up to 1920, the (very small) federal government budget was paid primarily by import duties and liquor taxes. Prohibition eliminated liquor taxes and this led Congress to create an income tax as an alternative revenue source. Later, during the years of Roosevelt’s “new deal”, income taxation was extended considerably and so was the government’s role in distributing from those with greater income/wealth to those will less so. Today, the federal government plays a bigger role in transfers than in providing goods, which explains the high income tax rates.

Regarding the relation between these rates and economic growth, the US historical evidence is not very clear. Table 12.1 shows that, historically, income tax rates have increased considerably: from 2.4%
at the turn of the 20th century to 10.7% through mid-20th century to 17.2% during late 20th century and early 21st century. The GDP growth rate, however, has remained steady at around 2%. Table 12.2 zooms in on the past 70 years. It shows that, compared the 1950–1980 period, several tax rates have been lowered in the past 30 years. The average top individual income tax rate, the average top estate tax rate, the average corporate tax rate — all declined from 50–80 percent to 34–47 percent. However, growth rates did not increase (as “predicted” by the Laffer curve), rather the opposite happened. That said, and as mentioned in Section 2.1, correlation is not causality, and absence of correlation is not absence of causality. In one case where we have evidence beyond pure correlation, a Congressional Research Service report found that the estate tax’s net impact on private saving is unclear — it causes some people to save more and others to save less — and that its overall impact on national saving, a critical determinant of the amount of capital available for private investment, is likely negative.

**CAPITAL VS LABOR TAXATION**

For most people, their primary job is also their primary income source. Some, however, also earn income from their businesses, the shares they own, the capital assets they sell, and so on. Different income sources are taxed differently, and this variety of treatments has both efficiency implications and distribution implications. Particularly important is the distinction between capital and labor taxation, the focus of this section.

If the government needs to raise $1, what’s the best way of doing so: taxing capital income or taxing labor income? From a distributional point of view, and to the extent that capital assets are more concentrated than labor income, there is an advantage in taxing capital rather than income. From an efficiency viewpoint, however, this is not necessarily the case. The economic theory on this fundamental problem is complex, long, and unfortunately largely unsettled. The results are very dependent on the particular assumptions regarding the structure of the economy and the behavior of agents (workers, consumers, firms, etc). Some argue that the substitution effect of capital taxation on savings is so large that capital should not be taxed at all. In other words, taxing capital would have such a negative effect
on savings and investment that, in the long run, it would amount to “killing the goose that lays the golden eggs.” However, others state that the income effect of capital taxation may be so large that it is optimal to set capital taxes in a progressive pattern.

In case you need to refresh your memory, the substitution and income effects of a price change are discussed in Section 4.1, but here’s the idea as these concepts apply to the case at hand. (This paragraph is a little technical and may be skipped on a first reading.) An increase in capital taxation amounts to a decrease in the “price” of consumption today. In fact, what I give up (in terms of future consumption) is lower the greater the tax on capital. The change in the “price” of today’s consumption has a substitution effect and an income effect. The substitution effect implies that I consume more today relative to future consumption. The income effect is the effect the price increase has on my real income. Since aggregate consumption is a normal good (more income implies more consumption), an increase in tax implies that I consume less today. If this income effect is sufficiently strong, then an increase in capital taxation leads me to save more (and consume less today). The idea is that I am so concerned about future consumption that I compensate for the increase in capital taxation by consuming less today. Sounds crazy but it’s actually quite possible.

So, the bottom line of the economic theory on capital taxation vs labor taxation is that there is no bottom line. Much research still needs to be done. And it gets worse: most prior analysis measures capital and labor as aggregate factors, but as we have seen multiple times neither capital nor labor are homogeneous production factors: A sophisticated AI system is not the same thing as a building or a sewing machine (but all count as capital). Similarly, a rocket scientist or an experienced trial lawyer are not the same as a worker with only basic instruction (but all count as labor).

How are capital and labor taxed in practice? Regarding labor, the answer is relatively easy. If you reside in the US, are unmarried and hold a job paying $50,000 a year, then you pay 22% in federal income tax; 12.4% in payroll tax (social security), half of which is paid by your employer on your behalf; and 2.90% for Medicare, half of which is paid by your employer on your behalf. As we know from the discussion on incidence, it really does not matter whether it’s you or your employer who pays: adding all up, a total of 37% is paid
to the federal government (and this excludes state and city government). This may seem like a lot, but it’s actually less than the average of OECD countries. Now suppose that instead of $50,000 you earn $150,000 a year. Then your income tax rate is a little higher, 24%, but your social security rate is actually lower: payments are capped at $8,537.40, which means the rate you effectively pay is about 5.7%. So, all in all, a 150k job is taxed at approximately the same rate as a 50k job.

Regarding capital taxation, the question is much more difficult to answer. There are multiple rules and exception and exemptions. All in all, I think it is fair to say that capital income is taxed at a lower rate than labor income. Partly for this reason and partly due to the skill-biased nature of technical change, a number of people — including, for example, Bill Gates — have proposed the creation of a robot tax. At some level, there is merit to the proposal: labor is heavily taxed, which implies that the system incentivizes the development of labor-saving innovations: automatic checkout machines, ATM machines, assembly-line robots, and so forth. The critics of the robot tax also have a point: It’s very difficult to define what a labor-substitution innovation is. In other words, most innovations fall somewhere in-between: they do substitute some workers but they also greatly enhance the productivity of other workers. It may simply not be a very practical tax, and ultimately may do more harm than good.

WEALTH TAXATION

Wealth taxation has become a hot topic in recent years, especially in the US, where former presidential candidates Sanders and Warren have made it as part of their political platforms. Sanders proposed to levy a 1 percent tax on wealth above $32 million, for married couples, and then slowly increase the tax for wealthier households: a 2 percent for wealth between $50 to $250 million; 3 percent for wealth from $250 to $500 million; 4 percent from $500 million to $1 billion, 5 percent from $1 to $2.5 billion, 6 percent from $2.5 to $5 billion, 7 percent from $5 to $10 billion, and 8 percent on wealth over $10 billion. Warren, who was advised by UC Berkeley economists Emmanuel Saez and Gabriel Zucman, proposed to levy a 2 percent tax on fortunes worth more than $50 million, and a 3 percent tax on fortunes worth more than $1 billion.)
The case for a wealth tax starts with the observation that, in the past few decades, we have observed a significant concentration of wealth among a small number of people. Moreover, at the current levels of wealth, interest rates, and estate tax rates we can easily fall into a trajectory where a new class emerges which lives off inherited wealth rather than acquired wealth.

To this, proponents of the wealth tax add that it’s the only really viable option. True, many economists would advocate a combination of a progressive income tax and an inheritance tax, rather than a tax on wealth. However, many of the super-rich actually have little income. For example, Warren Buffett and Mark Zuckerberg earn little more than they spend. For this reason, one may argue that

Their wealth increases as a result of capital gains, not saved income. And because such gains are taxable only when the corresponding assets are sold, their annual increase in wealth essentially escapes taxation.

Alternatively, one might think that an inheritance tax would eventually take care of this issue, but some believe that such a tax is politically unfeasible (“opinion polls consistently show that while economists love the idea, most voters hate it”). (In the US, the inheritance tax rate is very small and applies to a small number of inheritances.)

The case against a wealth tax is multi-dimensional. Similar to an income tax, one may argue that wealth taxes create the wrong incentives.

Imagine two very rich people. Joe is a spendthrift who buys yachts, fancy cars, diamonds, and other rapidly depreciating property. Because his spending habits keep his assets below the wealth tax threshold, he effectively is rewarded for his spending habits.

Jane makes the same annual income as Joe. But she is a saver and investor, prudently putting money away for the future or using her wealth to help create or expand businesses and job opportunities. Because she accumulates sufficient assets to be subject to the wealth tax, she effectively is penalized for saving and investing.
To this, one might add the claim that wealth taxes are unfair.

Most wealth has already been subjected to income and other taxes, perhaps multiple times. It doesn’t seem fair to the holders of that wealth to suddenly pay additional taxes on assets that they thought were in the clear, and such taxes would signal that previous policy has failed.

There are also implementation issues. Perhaps the most important one is tax competition. If country A creates a wealth tax, then wealthy citizens have a big incentive to move to country B. In this respect, Switzerland provides a useful example. While the wealth tax base is defined at the federal level, tax rates vary considerably across locations and over time. It can be shown that when two different cantons set different wealth tax rates we observe a movement of taxpayers to the canton with lower rates.

Finally, in order to tax wealth you have to first measure wealth. At some level, this seems doable: bank accounts, stocks, bonds, and other financial assets are relatively easy to measure. However, this only corresponds to a fraction (less than half?) of the total wealth in the US. What about all of the assets for which there is no market price (e.g., the Vermeer painting hanging in my living room)?

ADDITIONAL NOTES

A few final notes on principles of taxation. First, a principle we might refer to as the “second-best principle”. The above considerations and calculations typically assume that we start from an efficient market. But the world is more complicated: there are market distortions everywhere. In this context, taxes may actually create a “good distortion”. Examples include Pigou taxes and sin taxes. (In a similar vein, allowing for market power by unions, or setting a minimum wage, may correct for an existing distortion. Note: this is a very controversial idea.) Another important consideration for tax design is the ease with which taxpayers can avoid or evade a tax. This is particularly important when it comes to income or wealth taxes, but not exclusively. Finally, we must also consider the issue of a tax’s political feasibility. Nobody likes taxes, but there are taxes which people particularly dislike. (And, possibly, taxes which people love, especially if they apply to others.)
PHILANTHROPY

One additional argument against a wealth tax is that many (most?) wealthy people put their wealth to good use, in particular through philanthropy. A particular case in point is the Giving Pledge created by Bill Gates and Warren Buffett in 2010. In essence, this is a campaign to encourage extremely wealthy people to contribute a majority of their wealth to philanthropic causes. As of May 2019, it counted with 204 signatories (individuals or couples) from 22 countries. From the Giving Pledge website:

The goal is to talk about giving in an open way and create an atmosphere that can draw more people into philanthropy. ... The pledge does not involve pooling money or supporting a particular set of causes or organizations. ... The pledge is a moral commitment to give, not a legal contract.

Specifically, the (not legally binding) pledge taken by its signatories corresponds to the “commitment to give the majority of their wealth to address some of society’s most pressing problems.” Table 12.3 suggests that, if fulfilled, the pledges would correspond to substantial contributions, probably at levels at or above what a wealth tax would levy.

That said, one must also beware of the perils of billionaire philanthropy. Philanthropy may result in little more than self-serving policy advocacy. A number of wealthy donors (e.g., the Koch brothers) use their philanthropy to advance specific causes, which may or may not correspond to the will of most. An important share of char-
TABLE 12.3
Giving Pledge Top 10 by net worth (source: Giving Pledge)

<table>
<thead>
<tr>
<th>Name</th>
<th>Net worth ($b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill and Melinda Gates</td>
<td>101.0</td>
</tr>
<tr>
<td>Warren Buffett</td>
<td>87.4</td>
</tr>
<tr>
<td>Larry Ellison</td>
<td>62.5</td>
</tr>
<tr>
<td>Mark Zuckerberg and Priscilla Chan</td>
<td>62.3</td>
</tr>
<tr>
<td>Michael R. Bloomberg</td>
<td>55.5</td>
</tr>
<tr>
<td>MacKenzie Bezos</td>
<td>36.6</td>
</tr>
<tr>
<td>Azim Premji</td>
<td>22.6</td>
</tr>
<tr>
<td>Elon Musk</td>
<td>22.3</td>
</tr>
<tr>
<td>Jim and Marilyn Simons</td>
<td>21.5</td>
</tr>
<tr>
<td>Paul G. Allen</td>
<td>20.3</td>
</tr>
</tbody>
</table>

itable donations are directed at educational institutions. In this context, a danger posed by philanthropy is the distortion it may create in the education system: Some private foundations support highly exclusive schools in affluent school districts, thus compounding one important source of inequality and absence of social mobility.

But it gets worse (or may get worse): Recently, the US witnessed a major case of corruption in higher education, whereby wealthy parents bribed college coaches, test proctors, and others to rig college admissions, while funneling funds through tax deductible foundations. Also, some foundations pay family members to serve on boards, staff foundations, and subsidize family reunions, in the form of board meetings. In other words, sometimes philanthropy is nothing more than a form of self-dealing.

Obviously, this is not the case of all — or even most — philanthropic foundations. For example, the Gates Foundation has committed more than $7 billion to programs related to infectious diseases control and malaria control, an investment that would likely not be taken by any individual government.
KEY CONCEPTS

- laboratory experiment
- ultimatum game
- intertemporal favor-exchange
- altruism
- intrinsic reciprocity
- instrumental reciprocity
- Pareto frontier
- Pareto optima
- Second Welfare Theorem
- strict Pareto move
- potential Pareto move
- equality of opportunities
- equality of outcomes
- libertarians
- market-based, socially concerned
- regulated markets, socially concerned
- socialists
- nudge
- schooling
- pension systems
- Universal Basic Income
- incidence (tax)
- deadweight loss
- sin tax
- Laffer curve
Broadly speaking, we can think of two ways to solve the wide inequalities documented in Chapter 11. One is to redistribute income and wealth, as discussed in Chapter 12. An alternative approach is to create opportunities for upward economic mobility, the focus of this chapter. How important is a person’s starting point in terms of their life opportunities? Relevant factors include wealth, race and ethnicity, generational factors, family structure, and location (the country where they were born and the neighborhood where they grew up). The goal of economic opportunity is that a person’s options at birth be as much as possible independent of the above factors.

13.1. MIGRATION

Based on the media coverage we are exposed to, the word “migration” evokes images of refugee migrants. For example, in recent years millions of middle-eastern refugees fled war-stricken countries such as Syria seeking asylum in countries such as France or Germany. According to the UN, as of 2019 a total of 79.5 million of the world’s inhabitants — about 1 in 100 — were displaced for involuntary reasons. Table 13.1 displays the top source and the top host countries of refugees as of 2019. The source countries listed in Table 13.1 account for more than two thirds of the total. Three out of four refugees are hosted in neighboring countries, a fact that is reflected in a close par-
### MIGRATION AS ECONOMIC OPPORTUNITY

Figure 13.1 shows the income probability density for the world (blue) and the US (red). (See Section 11.1 for more details.) The first thing that stands out is the fact that US incomes are, by and large, higher than those of other countries. The second interesting observation is that the range of incomes in the world distribution is considerably higher than in the US. In the US practically no one has an income lower than a few thousand dollars. Elsewhere in the world your income may be as low as $100 a year or less. Income inequality is not just an issue within the US and within other high-income economies, in fact it’s even more important worldwide. In other words, if I had to predict a newborn’s economic prospects I would say that the place where he or she is born would explain a lot of the variation in predicted outcome. In a way, this is just a way of repeating the idea that migration is an important source of opportunity: The country where you are born may put you at a significant economic disadvantage with respect to people born in other countries, and migration may be a way of leveling that difference. In sum,
Migration is an important source of opportunity, namely economic opportunity.

As of 2014, 247 million people lived in a country not of their origin. Consistent with the idea of economic opportunity, broadly speaking migrants originate in developing countries and head to more economically developed economies. Table 13.2 shows the top 10 source countries and top 10 host countries in 2014. The US is, by a long shot, the country with the largest number of foreign-born residents: 47 million.

Migration has continuously grown over time. In the past few decades, it has essentially grown at the same rate as population has grown. However, the data shows that, between 2000 and 2018, the increase in foreign-born population accounted for more than three-quarters of the total population increase in European OECD countries, and for almost 40% of the increase in the United States.

From the data in Table 13.2 it’s easy to think of the US as the “land of opportunity”: the number of individuals who have moved to the US looking for a better life is astounding. However, in relative terms — that is, controlling for country size — the data paint a different picture. Figure 13.2 shows the number of foreign-born population as a percentage of the host country’s or host city’s total population.
TABLE 13.2
Top 10 migrant origins and destinations by the end of 2014 (source)

<table>
<thead>
<tr>
<th>Top migrant origins (million)</th>
<th>Top migrant destinations (million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>United States</td>
</tr>
<tr>
<td>16</td>
<td>47</td>
</tr>
<tr>
<td>Mexico</td>
<td>Germany</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Russia</td>
<td>Russia</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>China</td>
<td>Saudi Arabia</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Syria</td>
<td>United Arab Emirates</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Ukraine</td>
<td>Canada</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Pakistan</td>
<td>France</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Philippines</td>
<td>Australia</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Spain</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

for various countries and cities of destination. The contrast between Figure 13.2 and Table 13.2 is remarkable. Australia, which is a the ninth destination in absolute number, is the most significant in terms of percentage of domestic population. Figure 13.2 also shows that the relative importance of foreign-born population varies considerably from city to city, with Sidney, Miami and London being some of the world’s very large and very international cities.

Figure 13.2 is restricted to a selected number of (relatively large) countries. If we were to look at the whole sample we would find unusual cases such as Luxembourg. Luxembourg is a small European country with a mere 281 thousand foreign-born residents in 2018. However, this represents an enormous fraction of the domestic population, close to one half. In this sense, it would appear that Luxembourg is more of a “land of opportunity” than the US. Another interesting fact related to Luxembourg is that more than 80 percent of the foreign-born population was born in other European Union countries. The idea that migration is primarily a movement of individuals and families from developing to mature economies is not always accurate. That said, one must add that more than 30 percent of foreign-born Luxembourg residents were born in Portugal, which,
MIGRATION AS A MARKET OUTCOME

In various parts of the book I stressed the point that trade generates value, and that markets, by facilitating trade, create value too. To repeat the example considered in Chapter 7, suppose that Jane would be willing to pay up to $10 for a pound of golden delicious apples. It costs Old McDonald 50 cents per pound to grow apples. When Old McDonald sells one pound of apples for $3 per pound, the seller is $2.5 better off and the buyer is $7 better off. Trade can be a win-win operation.

Similarly, migration can in principle be a win-win situation. More generally, we can think about the world as a “market” where countries supply citizenship/residence and individuals demand citizenship/residence. To the extent that individuals have different preferences regarding location, political regime, economic opportunity, etc, there is the potential that movements of people across countries cre-
MIGRATION

ate value both for themselves and for the countries they move to. In addition to this market-based idea, one can also think of migration as a manifestation of an individual’s freedom. The pursuit of happiness frequently includes a choice of where to live, and we should strive to offer to all the ability to pursue their happiness.

Interestingly, this line of argument also suggests that differences across countries — for example, in terms of societal and political organization — may be a good thing. In Chapter 1 we talked about varieties of capitalism. More generally, we also observe varieties of democratic systems. Is the French presidential system better than the German parliamentary system? Is the more market-based US system better than the more socialist system of Denmark or France? In some sense, these may be the wrong questions, for they ignore the possibility that each system is more appropriate for the majority of each country’s residents. In this sense, rather than making all countries identical in terms of economic and political system, the world may be better off by allowing for differences across countries and, at the same time, providing the world’s citizens the freedom to choose between systems.

ECONOMIC EFFECTS OF MIGRATION

In Section 5.1 we introduced the concept of production function with two inputs. Table 13.3 provides a numerical example of such a production function when there are two production inputs, capital (K) and labor (L). For each value of K and L, Table 13.3 shows the value of output. In Section 5.1 we also introduced the concept of decreasing marginal returns of an input. As a reminder, consider for example the production function of restaurant meals. It requires two inputs, kitchen space (in square feet) and workers. Fixing the input “kitchen space”, you can see how the contribution of an additional worker would decline as more workers are added. In terms of Table 13.3, we can measure diminishing marginal returns by fixing K = 1, for example. The output added by the first, second, etc, worker is then given by 100, 41, 32, 27, 23, 21. Decreasing marginal returns.

Another feature of most production functions (including the one in Table 13.3) is that the marginal return from adding a worker (say, a second worker) is greater the greater the amount of capital. For example, as K changes from 1 all the way to 6, the added output
obtained from a second worker is given by $141 - 100 = 41$ if $K = 1$, $200 - 141 = 59$ if $K = 2$, ... all the way to $346 - 244 = 102$ if $K = 6$. In a way, this property is similar to diminishing marginal returns. Generally speaking, we might say that the output added by one worker is greater the greater the ratio of capital over labor, that is, the $K/L$ ratio.

What does this all have to do with migration? A lot: One of the main differences between developing and developed economies is precisely that developed economies are more capital intensive, that is, have a higher $K/L$ ratio. Going back to Table 13.3, it’s as if a migrant were to leave from a firm with $(K = 2, L = 4)$, so $K/L = \frac{1}{2}$, where that worker’s contribution was $282 - 244 = 38$, and moved to a firm with $(K = 5, L = 2)$, so $K/L = 2.5$, where the worker’s contribution is $387 - 316 = 71$. Following this line of argument, a McKinsey report estimates that, considering differences in worker productivity across countries, migrants contributed roughly $6.7$ trillion, or 9.4 percent, to global GDP in 2015 — some $3$ trillion more than they would have produced in their origin countries. North America captured up to $2.5$ trillion of this output, while up to $2.3$ trillion went to Western Europe.

The simple economics of production functions and input mix suggests that the overall economic gain from migration can be significant. But: are all parties better off? It seems reasonable to assume that migrants are better off. This is a classic example of what economists refer to as revealed preference: If migrants move of their own volition — and assuming that they are well informed about the conditions in
the host country — then the move signals that they are better off by migrating.

Regarding the country of origin, the question is not so simple. Going back to the previous example, the marginal contribution \(282 - 244 = 38\) is lost. However, the data shows that a substantial amount of the \(387 - 316 = 71\) created by the worker in the host country is sent to the country of origin in the form of cash remittances. In 2014 only, remittances totaled $580 billion (roughly 8.7 percent of the output generated by migrants). The largest inflows went to India ($70 billion), China ($62 billion), and the Philippines ($28 billion). In many cases, however, the losses may be greater than the gain. In fact, it is estimated that remittances correspond to about one half of the value migrants would have generated in their country of origin if they had not moved. This is especially true for high-skilled workers, the so-called brain drain problem, that is, the loss in human capital due to emigration. One study found that dozens of poor countries — mostly small countries in sub-Saharan Africa, developing Asia, and the tropics — were losing one third to one half of their college graduates.

Finally, the effects on the host country are generally estimated to be positive. Much of the economic contribution by immigrants is captured by their employers as well as by tax authorities. Moreover, the entrepreneurial and innovative effort by immigrants contributes significantly to the economic growth of their host economies. That said, when discussing the effects of immigration one cannot avoid the controversy regarding the impact on wages and unemployment: Aren’t immigrants simply replacing domestic workers and sending the latter to the ranks of the unemployed, or at least putting additional pressure to lower wages, especially wages among lower-skilled workers? A simple supply-and-demand analysis suggests that immigration, by pushing the supply curve to the right, leads to a new equilibrium with lower wage rates. In fact, one study estimates that a 10 percent increase in immigration leads to a 3 to 4 percent decline in wages.

However, two important remarks are in order. First, the standard supply-and-demand diagram assumes that labor is a homogeneous production factor. In reality, there are many different qualification levels, many different skills, etc. Earlier we saw that capital and labor may be complementary inputs, in the sense that a higher level of
capital benefits labor (specifically, increases the marginal product of labor). Something similar can happen between different types of labor. For example, it may be that native labor is better at tasks that require command of the local language, whereas immigrants are relatively better at tasks that do not require such language skills. To the extent that we have complementarity between types of labor, the effect of immigration could well be to increase the average wage of native workers. To give a simple example: Even Google, which hires a lot of high-skilled workers, needs janitors to clean up its buildings. If immigrants take up those jobs they will replace native workers. However, many of those workers might find a job as, say, checkout clerks, a job at which they might have greater relative advantage. Inevitably, here are adjustment cost to be paid, but it’s not impossible that all be made better off. In fact, one study that controls for type of work (as well as the particular commuting zone where immigrants are located) suggests that immigration has a zero or positive effect on the wages earned by native workers.

Naturally, the effects of immigration go well beyond the short-run effect on wages. As we saw in Chapter 2, immigration quotas, especially when applied to skilled immigrants, produce significant harm to the domestic economy. That’s a clear lesson from US economic history: Between 1921 and 1924 the US first adopted immigration quotas for “undesirable” nationalities, so as to stem the inflow of Eastern and Southern Europeans (ESE). It is estimated that, due to these quotas, 1,170 ESE-born scientists were missing from US science by the 1950s. This in turn led to a 68 percent decline in patenting in the fields where ESE immigrants researched. Moreover, these effects were still felt well into the 1960s. More generally,

The evidence broadly suggests that migrants of all skill levels make a positive economic contribution to their host country, whether through innovation, entrepreneurship, or freeing up natives for higher-value work.

Given this evidence, why is there so much resistance to immigration? First, as mentioned earlier, the benefits to the host country are not uniform across time or across the population. In particular, the case can be made that some domestic workers experience lower wages following an influx of immigrants. Second, the public percep-
13.2. INTER-GENERATIONAL MOBILITY

In the previous section we looked at the opportunity that an individual has of moving to a different country and thus improve their economic condition. Within a given country, a person’s opportunities also depend on his or her parents: We inherit genes, possibly...
13.2. INTER-GENERATIONAL MOBILITY

some wealth, and grow up where our biological or adoptive parents live. How important are these factors in determining an individual’s economic opportunities? A way of addressing these questions is to construct a mobility matrix. Figure 13.3 does so for the United States based on the 1980-82 birth cohorts. The way this matrix is constructed is as follows. First we divide the parents’ income into ten deciles: decile 1 corresponds to the bottom 10%, decile 2 to the second bottom 10%, and so forth, all the way to decile 10, which corresponds to the top 10% of the parents’ income distribution. Next, within decile 1 (bottom 10% of the parents’ income) we tabulate the children’s income level when adults. Specifically, we assign the child’s income to each of the 10 deciles of the overall distribution of children’s income. Instead of writing the fraction of children’s income corresponding to each decile, in Figure 13.3 we color each square in the first column according to the value of each cell. The fact that the (1,1) cell is very dark signifies that, conditional on the parent’s income falling into the first decile, it is very likely that the child’s income will also fall into the first decile. By contrast, the fact that the cell (1,10) is shaded white signifies that the probability that the child’s income falls into decile 10 (the richest) given that the parent’s income falls into the lowest decile is very close to zero.

More generally, Figure 13.3 suggests that inter-generational mobility in the US is rather limited. If the parent’s income is at or close to the bottom decile then it’s very likely the child’s income will fall at or close to the bottom decile. Conversely, if the parent’s income is at or close to the top decile then it’s very likely the child’s income will fall at or close to the top decile. In the limit when the child’s income ranking is exactly the same as the parent’s income ranking, we would observe the cells along the main diagonal in black and the off-diagonal cells in white. The pattern in Figure 13.3 differs from that extreme, but we do observe the general pattern that off-diagonal cells are shaded lighter than cells close to the main diagonal.

Another way of expressing the degree of inter-generational mobility — or lack thereof — is to correlate the child’s ranking (vertical axis) with the parent’s ranking (horizontal axis). As mentioned earlier, zero mobility would yield a relation like the blue line in Figure 13.3: the child’s rank is equal to the parent’s rank. At the opposite extreme, if your parent’s rank says nothing about your own rank, then the child’s expected rank is 50% regardless of the parent’s rank. This
13.2. INTER-GENERATIONAL MOBILITY

The so-called Great Gatsby curve: Income inequality and intergenerational elasticity (x source and y source)

corresponds to the green line in Figure 13.3. The actual regression based on US data corresponds to the red line. The slope of this line is given by .341. For comparison purposes, the slope of the same line estimated for Denmark is .180, for Canada .174. Both of these values are considerably closer to zero, that is, closer to the green line in Figure 13.3, that is, the line corresponding to equal opportunity. This suggests that inter-generational mobility is considerably lower in the US than in Canada or Denmark.

A similar statistic of inter-generational mobility is given by the rank elasticity: if we increase the parent’s rank by 1%, how much does the child’s rank increase. An elasticity of zero denotes equal opportunity: the parent’s rank has no effect on the child’s rank. At the opposite end, an elasticity of 1 corresponds to no mobility. One meta study combined several estimates for several countries and reached a series of meta-estimates for twenty-one countries. These are plotted in Figure 13.4, together with each country’s Gini coefficient, which, as seen in Section 11.1, measures the degree of income inequality.

Figure 13.4 suggests three observations. First, there is a positive correlation between inequality (Gini coefficient, on the x axis) and lack of mobility (rank elasticity, one the y axis). Norway, Finland and Denmark are countries with relatively little inequality and where your parents’ position on the income distributions says rela-
13.2. INTER-GENERATIONAL MOBILITY

Across countries, there is a positive correlation between measures of inequality and measures of inter-generational inertia. Within developed economics, the US shows one of the highest levels of both inequality and lack of mobility.

In other words, the US is much less of the “land of opportunity” that is often pictured. The difficulty of Americans rising above their parents’ economic conditions is, in some way, a new phenomenon. Figure 13.5, aptly titled “The fading American dream,” plots the probability that, at age 30, an American’s income is greater than his or her


In Section 13.1 we documented the significant variation in income levels across the world. A significant variation can also be found within each country. A series of studies by a group of scholars in collaboration with the US Census Bureau has led to a rich data set which allows us to visualize economic measures by each of 70,000 neighborhoods across America: the so-called Opportunity Atlas. The top panel in Figure 13.6 presents a heat map of household income levels in the US. In some parts of the country — colored in red — average household income is lower than $35,000. In other parts of the country — colored in blue — it is higher than $55,000. The data on the top panel is presented at the commuter zone level. However, as we zoom in on a particular area, we find that average household income varies at a very granular level. The bottom panel in Figure 13.6 focuses on New York City (and its environments), with data plotted at the Census track level, a much more granular division than commuting zone. In the top panel of Figure 13.6 New York City is colored green, as it is part of a commuting zone with above-average income. However, as the bottom panel shows, there are many neighborhoods
13.3. HOUSING, SCHOOLING AND FAMILY

FIGURE 13.6
Average household income by US commuting zone (top panel) and by New York City census track (bottom panel). Source: Opportunity Atlas

in New York city with income levels below $35,000.

A second important pattern we learn from the Opportunity Atlas is that the degree of mobility also varies greatly across the US. Figure 13.7 shows the probability of staying in the same commuter zone as an adult. In some commuting zones, such probability is greater than two thirds, whereas in many others it is less than one third — a
considerable variation. The striking feature about the spatial distribution of these probabilities is the high correlation with average income, as per the top panel in Figure 13.6: US commuting zones with lower household income (red in Figure 13.6) are also US commuting zones with higher inertia (blue in Figure 13.7).

There is considerable variation in income levels across US commuting zones. Moreover, within each commuting zone there is significant variation across Census tracks. Finally, there is a positive correlation across locations between income levels and the probability of moving away from the location where one grows up.

What explains these striking spatial variations? At first, one might think that location is simply a proxy for other variables, such as education level or race. However, Figure 13.8 replicates the heat map of the top panel of Figure 13.6 for black households only. Several commuting zones are colored gray, denoting that there is not enough data to make the probabilities statistically significant. With that qualifica-
FIGURE 13.8
Average black household income by US commuting zone. Source: Opportunity Atlas

...tion in mind, it is remarkable how highly correlated the color patterns in Figure 13.7 and the top of Figure 13.6 are.

Overall, Figures 13.6 and 13.7 suggest that the area where you grow up has much to say about your economic condition. Why is there such a variation across neighborhoods? Is it the schools? Is it the social fabric? And what causes what: Is it that certain neighborhoods nurture success (or failure), or rather is it that some neighborhoods just attract those who would succeed (or fail) anyway? As frequently is the case, the empirical challenge is to distinguish correlation from causality. We next turn to the challenging but crucial question of the role played by address (housing, schooling, etc) in determining one’s economic opportunity.

MOVING TO OPPORTUNITY

The United States Department of Housing and Urban Development (HUD) is a Cabinet department in the executive branch of the US federal government. Over the decades, HUD has developed a number of programs related to mortgage and loan insurance, community development, rental assistance, subsidized housing, etc. A particu-
larly important program for economics researchers was the Moving to Opportunity for Fair Housing (MTO), a 10-year program created in the 1990s (1994-1998). The reason why researchers were so interested in the program is that the program consisted of a randomized controlled trial (RCT). The idea is to apply the program to a randomly chosen set of families and then compare results with a control group (namely, families that were not offered the same conditions). From a research point of view, this is a plus because it avoids the usual problem of confounding correlation and causality. In some ways, these RCT are the “gold standard” of empirical research, for they manage to “distill” causal effects in a “clean” way.

The MTO program was offered to low-income families with children living in high-poverty public housing projects in Baltimore, Boston, Chicago, Los Angeles, and New York City. About 4,600 families volunteered for the program, mostly families headed by African-American or Hispanic single mothers. These families were then randomly assigned to three different groups. One group received housing vouchers that could be used only in low-poverty areas for the first year as well as counseling to help them find units there. After a year, they could use their vouchers anywhere. One group received vouchers that could be used anywhere but no counseling. A third (control) group did not receive vouchers but remained eligible for any other government assistance to which they otherwise would have been entitled.

Having started a quarter of a century ago, by 2020 the MTO program provides a rich set of longitudinal data regarding the eventual fate of the children involved in each of the three treatments. (By linking each family to tax records, it is possible to obtain longitudinal information of close to 100 percent of all subjects.) The results are astounding. Compared to the control group — which received no extra benefits — those who received housing vouchers with no strings attached (and no counseling) experienced small or no significant long-term effects. Most of them remained in the same neighborhood where they lived before the program went into effect. (The compliance rate in this group was 66 percent.)

By contrast, the long term effects of offering housing vouchers that required families to move to low-poverty areas were quite significant. (The compliance rate in this group was 48 percent.) Specifically, if the children in the family were younger than 13 at the time
The evidence suggests that a child’s economic prospects are largely determined by the neighborhood where they grow up.

of the move, then we observe that, once they become adults, annual income is $3,477 higher than the control group. This represents a 31 percent increase with respect to the $11,270 average of the control group. Although there is variation in income levels, the statistical probability that the neighborhood move had no effect on the income of the children once they become adults is as low as 1.4 percent. Other indicators also show significant signs of improvement. For example, the probability of college attendance increases by 21.7 percent with respect to the control group (and the probability that an increase is simply the result of noise is as low as 2.8 percent).

Another important result is that the effect depends drastically on the age of the children at the time of the move. For example if the child in question is female and younger than 13 then the probability of becoming a single mother — specifically, to give birth with no father present — decreases from 30 to 23 percent. However, if the mover is between 13 and 18 years old at the time of the move, then the probability of becoming a single mother increases from 4.14 to 51.8 percent.

These results prompt two questions. First, what are the “mechanisms” that make a move to a different neighborhood so effective in terms of long term effects such as college attendance and earnings. Second, if the benefits are so significant, then why do we not observe more moves apart from these controlled programs? Let us start with the second question. One possible explanation for inertia is that low-income families prefer low-income neighborhoods not so much because of barriers to mobility as for reasons such as affordability or proximity to family and jobs. A more recent study/program, sim-
ilarly to the MTO program, provided services to reduce barriers to moving to high-upward-mobility neighborhoods: customized search assistance, landlord engagement, and short-term financial assistance. This intervention increased the fraction of families who moved to high-upward-mobility areas from 14 percent in the control group to 54 percent in the treatment group. The evidence suggests that families induced to move to higher opportunity areas by the treatment do not make sacrifices on other dimensions of neighborhood quality and report much higher levels of neighborhood satisfaction. In other words, most low-income families do not have a strong preference to stay in low-opportunity areas. Rather, locational inertia results from barriers in the housing search process. This shows the importance of providing customized assistance in housing search as a means to help low-income families move to high-mobility areas.

We now turn to the first question: What are the mechanisms whereby moving families achieve such long-term results? First, the results show that neighborhoods — their schools, community, neighbors, local amenities, economic opportunities and social norms — play an important role in shaping a child’s outcomes. Beyond that, it’s difficult to pinpoint the exact mechanism, for the simple reason that many of these neighborhood characteristics are highly correlated with each other.

A recent study relates the gains from a family’s move to the characteristics of the neighborhood the family moves to. The study compensates for the absence of a clean randomized experiment by analyzing a very large number of children, close to 25 million. The problem of not having a randomized experiment is that moves are endogenously determined by families and are potentially dependent on events that are not observed by the analyst. Based on the results from the MTO project, the authors are particularly interested in measuring the effect of the time children are exposed to “better” neighborhoods. For this reason, the important assumption required for correct identification of the relevant effects is that the probability of a move is not determined by the age of children. This is a strong assumption, but one that the authors check by looking at families with multiple children. The results reiterate the prior results from the MTO program: moving to a better neighborhood has a significant long-term effect on children, especially when the children move at an early age. Specifically, each year of childhood exposure to a
In the 1970s, a group of African-American tenants sued the Chicago Housing Authority (CHA) for assigning tenants to public housing on the basis of race. Although the District Court dismissed the plaintiffs’ complaint, the Court of Appeals reversed and ordered the District Court to enter summary judgment for the plaintiffs, holding that there had been a violation both of the Fifth Amendment and of the Civil Rights Act of 1964. This led to another case, namely regarding the remedy. The original plaintiffs wanted to consider a move within the Chicago metropolitan area, whereas the CHA wanted them to remain within the city limits. Eventually, the US Supreme Court, in *Hills v Gautreaux*, ruled in favor of the tenants, many of whom used their Section 8 vouchers to move to the Chicago suburbs.

Subsequent research showed that the families who moved away from high-poverty neighborhoods in Chicago fared considerably better than the ones that remained behind. However, it was not clear whether there was any causal effect. After all, it could simply be that the families that decided to move had better abilities to begin with and decided to move precisely because of those better abilities.

The more recent research (cf p. 432 and following pages) suggests there was a causal effect from the move. In fact, it is estimated that DuPage County — the western suburbs of Chicago — “produces the best outcomes for children from below-median income families among the 100 largest counties. Growing up from birth in DuPage County would increase a child’s income by 16.0% relative to the average county ... Moving from Chicago proper to the western suburbs of Chicago at birth would increase a child’s household income by $7,510 a year on average, a 28.8% increase.”

Box 13.1: Moving out of Chicago Poverty Areas

In the 1970s, a group of African-American tenants sued the Chicago Housing Authority (CHA) for assigning tenants to public housing on the basis of race. Although the District Court dismissed the plaintiffs’ complaint, the Court of Appeals reversed and ordered the District Court to enter summary judgment for the plaintiffs, holding that there had been a violation both of the Fifth Amendment and of the Civil Rights Act of 1964. This led to another case, namely regarding the remedy. The original plaintiffs wanted to consider a move within the Chicago metropolitan area, whereas the CHA wanted them to remain within the city limits. Eventually, the US Supreme Court, in *Hills v Gautreaux*, ruled in favor of the tenants, many of whom used their Section 8 vouchers to move to the Chicago suburbs.

Subsequent research showed that the families who moved away from high-poverty neighborhoods in Chicago fared considerably better than the ones that remained behind. However, it was not clear whether there was any causal effect. After all, it could simply be that the families that decided to move had better abilities to begin with and decided to move precisely because of those better abilities.

The more recent research (cf p. 432 and following pages) suggests there was a causal effect from the move. In fact, it is estimated that DuPage County — the western suburbs of Chicago — “produces the best outcomes for children from below-median income families among the 100 largest counties. Growing up from birth in DuPage County would increase a child’s income by 16.0% relative to the average county ... Moving from Chicago proper to the western suburbs of Chicago at birth would increase a child’s household income by $7,510 a year on average, a 28.8% increase.”
Counties with less concentrated poverty, less income inequality, better schools, a larger share of two-parent families, and lower crime rates tend to produce better outcomes for children in poor families. Boys’ outcomes vary more across areas than girls’ outcomes, and boys have especially negative outcomes in highly segregated areas.

The authors also show that — not entirely surprising — house prices are on average higher in better neighborhoods (where better corresponds to the above characteristics). However, there are many “opportunity bargains,” that is, places that generate good outcomes but are not very expensive.

By way of conclusion, we might say that the study of US neighborhoods results in both good news and bad news. The bad news is that not only economic conditions vary enormously across neighborhoods but also the neighborhoods with worse economic conditions are also neighborhoods of lower mobility. The good news is that moving to a better neighborhood has significant and long-lasting effects on the children of low-income families. The good news is also that the benefits of such move largely compensate the costs. Finally, the bad news is that many beneficial moves simply do not take place, but the good news is that proper counseling has a significant effect on the families’ decision to move.

EDUCATION AND OPPORTUNITY

Economists, policymakers, parents — all seem to agree that education can be an important source of opportunity. However, underneath such general statement lurk a number of difficult questions. What should we do to improve the opportunities offered by education? If the government had $1 to spend on education, where would that dollar go? The problem for an analyst is that there are many levels of education, and at each level there are many variables one has to consider. To make things worse, these variables tend to be highly correlated: across communities, when education works well it seems as everything works well: students attainment, teacher quality, parental education and involvement, household income, and so on. When all measurable variables vary in the same direction, it’s difficult to determine what causes what; we must therefore find strategies to tease
out correlation from causality. We have seen many instances of this throughout the book, but there is a sense in which education poses particularly difficult problems, namely the fact that there are so many correlations.

**COLLEGE EDUCATION**

To what extent is college education a source of opportunity, an engine of mobility? A recent [study](#) looks at the over 30 million college students who enrolled in US colleges from 1999-2013. Four main results come out of this study. First, access to colleges varies greatly by parent income. For example, children whose parents are in the top one percent of the income distribution are 77 times more likely to attend an Ivy League college than those whose parents are in the bottom 20 percent. Second, children from low and high-income families have similar earnings outcomes conditional on the college they attend. Third, rates of upward mobility — the fraction of students who come from families in the bottom income quintile and reach the top quintile — differ substantially across colleges, largely because low-income access varies significantly across colleges with similar earnings outcomes. Fourth, between 2000-2011 the fraction of students from low-income families enrolled at elite private colleges did not change substantially, but the same fraction at colleges with the highest rates of bottom-to-top-quintile mobility declined sharply. In other words, college is less of an engine of mobility that it has been in the past.

As mentioned earlier, there is considerable variation across colleges regarding their role in promoting mobility in the income distribution. Table 13.4 plots two top 10 mobility lists. The top panel refers to the top 10 colleges in terms of bottom quintile to top quintile mobility. The bottom-to-top-quintile mobility rate is the fraction of students whose parents were in the bottom quintile of the parent household income distribution (when they were aged 15-19) and whose own earnings (at ages 32-24) place them in the top quintile of the children’s income distribution. The mobility rate equals the product of the fraction of children at a college with parents in the bottom quintile of the income distribution (column “Access”) and the fraction of children with parents in the bottom quintile of the income distribution who reach the top quintile (top panel) or the top per-
TABLE 13.4

Colleges with the Highest Mobility Rates

### Top 10 Colleges by Bottom-to-Top-Quintile Mobility Rate

<table>
<thead>
<tr>
<th>Rk</th>
<th>Name</th>
<th>Mobility</th>
<th>Access</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cal State, LA</td>
<td>9.9</td>
<td>33.1</td>
<td>29.9</td>
</tr>
<tr>
<td>2</td>
<td>Pace University — New York</td>
<td>8.4</td>
<td>15.2</td>
<td>55.6</td>
</tr>
<tr>
<td>3</td>
<td>SUNY — Stony Brook</td>
<td>8.4</td>
<td>16.4</td>
<td>51.2</td>
</tr>
<tr>
<td>4</td>
<td>Technical Career Institutes</td>
<td>8.0</td>
<td>40.3</td>
<td>19.8</td>
</tr>
<tr>
<td>5</td>
<td>University of Texas — Pan American</td>
<td>7.6</td>
<td>38.7</td>
<td>19.8</td>
</tr>
<tr>
<td>6</td>
<td>CUNY System</td>
<td>7.2</td>
<td>28.7</td>
<td>25.2</td>
</tr>
<tr>
<td>7</td>
<td>Glendale Community College</td>
<td>7.1</td>
<td>32.4</td>
<td>21.9</td>
</tr>
<tr>
<td>8</td>
<td>South Texas College</td>
<td>6.9</td>
<td>52.4</td>
<td>13.2</td>
</tr>
<tr>
<td>9</td>
<td>Cal State Polytechnic — Pomona</td>
<td>6.8</td>
<td>14.9</td>
<td>45.8</td>
</tr>
<tr>
<td>10</td>
<td>University of Texas — El Paso</td>
<td>6.8</td>
<td>28.0</td>
<td>24.4</td>
</tr>
</tbody>
</table>

### Top 10 Colleges by Bottom-Quintile-to-Top-Percentile Mobility Rate

<table>
<thead>
<tr>
<th>Rk</th>
<th>Name</th>
<th>Mobility</th>
<th>Access</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>University of California — Berkeley</td>
<td>0.76</td>
<td>8.8</td>
<td>8.6</td>
</tr>
<tr>
<td>2</td>
<td>Columbia University</td>
<td>0.75</td>
<td>5.0</td>
<td>14.9</td>
</tr>
<tr>
<td>3</td>
<td>MIT</td>
<td>0.68</td>
<td>5.1</td>
<td>13.4</td>
</tr>
<tr>
<td>4</td>
<td>Stanford University</td>
<td>0.66</td>
<td>3.6</td>
<td>18.5</td>
</tr>
<tr>
<td>5</td>
<td>Swarthmore College</td>
<td>0.61</td>
<td>4.7</td>
<td>13.0</td>
</tr>
<tr>
<td>6</td>
<td>Johns Hopkins University</td>
<td>0.54</td>
<td>3.7</td>
<td>14.7</td>
</tr>
<tr>
<td>7</td>
<td>New York University</td>
<td>0.52</td>
<td>6.9</td>
<td>7.5</td>
</tr>
<tr>
<td>8</td>
<td>University of Pennsylvania</td>
<td>0.51</td>
<td>3.5</td>
<td>14.5</td>
</tr>
<tr>
<td>9</td>
<td>Cornell University</td>
<td>0.51</td>
<td>4.9</td>
<td>10.4</td>
</tr>
<tr>
<td>10</td>
<td>University of Chicago</td>
<td>0.50</td>
<td>4.3</td>
<td>11.5</td>
</tr>
</tbody>
</table>

By means of example, consider the numbers for New York University. 6.9 percent of the accepted students come from households in the bottom quintile (i.e., bottom 20 percent) of the income distribution. Of these students, 7.5 percent end up in the top 1 percent of the income distribution. Multiplying 6.9 by 7.5 we conclude that 0.52
percent of the students admitted at NYU (approximately 1 in 200) move from the bottom quantile to the top percentile of the income distribution.

The two panels in Table 13.4 suggest a number of observations. First, when it comes to moving to the very top of the income distribution, elite universities seem to be the most relevant vehicles. For example, if your parents are in the bottom quintile of the income distribution and you are enrolled at Stanford, there is a 18.5 percent change you will get to the top 1 percent of the income distribution. Second, the most successful colleges in moving students from the bottom quintile to the top quintile are a mixture that includes research universities and community colleges, private and public institutions, well-known names and not-so-well-known names. The table also suggests that the elite schools accept a fairly low number of lower-income students. Specifically, the access rates in the bottom panel are considerably lower than those in the top panel.

The values in Table 13.4 suggest that lower-income students may be under-represented at top colleges. For example, the access rates at the schools listed in the bottom panel are uniformly lower than 10
percent. However, we must be careful about jumping to such conclusion. First, the list of schools in the Table 13.4 form a very selective sample. Second, to the extent that higher-ability students have on average higher income we might observe a low percentage of low-income students even if there equal opportunity based on abilities. Figure 13.9 addresses the issue of selection and opportunity head on. We fix a student’s ability as measured by their SAT score and then compute the distribution by income quantile conditional on a given level of ability.

The blue bars in Figure 13.9 show the fraction of students by parental income quintile who attend selective (i.e., non-open-enrollment) colleges, among students who have an SAT score of exactly 1080 (the median score among students who attend selective colleges). The chart shows that children from lower-income families are under-represented relative to children from high-income families despite having the same test scores: If students were selected solely based on SAT scores, then we would expect the distribution by percentiles to be approximately constant, which it is not. The red bars, in turn, show how the fraction of students who attend elite private colleges (the Ivy-League plus Chicago, Duke, MIT, and Stanford) varies with parental income, among students who scored exactly a 1400 on the SAT (the median score among Ivy-Plus students). (Since the frequencies are uniformly lower than those of 1080 SAT students, the values are scaled up and should be read on the right-hand scale.) As can be seen, middle-class students are especially underrepresented at Ivy-plus colleges. Somewhat surprisingly, students from the lowest income (bottom quintile) families are only slightly under-represented at elite private colleges.

The above analysis is essentially descriptive. In particular, it does not identify a college’s causal effects on students’ outcomes. The difficulty, as mentioned before, is to distinguish correlation from causality. When we compare the income stream of a student who went to college with that of a student who did not go to college we get an uncorrected college premium (see, for example, Figure 11.8). However, it is likely that the innate unobservable skills of the student who went to college are higher than those of the student who did not. Therefore, a part of the uncorrected college premium is not a premium in the causal sense of the word, that is, is not a measure of the income return from going to college. Similar to the problem of estimating
a demand curve (cf Section 6.2), causal estimates can be obtained if we observe a variable that shifts the supply of skills but not the demand for skills (for example, a change in the ability to apply for college loans). Research based on such empirical strategies estimates a return of college education of about 9%. Two notes are in order, however. First, as suggested by Figure 11.8, the return on college education has somewhat decelerated. Second, and related, there is a significant difference between the average return and the marginal return to the lower-skilled college enrollees (the latter can be as low as 1 percent). This distinction, which is yet another instance of decreasing marginal return, is important as we try to reconcile the effects of the wage distribution on increasing inequality with the deceleration of the college premium.

PRIMARY AND SECONDARY EDUCATION

When it comes to primary and secondary education, one of the most important questions asked by economists is whether teachers have an impact on students, both in terms of their success in school and in terms of long-term outcomes. The empirical challenge is that a teacher with low-achievement students may nevertheless be a very good teacher: had the students had a different teacher, their academic scores would have been even worse. One possible solution to this identification problem is to measure a student change in grades when studying under a certain teacher, what is usually referred to as the teacher’s value added (VA). However, even then there are two problems. First, a high VA may simply result from sorting: For example, a certain teacher becomes popular for some reason unrelated to his or her abilities, many students try to study under this teacher, and the teacher in turn is able to select students with better abilities, which in turn is reflected in higher gains in terms of grades. In other words, differences in VA may result from sorting rather than a causal effect of teacher quality. The second problem with VA measures is that higher-VA teachers may simply teach to the test, in which case higher grades may not reflect actual long-term gains.

A series of studies addresses both of these criticisms of the use of VA as a means of valuing teachers. Based on a very large number of student and teacher observations (more than a million) the data indicate that VA does measure teacher qualities rather than sort-
Moreover, it can be shown that high-VA teachers are not simply teaching to the test, rather they have a long-term effect on their students’ outcomes: Students assigned to high-VA teachers are more likely to attend college, earn higher salaries, and are less likely to have children as teenagers. Replacing a teacher whose VA is in the bottom 5 percent with an average teacher would increase the present value of students’ lifetime income by approximately $250,000 per classroom.

Given all of this variation in teacher AV, it follows that studying under the right teacher — or attending the right school — may have a significant impact on a student’s long-term opportunities. This leads to a natural follow-up question: Just like moving to a different neighborhood, so the ability to choose a different school may be a significant engine of mobility. In the US, there are two types of programs that have attempted to increase the ability of parents — especially low-income parents — to choose their children’s school: voucher programs and charter schools.

A school voucher is government-supplied coupon that is used to offset tuition at an eligible private school. The main argument in favor of choice in general and voucher programs in particular is that they provide disadvantaged students the opportunity to attend a better school, similar to what programs such as MTO attempt to achieve with family relocation. Moreover, by increasing competition between schools, better outcomes are achieved (more motivated teachers, etc). Last but not least, to the extent that both parents and schools differ in their education philosophies, allowing families to choose leads to a better match between supply and demand. The main argument against voucher programs is that they lead to sorting — sometimes referred to as stratification. The idea is that the students with better abilities are more likely to leave public schools, leaving the latter with a low-ability pool. So, while student migration may improve the outcomes of the outgoing students, the ones who stay behind are likely to suffer from the program (both because of so-called peer effects and because of the stigma associated with studying at a lower-achievement school). Also, to the extent that private schools may be able to charge tuition add-ons, vouchers may also contribute to income stratification.

Even if we restrict to the US, the programs that distribute school vouchers vary in several dimensions: who is eligible to receive them,
the source of funding for the program, the criteria for private-school participation, and so forth. For this reason, it is difficult to make a general assessment of the effects of voucher programs. That said, with regard to the effect on the students enrolled in the program, the empirical evidence is mixed. For example, research of the Milwaukee voucher program — one of the earliest in the US — reports that voucher recipients had faster math score gains than, but similar reading score gains to, the comparison groups. However, research on a Louisiana voucher program shows that participants saw their math scores decline, a pattern the authors attribute to poor school selection on the part of the parents. There are also many studies reporting very low effects on participating students.

Several studies suggest that there is some sorting effect, not so much on income — after all, the programs are typically targeted at lower-income families — but in terms of ability. In other words, better ability students are disproportionately more likely to take advantage of voucher programs. There is also evidence that the threat of losing students to voucher programs leads public schools to improve their performance.

I should also add that voucher programs have become a bit of a political hot button, especially in the US. In addition to the points raised above, many feel uncomfortable with the idea of government funds being used for non-public schooling. In this sense, charter schools — an alternative school choice policy — provides a more consensus approach. The idea behind the charter school movement is to maintain universal access as in the traditional public system but giving each school the freedom to organize, a freedom that is provided in exchange for greater accountability.

The first US charter school opened in Minnesota in 1992. Since then, the scale of the charter movement has grown to nearly 7,000 schools and three million students in 43 states plus the District of Columbia as of the 2016-17 school year. Charter enrollments constitute about 6 percent of annual public school enrollments in the United States with enrollment shares as high as 43 percent in the District of Columbia. Table 13.5 provides some descriptive statistics of US charter schools, in particular as they compare to traditional public schools. As can be seen, charter schools focus primarily on urban neighborhoods with mostly black and hispanic populations.

How effective are charter schools? One advantage of the data
TABLE 13.5
Characteristics of traditional public schools and of charter schools, 2015-2016 (source)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>public</th>
<th>charter</th>
</tr>
</thead>
<tbody>
<tr>
<td>White students</td>
<td>49.9</td>
<td>33.1</td>
</tr>
<tr>
<td>Black students</td>
<td>14.7</td>
<td>26.8</td>
</tr>
<tr>
<td>Hispanic students</td>
<td>25.6</td>
<td>31.7</td>
</tr>
<tr>
<td>75%+ students eligible for lunch program</td>
<td>23.9</td>
<td>32.6</td>
</tr>
<tr>
<td>Urban location</td>
<td>24.9</td>
<td>56.5</td>
</tr>
</tbody>
</table>

available from various schools is that when oversubscribed — which happens often — their admissions are based on a lottery system. This effectively provides researchers with a randomized control trial with which to estimate causal effects. The results from these studies tend to show positive and significant effects. A study based on Harlem data finds that attending an HCZ (charter) middle school effectively closes the black-white achievement gap in mathematics. A study based on Charlotte-Mecklenburg data finds a significant overall increase in college attainment among lottery winners who attend their first-choice school, and that gains in attainment are concentrated among girls. A study based on Boston data shows significant score gains for charter students in middle and high school. Interestingly, it also compares charter schools with pilot schools. Like charter schools, pilot schools are answerable to independent governing boards and determine their own budgets, staffing, curricula, and scheduling. Unlike charter schools, however, pilot schools remain part of the Boston school district and their teachers are part of the teachers’ union. Lottery estimates for pilot school students are mostly small and insignificant, with some significant negative effects.

The internal validity of these studies is unassailable: In a sense, randomized control trials are the gold standard of social science. However, the sample of schools used for these RCT is necessarily biased: oversubscribed schools are likely to be better than average. There is also ample evidence of poorly performing or even failing charter schools. In this regard, a particularly important study, based on data from Texas, finds that average school quality in the charter sector is not significantly different from that in regular public schools.
after an initial start-up period but that there is considerable heterogeneity. Moreover, parental decision to exit a charter school is significantly related to school quality and more so than in the regular public school sector. This is consistent with the notion that the introduction of charter schools substantially reduces the transactions costs of switching schools.

The heterogeneity in the effect of charter schools also seems related to their location. A consistent pattern seems to be that in urban areas — where students are overwhelmingly low-achieving, poor and nonwhite — charter schools tend to do better than other public schools. By contrast, outside of urban areas, where students tend to be white and middle class, charters do no better and sometimes do worse than public schools. In this regard, an important feature of the charter school system is that their independence comes at the cost of regular review: failing charter schools are closed much more easily than traditional public schools are. This is a very important feature. Earlier I mentioned that, by replacing a teacher whose VA is in the bottom 5 percent with an average teacher would increase the present value of students’ lifetime income by approximately $250,000 per classroom. These large gains result from the large heterogeneity in VA across teachers. Something similar happens with charter schools: The data show considerable variation in their performance. At some level, this is a bad thing (nobody likes poor performing schools). But together with the flexibility the system has — namely in terms of closing failing schools — the end result may be quite good.

**EARLY YEARS**

In the late 20th century, two psychology experts in early development decided to follow a number of families from different socioeconomic backgrounds. They were interested in understanding why, by age 4, there were already such marked differences in the children’s ability to absorb new vocabulary learned in pre-school and kindergarten. They assembled a sample of 42 Kansas families from different socioeconomic status (SES). There were African-American families in each SES category, in numbers roughly reflecting local job allocations. Each month, for 2.5 years, the researchers spent one hour with each family simply observing their habits, in particular the interaction between parents and children.
The results were remarkable. First, perhaps not surprisingly, children at an early age are very much a reflection of their parents.

Before children can take charge of their own experience and begin to spend time with peers in social groups outside the home, almost everything they learn comes from their families.

In quantitative terms, 86 to 98 percent of the words recorded in each child’s vocabulary consisted of words also recorded in their parents’ vocabularies. Figure 13.10 provides a graphic description of this correlation. On the horizontal axis we measure word counts for parents, on the vertical axis word counts for their children. Two different concepts are plotted: the average number of words per hour and the average number of different words per hour. (Naturally, the latter is smaller than the first one.) A first striking feature of Figure 13.10 is the high correlation between parents and children. Drawing a line through the red dots (number of words) or the green dots (number of different words) shows a close-to-proportional relation between parents and children.

A second remarkable feature of Figure 13.10 is the significant difference in number of words across socio-economic groups. On average, a professional parent speaks 2.77 more words (2.29 more different words) than a parent on welfare. In other words, it’s not just the
Education — especially during a child’s early age — can be an important path of social and economic mobility.

fact that higher SES parents have a better vocabulary, it’s also the fact they talk more to their children.

As frequently is the case with studies of this sort, the question arises as to whether there is a causal relation or simply a correlation implied by genetic factors. In other words, since the parents’ and the children’s genetic features are correlated, then what we observe in Figure 13.10 may result solely from spurious correlation. This view, while popular in certain circles, has been largely discredited by recent studies on cognitive abilities. For example, a recent study finds very minor differences in test outcomes between black and white infants. In other words, the data leads one to believe that correlations like that in Figure 13.10 correspond to causal relations dictated by a child’s growing environment. Moreover, the evidence from the Kansas study suggests that the differences across SES are more than purely cognitive.

We can extrapolate similarly the relative differences the data showed in children’s hourly experience with parent affirmatives (encouraging words) and prohibitions. The average child in a professional family was accumulating 32 affirmatives and five prohibitions per hour, a ratio of 6 encouragements to 1 discouragement. The average child in a working-class family was accumulating 12 affirmatives and seven prohibitions per hour, a ratio of 2 encouragements to 1 discouragement. The average child in a welfare family, though, was accumulating five affirmatives and 11 prohibitions per hour, a ratio of 1 encouragement to 2 discouragements.
In sum, the evidence suggest that there is much to be gained from policies that help children and their parents. Since the above-mentioned Kansas research took place, several other studies have identified parenting practices that play an important role, including doing enriching activities with children, getting involved in their schoolwork, providing educational materials, and exhibiting warmth and patience. In fact, a recent study claims that

Parental behavior interpreted in this way probably accounts for around half of the variance in adult economic outcomes, and therefore contributes significantly to a country’s intergenerational mobility.

In this regard, one important program (and research study) is the Perry Preschool program, a program that served disadvantaged African American kids in Ypsilanti, Mich. Essentially, the program consisted of regular home visits during the kids’ preschool years, involving both the kids and their parents. Similarly to the MTO program mentioned earlier, families were randomly assigned to the program (among the families that initially volunteered). Because the Perry Preschool Program was conducted in the 1960s, researchers have been able to follow the children who went through the program through adulthood. The results suggest the effects of the program were enormous: Adults from the treatment group were “much more likely to graduate high school, much more likely to make earnings, much more likely to go on to college, much less likely to commit crime.” More: the benefits of the program extended beyond the first generation: children of participants in the program appear to have benefitted as well.

The children of the participants are healthier. The children of the participants are also earning more. They have better social and emotional skills, are more likely to graduate high school and go on to college, less likely to engage in the criminal justice system, so they’re less likely to be incarcerated or even have ever been arrested.

The studies surveyed above represent a small subset of the research on the effects — including the economic effects — of education, and of education as an engine of economic and social mobility. Three
common themes what seem to summarize the evidence: First, it pays to invest in education. It pays both for individuals and for public policy. Moreover, the return on investment seems to be greater the earlier these investments are made. Second, school quality varies from school to school, and within a given school teacher quality varies too. For this reason, mechanisms that allow for selection (e.g., replacing poorly performing teachers with average performance teachers) can have significant positive effects. Third, families matter a great deal, both in terms of household structure and in terms of parental involvement in the kids’ education.

*Education — especially during a child’s early age — can be an important path of social and economic mobility.*

As frequently is the case, one can see the glass half full or half empty. The current scope for economic and social mobility in the US is rather low, especially when compared with other developing countries. However, there is ample room for improvement and we seem to know a number of ways in which improvement can be attained.
KEY CONCEPTS

- brain drain
- mobility matrix
- rank elasticity
- heat map
- randomized controlled trial
- college premium
- teacher's value added
- school voucher
- stratification
- charter school
demand, 177
crowd out, 261
cumulative distribution function, 320
deadweight loss, 254, 400
decreasing marginal benefit, 74
decreasing marginal rate of substitution, 98
decreasing marginal returns, 74, 147, 442
decreasing marginal utility, 74
decreasing returns to scale, 163
demand curve, 207, 210
demand function, 210
difference-in-differences, 268
discount rate, 321
discrimination, 28
diversity traps, 377
division of labor, 16
dominant firm, 276
dominant strategy, 66, 278
dominated strategy, 66
economic model, 48
efficiency, 253
    allocative, 253
deadweight loss, 254
    productive, 255
    X-inefficiency, 256
elasticity rule, 179
equality of opportunities, 392
equality of outcomes, 392
excess demand, 227
excess supply, 227
excludable, 313
expected value, 130
experience goods, 344
externality, 304
feasible set, 72, 88
fertility rate, 134
firms, 13
First Welfare Theorem, 253
fixed cost, 188
forward reasoning, 285
free-rider problem, 312
GAFA, 297
game, 67, 278
    intertemporal favor-exchange, 386
    ultimatum, 383
    zero sum, 75
game theory, 64
Gini coefficient, 356
globalization, 18
Gross Domestic Product (GDP), 6
growth rates, 6
health mandate, 335
heat map, 429
high-powered incentive scheme, 342
history’s hockey stick, 12
homo economicus, 58
homogeneous product, 224
household responsibility system, 24
human capital, 364
identification, 53, 216
identifying assumption, 53
implicit collusion, 295
implicit discrimination, 375
import quota, 264
import tariff, 264
incentive scheme
    high powered, 342
low powered, 342
incentives, 339
incidence (tax), 399
income effect, 112
income elasticity of demand, 176
increasing returns to scale, 163
indifference curves, 93
industrial revolution, 15
inferior good, 109
innovation, 19
institutional discrimination, 61
instrumental reciprocity, 387
insurance premium, 131
intermittency problem, 325
intertemporal favor-exchange, 386
intrinsic reciprocity, 387
inverse demand curve, 209
isocost, 154
isoprofit, 167
isoquant, 151
killer acquisitions, 293
labor, 15
labor productivity, 164
labor supply, 119
laboratory experiment, 383
Laffer curve, 407
law of demand, 113, 210
law of large numbers, 212
law of supply and demand, 228
learning by doing, 17
leniency program, 294
Leontief production function, 153
libertarians, 393
life-cycle optimization, 124
lobbying, 265
long run, 198
long-run supply curve, 198
Lorenz curve, 355
low-powered incentive scheme, 342
macroeconomics, 48
management complexity, 163
margin, 178
marginal, 68, 123
marginal cost, 188
marginal external benefit, 307
marginal external cost, 305
marginal product, 147
marginal rate of substitution (MRS), 97
marginal rate of technical substitution (MRTS), 152
marginal rate of transformation (MRT), 89
marginal revenue, 171, 195
marginal social benefit, 307
marginal social cost, 306
market, 13, 47
market economies, 14
market equilibrium, 226
market exchange, 17
market power, 365
market value, 77, 250
market-based, socially concerned, 393
market-clearing price, 228
merger policy, 295
merger synergies, 295
microeconomics, 48
mixed economies, 22
mobility matrix, 426
monopoly, 276
monopsony, 365
monotonicity, 93
moral hazard, 338
Nash equilibrium, 279
natural monopoly, 163
negative externalities, 305
network effects, 277
non-excludable, 313
non-rival goods, 313
normal good, 108
normative analysis, 56
nudge, 347, 395
Obamacare, 335
oligopoly, 276
opportunity cost, 70
Pareto frontier, 389
Pareto optima, 389
payoffs, 278
pension systems, 396
per-capita GDP, 6
perfect complements, 99, 152
perfect information, 225
perfect substitutes, 99, 153
players, 278
pluralistic ignorance, 379
pooling equilibrium, 335
positional good, 60
positive analysis, 56
positive externalities, 305
post hoc fallacy, 52
potential Pareto move, 391
price elasticity of demand, 174
price gouging, 270
price taker, 116, 225
principal, 341
principal-agent problem, 341
prisoner’s dilemma, 67
private property, 13
process innovation, 19
producer surplus, 245
product innovation, 19
production function, 19
productivity, 15, 255
property rights, 225
public goods, 314
public policy, 48
purchasing power parity, 7
randomized control trial, 445
randomized controlled trial, 433
rank elasticity, 427
rational agents, 92
real income, 113
regulated markets, socially concerned, 393
repeated game, 280
representative agent, 120
returns to scale, 162
revealed preference, 78
risk aversion, 130
risk neutral, 130
risk premium, 132
rival goods, 313
rules, 278
sample selection effect, 55
savings function, 128
scale economies, 16
school voucher, 443
schooling, 395
search goods, 344
Second Welfare Theorem, 390
selection, 28
short run, 198
short-run supply curve, 198
sin tax, 404
single payer system, 336
skilled tradable services, 369
skilled-biased technical change, 361
social cost of carbon, 320
social dilemma, 64
social norms, 380
social status, 60
socialists, 393
sociology, 59
special interests, 265
specialization, 16
state of the world, 129
statistical discrimination, 61
strategies, 278
stratification, 443
strict Pareto move, 391
substitution effect, 111
sunk cost, 70
superstar effect, 366
superstar firms, 297, 369
supply curve, 200
  long run, 198
  short run, 198
supply function, 195, 200
surprise billing, 343
sustainable development, 33
Sustainable Development Goals, 33
sustainable economy, 33
sustainable resource use, 38
take-off, 13
taste discrimination, 61
tax
  carbon, 113
  incidence, 399
  sin, 404
teacher’s value added, 442
tech giants, 297
technology progress, 15
total cost, 188
total factor productivity, 165
tragedy of the commons, 304
transitivity, 93
tree game, 284
truth in advertising, 344
ultimatum game, 383
Universal Basic Income, 397
universal healthcare, 335
unknown unknowns, 322
utility, 91
value, 17
value in use, 77, 250
value of life, 42
variable cost, 188
varieties of capitalism, 22
welfare
  First Theorem, 253
  Second Theorem, 390
well-defined property rights, 225
willingness to pay, 208, 247
X-inefficiency, 256
zero-sum game, 75