Production

Jón Steinsson*
University of California, Berkeley

February 2, 2026

Science can be described as consisting of two activities: theory and empirics. Scientists gather data about the world and they seek to explain that data. These two activities are symbiotic. Measurement needs to be grounded in theory because there are too many things that can be measured about the world for it to be productive to measure things at random. Likewise, theory needs to be grounded in empirical evidence because there are too many theories that can be developed for it to be productive to theorize at random. For these reasons, all good science involves an interplay between theory and empirics. Many scientists specialize to some degree in either theory or empirics. But the symbiotic nature of theory and empirics implies that good scientists need to know some of both.

One goal of this chapter is to provide a first introduction to macroeconomic theory. An important tool economists use heavily when they theorize is formal modeling. This consists of representing some aspect of the economy (sometimes the entire economy) by a set of mathematical equations. These equations are called a *model*. The theorist starts by making some assumptions about the environment and about human behavior. They then derive a set of equations implied by these assumption. Finally, they solve the equations to see what the model implies about how the economy works.

Not all economic theory is formal. Some theory – even some of the most influential theory in the field – is informal (completely verbal without any mathematics).

*I would like to thank Alex Blumenfeld and Alessandro Dell'Acqua for excellent research assistance and Anand Bharadwaj and Emi Nakamura for valuable comments and discussions. I would like to thank Catherine Carlson, Michael Lee, James McAuliffe, Isaac Miller, Alexander Monissen, Alberto Undurraga-Flotts, Xiao Zhang, and Joshua Zheng for finding errors and typos. First posted in February 2026.

We will encounter such theory in many places in this book. But a substantial portion of economic theory is formal (mathematical) and it is thus an important part of the skill set of a well-trained economist to understand how to read and write formal economic theory. In this chapter, we take our first steps in this direction.

The model we develop in this chapter is very simple. It abstracts from many aspects of reality. But it nonetheless introduces several core concepts of economic modeling that we will use over and over again. These include the production function, the idea that firms maximize profits, and the idea of an equilibrium. In subsequent chapters, we will add additional features to make the models we analyze more realistic in various ways.

Another goal of this chapter is to analyze several important substantive questions. The most important question we consider is the effect of technological change on workers. This is a complex issue. On the one hand, technological change is the source of much of the increase in living standards humans have experienced over the past few centuries. On the other hand, technological change is disruptive and can destroy jobs. We will use the model we develop to shed light on these issues. But we will also use empirical evidence to think about whether the model is consistent with salient features of the real-world economy.

1 Some Preliminaries about Economic Models

The simplest and most canonical model in economics is a model of a single market consisting of two equations: one representing demand and another representing supply. Suppose for concreteness that we are analyzing the market for soy beans. Here are two equations that represent supply and demand in this market:

$$\begin{aligned} q_t &= ap_t + w_t \\ q_t &= -bp_t + p_t^{xyz}. \end{aligned}$$

The quantity of soy beans sold is denoted by q_t, while p_t denotes the price of soy beans. a and b are positive coefficients (i.e., numbers), while w_t and p_t^{xyz} are other variables that influence supply and demand.

These two equations look very similar in that they both give relationships between price and quantity. How can we tell which one is the supply curve and which one is the demand curve? Recall that supply curves slope upward, while demand curves slope downward. Since a and b are positive, the signs in front of a and b tell us

that the first equation is the supply curve, while the second equation is the demand curve. The variable w_t might then represent the weather, which affects the supply of soy beans, while p_t^{xyz} might represent the price of a substitute for soy beans, such as rapeseed.

When thinking about any economic model, it is crucial to distinguish between two groups of objects:

1. Endogenous variables

2. Exogenous variables and parameters

The endogenous variables are the objects of primary interest. These are the variables that one is seeking to solve for when one solves an economics model, i.e., the "unknown" in the problem. The word "endogenous" means "from within." The value of the endogenous variables is determined by solving the model. So, they can be said to emerge "from within" the model.

The exogenous variables are variables one is taking as given in the analysis. They are considered known when one is solving the model. The word "exogenous" means "from outside." The parameters are coefficients that are also considered known when one is solving the model. It is conventional to distinguish between exogenous variables and parameters even though both are known when solving the model (and therefore not fundamentally different). They are essentially two categories of known objects that, by convention, are referred to by different labels.

Whenever one is thinking about an economic model, one of the first things to consider is: what are the endogenous variables of the model? In the supply and demand model above, the endogenous variables are p_t and q_t. In such a model, one is solving for the price and quantity in the market taking other objects as given (i.e., the parameters a and b and the exogenous variables w_t and p_t^{xyz}). If you don't know what the endogenous variables of a model are, you have no hope of solving the model since you don't even know what you are solving for!

The supply-demand model discussed above is an example of a *partial equilibrium* model. This means that it is a model of a part of the economy that takes as given outcomes in other markets and assumes that outcomes in the markets being studied do not affect other markets in ways that feed back on the markets being studied. For example, by treating p_t^{xyz} as an exogenous variable, we are implicitly assuming that outcomes in the soy bean market do not affect p_t^{xyz}. Often such assumptions are unlikely to be literally true, but are reasonable approximations which simplify the

analysis. Partial equilibrium analysis is useful for many purposes and we will see many examples of it in this book.

Much of macroeconomic theory, however, involves studying *general equilibrium* models. A model is a general equilibrium model if prices and quantities in all markets are considered endogenous variables. In this chapter, we will encounter our first general equilibrium model. This model will consist of three markets: a labor market, a capital market, and a goods market. To keep things simple, we assume there is a single type of labor, a single type of capital, and a single consumption good in the economy. We also – for simplicity – ignore the passage of time. This means that we ignore how decisions at one point in time affect the economy later on. We simply assume there is one "time period." With these simplifying assumptions, our three-market model is a general equilibrium model of the economy.

In later chapters, we will relax these assumptions and consider richer models. But it is important to start with a relatively simple model since there are a lot of subtle issues that arise even in very simple models and it is best to master these before moving on to more complex settings. Also, economic models are rarely meant to capture reality in all its complexity. Most of the time, economic models are meant to provide insight into how the world works. This is best done with simple models that focus on a small subset of all the things going on in the economy.

It is useful to provide a rough roadmap of what a general equilibrium model consists of. Typically one starts with a set of agents (households, firms, governments, etc.) and makes assumptions about how these agents behave. In this chapter, we will assume that firms maximize profits. In the next chapter, we will assume that household maximize utility. These assumptions combined with additional assumptions about the environment (technology, available resources, markets, etc.) yield equations describing the behavior of the agents in the model. Typically, these equations can be interpreted as demand and supply curves in each market in the economy.

One therefore ends up with a demand curve and a supply curve for each market in the economy. There are two general caveats to this rule. The first caveat is that only relative prices matter. This implies that one can choose the overall level of prices in the economy to be whatever one wants. In particular, one can choose the overall level of prices such that one of the prices is equal to one. This is a common choice. In this case, the good whose price is one is called the *numeraire* good in the economy. Choosing the overall level of prices is simply a matter of units. Does one want to denote prices in dollar, or in cents, or in euros, or in pounds sterling. This

doesn't matter. So, one might as well choose convenient units (units which result in one of the prices being one).

The second general caveat is called *Walras' Law*. It states that if all but one market in the economy are in equilibrium then the last market must also be in equilibrium. Walras' Law follows from the basic notion that the agents in the economy will use all their resources in some way. This means that if one knows what they do in all but one market, one knows that they do in the last market (the remaining resources are spent in that market). Walras' Law means that one need only write down equations for all but one market in the economy. What happens in the last market will then be implied by the resource constraint in the economy (the fact that all resources will be used in some way).

Let's go back to our example of an economy with three markets: a labor market, capital market, and goods market. How many endogenous variables does this model have? The naive answer would be six: a price and a quantity in each of the three markets. But this answer doesn't take account of the fact that one of the prices can be set to one. So, really, there are only five endogenous variables. For example, perhaps the price of goods is set to one (which means that all other prices are denoted in units of goods). Then the endogenous variables are the quantity of goods Y, the quantity of labor L, the price of labor w, the quantity of capital K, and the price of capital r.

To solve for five unknown variables, one typically needs five questions. One might think one has six equations: a demand and supply curve in each of three markets. But Walras' Law implies that the demand and supply curves in the last market are implied by the demand and supply curves in the first two markets. So, demand and supply curves only give four equations. But there is typically a natural fifth equation. For example, the fifth equation might equate total income and total spending ($Y = wL + rK$) or it might be an aggregate production function ($Y = F(L, K)$). Adding a fifth equation of this type then allows one to solve the model. We will discuss this further below.

2 The Production Function

The process by which goods and services are produced is in many cases very complex. Perhaps somewhat surprisingly, much of economics abstracts from most of this complexity and models production in a very broad brush way using simple *pro-*

duction functions. This production function is meant to capture in a simple manner the technology society has at its disposal to produce goods. This approach has the advantage that it captures certain basic economic forces without adding too much complexity that might not be pertinent to the question at hand.

While large parts of economics employ very simple production functions, the idea of a production function is quite general and can be used to model production in more detail. In fact, one can imagine a complex multi-stage production function that captures the complexity of production in arbitrary detail. Sometimes economists incorporate some of this complexity into their modeling. But much of the time this is considered overkill. (It is thought not to yield additional insight.)

In this section, we consider a very simple production function that is a popular choice in economic modeling. Later in the chapter, we discuss some of the limitations of this production function. This will help us understand in what types of circumstances more complex production functions are needed.

Consider the production of ice cream and suppose for simplicity that making ice cream involves two inputs to production: workers and ice cream machines. Clearly, this is a vast simplification. The production of ice cream also involves ingredients (milk, sugar, eggs, vanilla, etc.), electricity, a building, etc. But let's keep things simple and assume there are only two inputs. We then represent the process of producing ice cream by a function

$$Y = F(K, L), \tag{1}$$

where F is the function, Y denotes the amount of ice cream produced (perhaps measured in pounds), K denotes the amount of capital employed (number of ice cream machines), and L denotes the amount of labor employed (number of workers or hours of work). We say that capital and labor are *inputs to production* or *factors of production*. The amount of ice cream produced is the output produced. We say that the function F maps inputs to production into output produced.

What properties should a reasonable production function have? Consider first the first derivatives. It seems reasonable that these should be positive (at least weakly positive): employing more workers for a given number of ice cream machines yields more ice cream and installing more ice cream machines for a given number of workers also likely yields more ice cream. Mathematically, these assumptions can be written as

$$\frac{\partial F}{\partial L} \geq 0 \quad \text{and} \quad \frac{\partial F}{\partial K} \geq 0. \tag{2}$$

Next consider the second derivatives. Suppose an ice cream shop that starts off with one worker and one machine adds a second worker. This will likely yield quite a bit of extra ice cream production. The first worker spends some time interacting with customers, cleaning, and performing other tasks. At those times, the machine is underutilized and a second worker could add to production. Suppose then that the ice cream shop adds a third worker, and a fourth, and so on, while holding the number of machines constant at one. It seems likely that the third worker will yield less extra ice cream than the second, the fourth worker less extra ice cream than the third, and so on. As the number of workers rises, the fact that the ice cream shop has only one ice cream machine becomes more and more of a bottleneck for production. This implies that the *marginal product of labor* $\partial F/\partial L$ – i.e., the extra amount of output produced per unit of extra labor employed at the margin – falls in the amount of labor employed.

A similar thought experiment suggests that the *marginal product of capital* $\partial F/\partial K$ – i.e., the extra amount of output produced per unit of extra capital employed at the margin – also falls in the amount of capital employed. Starting from one worker and one ice cream machine, adding another ice cream machine will likely add quite a bit to production. The worker may be able to operate both machines at the same time. Perhaps attempting to operate two machines at once will involve some downtime for each machine. But it will still likely be better than only having one machine (which will sometimes need to be cleaned, refilled, or maintained in various ways). If the ice cream shop adds a third machine, this will likely add less production than the second machine, and a fourth machine would likely add even less than the third, and so on. In this case, it is the worker that is becoming a bottleneck in production. The worker can only work so hard, which means that at some point extra machines will mostly stand idle.

Mathematically, the arguments in the last two paragraphs can be written as

$$\frac{\partial^2 F}{\partial L^2} \leq 0 \quad \text{and} \quad \frac{\partial^2 F}{\partial K^2} \leq 0. \tag{3}$$

They state that there is *diminishing return* to each factor of production holding the other factor fixed.

What about the cross-partial $\frac{\partial^2 F}{\partial L \partial K}$? Is it positive? Or is it negative? This is less clear. What does this cross-partial represent? One way to describe it is: how does the marginal product of labor change when the amount of capital employed increases ($\frac{\partial}{\partial K}\frac{\partial F}{\partial L}$)? When the ice cream shop adds another machine, does this increase the

marginal product of labor (i.e., make adding another worker more valuable), or does it decrease the marginal product of labor (make adding a worker less valuable).

Intuitively, this will depend on the nature of the machine being added. In the case discussed above where we were thinking of adding more machines that were identical to the machines already employed, it seemed intuitive that the marginal product of labor would increase. If so, the workers and the machines are complements. But some machines will replace workers, i.e., perform tasks that workers performed before. Think of ATMs or robots or software that automates various tasks. These types of machines may decrease the marginal product of labor. If so, the workers and the machines are substitutes. Whether machines and workers are complements or substitutes is importance since it determines whether more machines raise worker wages or lower worker wages.

2.1 The Cobb-Douglas Production Function

The most commonly used production function in economics is the Cobb-Douglas productions function:

$$Y = AK^a L^{1-a}, \tag{4}$$

where $A > 0$ is typically referred to as total factor productivity (TFP) or sometimes simply as productivity, and $0 \leq a \leq 1$. This production function satisfies the four conditions in (2) and (3):

$$\frac{\partial Y}{\partial L} = (1-a)AK^a L^{-a} \geq 0 \quad \text{and} \quad \frac{\partial Y}{\partial K} = aAK^{a-1}L^{1-a} \geq 0, \tag{5}$$

$$\frac{\partial^2 Y}{\partial L^2} = -a(1-a)AK^a L^{-a-1} \leq 0 \quad \text{and} \quad \frac{\partial^2 Y}{\partial K^2} = -a(1-a)AK^{a-2}L^{1-a} \leq 0. \tag{6}$$

It therefore features positive and diminishing returns to both labor and capital.

What does the Cobb-Douglas production function imply about $\frac{\partial^2 F}{\partial L \partial K}$? Taking the partial derivative of $\partial F/\partial L$ with respect to K yields

$$\frac{\partial^2 Y}{\partial K \partial L} = a(1-a)AK^{a-1}L^{-a} \geq 0.$$

In other words, the Cobb-Douglas production function implies that capital complements labor (and vice versa): an increase in the amount of capital increases the marginal product of labor, i.e., makes workers more productive. We will see below that if labor markets are competitive, this implies that an increase in capital increases wages of workers.

2.2 Returns to Scale

How much would output increase were we to double both capital and labor? With the Cobb-Douglas production function discussed above, we have the following important result

$$\begin{aligned} F(2K, 2L) &= A(2K)^a(2L)^{1-a} \\ &= A2^aK^a2^{1-a}L^{1-a} \\ &= 2AK^aL^{1-a} \\ &= 2F(K, L). \end{aligned}$$

In other words, doubling both capital and labor doubles output. A production function that has this property – that doubling all inputs to production doubles output – is said to exhibit *constant returns to scale*.

What property of the production function $Y = AK^aL^{1-a}$ implies that it is constant returns to scale? The derivation above shows clearly that the key property is that the exponents on capital and labor add to one. Suppose instead that the production function were $Y = AK^aL^b$. In that case, a derivation analogous to the one above would yield

$$F(2K, 2L) = 2^{a+b}F(K, L). \tag{7}$$

Only with $a + b = 1$ do we get that $2^{a+b} = 2$ and $F(2K, 2L) = 2F(K, L)$. This is the case of constant returns to scale. If $a + b < 1$, we have $F(2K, 2L) = 2^{a+b}F(K, L) < 2F(K, L)$. In this case, we say that the production function exhibits *diminishing return to scale*. If $a + b > 1$, we have $F(2K, 2L) = 2^{a+b}F(K, L) > 2F(K, L)$. In this case, we say that the production function exhibits *increasing returns to scale*.

In macroeconomics, we often assume that the aggregate production function for the economy as a whole is constant returns to scale. What justifies this assumption? For a single factory, constant returns to scale is actually not a very appealing assumption. Factories typically become more efficient as they grow in size up to some point. Small factories do not produce enough to justify the use of specialized machinery. As the factories grow, it pays to invest in more such machinery, which increases their efficiency.

At some point, however, the factory has reached a scale where it already employs top-of-the-line machinery across the board. At that point, increasing its size no longer increases efficiency. The factory's level of efficiency may then level off with further growth, or it may even fall, for example, if the factory becomes difficult to manage because of its large size.

This logic implies that factories exhibit increasing returns to scale when they are small, and that the returns to scale eventually fall to a point where they are constant or diminishing. As a consequence, factories have an optimal size: to reach maximal efficiency, they should grow enough to exhaust the increasing returns. The same logic holds for grocery stores, distribution centers, law firms, hospitals, and other types of production units in the economy. The optimal size of these different units will differ widely depending on the technology and types of workers they employ.

The entire economy of a country is typically much larger than a single factory. National economies consist of many factories, many grocery stores, many banks, many schools, etc. National economies are therefore large enough that each production unit can operate at its optimal size. (Of course, some production units are small because they are poorly run or haven't had time to grow to their optimal size.) If such an economy were to double in size, this would occur primarily by the establishment of new production units that would grow to their optimal size. As a consequence, the economy would be roughly as efficient after it doubled in size as before.

This basic idea is called the *replication argument*: to double the size of an economy, one can simply replicate each production unit in the economy. This will use twice as much of each input and it will yield twice as much output. As a consequence, production at the national level will be constant returns to scale.

There are a number of reasons why the replication argument may not hold in reality. One reason is that it may not be possible to double all inputs to production. Some inputs to production are naturally fixed. Take, for example, land. There is a fixed amount of land on our planet. When the population of the planet doubles, it is not possible to also double the amount of land. This tends to push the economy towards diminishing returns to the other factors (capital and labor). For the last 150 years, this force has been rather weak, since land has not been a very important factor of production. Prior to the Industrial Revolution, land was an important factor of production and the economy as a whole clearly suffered from diminishing returns to scale as we discuss in detail in chapter XX [Malthus chapter]. But since then the importance of land has fallen to the point where it is sufficiently small that we often ignore it altogether.

Strictly speaking, fixed factors do not speak to whether the production function is constant returns to scale. The idea of the replication argument is to ask what would happen to output *if* all inputs were doubled. The fact that it is not possible to double all inputs is a different notion than the production function deviating from

constant returns. Nevertheless, from a practical point of view, these problems have similar implications, and fixed factors typically give rise to production functions that have diminishing returns to scale in the remaining inputs as we will see in chapter XX [Malthus chapter].

Another potentially important reason why the replications argument may not hold is externalities. A first-order fact about economies is that most economic activity is clustered in cities. Why? A natural explanation for this is that production units may have positive externalities on each other making it efficient for them to locate close to each other. Clustering of economic activity, of course, also yields congestion, a negative externality. For production to be constant returns to scale at the national level, these opposing externalities must exactly offset each other. It may well be that they do not.

A simple fact suggesting that the constant returns to scale assumption is not too far off is that large economies are not systematically richer than small economies. In particular, economies of vastly different sizes have attained levels of income per capita close to the level of the richest country at any given point in time. For example, Denmark and Germany have similar GDP per capita even though Denmark is about 15 times smaller than Germany. A more extreme comparison is Iceland and the United States, which also have similar GDP per capita even though Iceland is 1000 times smaller than the United States. The optimal size of production units in certain industries are sufficiently large that Iceland cannot compete in these industries (for example, automobiles), but there are enough industries in which the optimal size of production units are smaller for Iceland to reach a high level of income (given that it can trade with other countries).

The idea that national economies are roughly constant returns to scale has a number of important implications. One implication is that over the long run immigration neither increases nor decreases output per capita (as long as the immigrants assimilate in terms of, for example, levels of education). If the population increases by 10% due to immigration, the capital stock will eventually also increase by roughly 10% and output will increase by roughly 10%. The immigrants will take a lot of jobs, but they will also create a lot of jobs (since they will use their income to purchase goods and services produced by others).

3 Firm Behavior

Most production is performed by firms. These firms have access to various technologies that allow them to produce goods and services. The production function discussed above is meant to describe the technology firms have access to. But how do the firms make use of this technology? In other words, what assumption should we make about how firms behave? The assumption made in much of economic analysis is that firms maximize profits.

The combination of the production function and the assumption that firms maximize profits makes for an extraordinarily simple model of production, as we will see below. However, this model also abstracts from a great deal of interesting economics. Firms have shareholders, managers, employees, customers, suppliers, and usually also creditors. Each of these parties has interests of their own which in many cases conflict with the aim of maximizing the profits of the firm.

The shareholders are the owners of the firm. They may indeed want the firm's profits to be maximized (although they may also have other aims for the firm). The shareholders must, however, hire employees to carry out various production tasks. These employees do not benefit directly from higher firm profits. They have various aims of their own. These aims may include gaining experience, building a strong track record, and performing their duties faithfully. But the employees also have an incentive to shirk and relax, to the extent that they can get away with such behavior.

The owners will try to align the employees' incentives with their own aims by paying hard-working employees more and promoting these workers to better jobs. But in most cases there will exist a conflict of interest between the employees and the owners of the firm with the owners having a stronger interest in the employees working hard. This conflict of interest is called the principal-agent problem and is a central problem in the fields of contract theory and corporate finance.

In large firms, there are many layers of principal-agent problems. The shareholders appoint a board of directors. The board of directors hire top managers (a CEO, CFO, COO, CTO, etc.). The top managers hire a top layer of middle managers. The top layer of middle managers hire another layer of employees, and so on. Each such layer involves a principal-agent problem. In each case, the boss wants their workers to work hard since this will make their part of the firm perform well and thus reflect well on them. But the workers below the boss have an incentive to work less hard to the extent they can get away with this.

The relationship between the firm and its creditors also involves a principal-

agent problem. Most firms are limited liability corporations. This means that the shareholders' liability is limited to the funds they originally supplied the firm with. If the firm enters into debt that it cannot pay back, the creditors will face loss. This gives rise to a conflict of interest between the owners of the firm and the creditors of the firm. In particular, the owners have an incentive to take excessive risk when the firm's financial situation is weak. If such a gamble pays off, the profits accrue to the firm owners, while if the gamble does not pay off, the losses largely fall on the creditors. Head I win, tails you lose.

This discussion should make clear that the assumption that firms maximize profits is a simplifying assumption that sweeps various important issues under the rug. For a number of questions, it is critical to delve into these issues more carefully. But in many cases, the profit maximization assumption captures the essence of firm behavior in a simple manner and is therefore a good way to model firm behavior.

3.1 The Firm's Problem

Proceeding under the assumption that firms maximize profits, we can state the firm's problem as follows:

$$\max_{K,L} F(K,L) - rK - wL. \tag{8}$$

Here, $\max_{K,L}$ denotes that the firm chooses K and L to maximize the function stated to its right. The function $F(K,L)-rK-wL$ is the firm's profits. The first term in this function, $F(K,L)$, is the firm's revenue. You might think that firm revenue should take the form PY where P is the price of the firm's output and Y is the quantity of firm output. But recall that we can choose one price in the economy to be the numeraire. Here we are choosing the price of the firm's output to be the numeraire. This means that firm revenue is simply equal to firm output, which is $F(K,L)$. The second and third terms, $-rK-wL$, are firm costs. We imagine (for simplicity) that the firm rents the capital it uses at a rental rate of r and hires workers at a wage w.

An important simplifying assumption we make is that the firm takes the rental rate of capital and the wage as given. This means that we assume that the firm believes that its own actions do not affect these prices. Whether the firm rents a lot of capital or a little capital, the rental rate will stay fixed at r. Likewise, whether the firm hires many workers or few workers, the wage will remain fixed at w.

Why might this be a reasonable assumption? If the firm is small relative to the overall market for capital and labor, its influence on the prices in these markets will

be negligible. In the limit in which each firm is infinitesimally small in the capital and labor markets, we say that these markets are *perfectly competitive*. In this case, the actions of any one firm will not affect the price in these markets. Our assumption that the firm takes r and w as given, thus, amounts to assuming that the labor and capital markets are competitive. (Actually, we are also implicitly assuming that the goods market is perfectly competitive.)

In reality, labor, capital, and goods markets are often far from competitive. Firms in many cases have some degree of monopoly power in the goods market and monopsony power in capital and labor markets. This will affect their behavior (lead them to produce less, which will raise the price of goods and lower wages and the rental rate on capital). Here, we ignore this for simplicity.

3.2 Solving the Firm's Problem

The firm's problem is to choose the number of workers to hire and amount of capital to rent to maximize profits. How do we solve this problem? It turns out that the solution to this problem is found by separately maximizing profits with respect to K holding L fixed and maximizing profits with respect to L holding K fixed. This yields a system of two equations in the two unknown variables K and L. The solution to this system of equations is the firm's optimal level of capital and labor.

The algorithm described in the last paragraph is an application of a simple but powerful mathematical result:

$$\max_{x,y} f(x,y)$$

is given by the solution to the following two equations:

$$\frac{\partial f(x,y)}{\partial x} = 0 \tag{9}$$

$$\frac{\partial f(x,y)}{\partial y} = 0. \tag{10}$$

The right hand sides of these two questions are the *partial* derivatives of $f(x,y)$ with respect to x and y, respectively. The result states that to maximize a function of two variables x and y, one first solves for the x that maximizes $f(x,y)$ for each value of y (equation (9))—this involves differentiating $f(x,y)$ with respect to x holding y constant—then one solves for the y that maximizes $f(x,y)$ for each value of x (equation (10))—this involves differentiating $f(x,y)$ with respect to y holding x constant. Doing this yields two equations in two unknown variables (x,y). (Notice that the

partial derivatives on the left hand sides of equations (9) and (10) are functions of x and y.) The solution to this system of equations is the point at which $f(x, y)$ is maximized.

This idea works for functions of more than two variables. In general, one also needs to check second-order conditions. In the problems we encounter in this chapter (and for the most part in this book), we will set things up in such a way that the second order conditions will hold (e.g., assume that the function $f(x, y)$ is globally concave). We will therefore rarely discuss the second order conditions. But it is important to remember that, in the background, assumptions are being made that allow us to ignore these conditions.

We can now apply this mathematical result to the problem of the firm. First, we replace the general production function $F(K, L)$ with a particular production function such as the Cobb-Douglas production function AK^aL^{1-a}. This substitution yields

$$\max_{K,L} AK^aL^{1-a} - rK - wL. \tag{11}$$

The optimal level of capital for a given level of labor is then found by differentiating this function with respect to K holding L fixed and setting the resulting expression equal to zero. We have that

$$\frac{\partial}{\partial K}\left[AK^aL^{1-a} - rK - wL\right] = aAK^{a-1}L^{1-a} - r.$$

Setting this equal to zero yields

$$aAK^{a-1}L^{1-a} - r = 0.$$

Rearranging then yields

$$aAK^{a-1}L^{1-a} = r \quad \text{or} \quad a\frac{Y}{K} = r \tag{12}$$

Likewise, the optimal level of labor for a given level of capital is found by differentiating the function in (11) with respect to L holding K fixed and setting the resulting expression equal to zero. We have that

$$\frac{\partial}{\partial L}\left[AK^aL^{1-a} - rK - wL\right] = (1-a)AK^aL^{-a} - w.$$

Setting this equal to zero yields

$$(1-a)AK^aL^{-a} - w = 0.$$

Rearranging then yields

$$(1-a)AK^aL^{-a} = w \quad \text{or} \quad (1-a)\frac{Y}{L} = w \tag{13}$$

The preceding analysis shows that a firm's choices of K and L must satisfy equations (12) and (13) for the firm to be maximizing profits. Actually, these two equations fully describe the behavior of a profit maximizing firm in our setting. These are two equations in two unknown variables (K and L). One can easily solve this system of equations to express optimal K and L as a function of A, r, w, and a, all of which the firm takes as given.

At first blush, equations (12) and (13) may seem inscrutable. But they actually have simple economic interpretations. Consider first equation (12). Notice that the right-hand side of this equation is the rental price of a unit of capital. What about the left-hand side? It is the marginal product of capital $\partial Y/\partial K$ (see equation (5)). Equation (12) therefore states that a profit maximizing firm should rent capital to the point where the marginal product of capital is equal to the price of capital.

Recall that the marginal product of capital is the extra amount of output produced per unit of extra capital when the firm employs a small amount of extra capital holding other inputs to product fixed. It is therefore the marginal benefit of adding extra capital to the firm. The rental rate r is the marginal cost of adding extra capital to the firm. Equation (12), therefore, states that the firm should choose the level of capital at which the marginal benefit of adding more capital is equal to the marginal cost of adding that capital.

As we discussed above, the marginal product of capital is falling in the level of capital (see equation (6)). If the firm rents very little capital, the marginal product of capital will be high. As the firm rents more capital, the marginal product of capital falls. As long as the marginal product of capital is larger than the rental price of capital r, it pays the firm to rent more capital. At some point, the marginal product of capital will have fallen enough to equal r. At this point, the firm should not rent more capital since it will start to cost more to rent that extra capital than that extra capital will contribute in extra revenue.

Next consider equation (13). The logic is similar, but for labor. The right-hand side of this equation is the price of labor (the wage). The left-hand side is the marginal product of labor $\partial Y/\partial L$. Equation (13) states that a profit maximizing firm should hire workers to the point where the marginal product of an extra worker is equal to the wage the firm must pay that worker. If the firm hired very little labor,

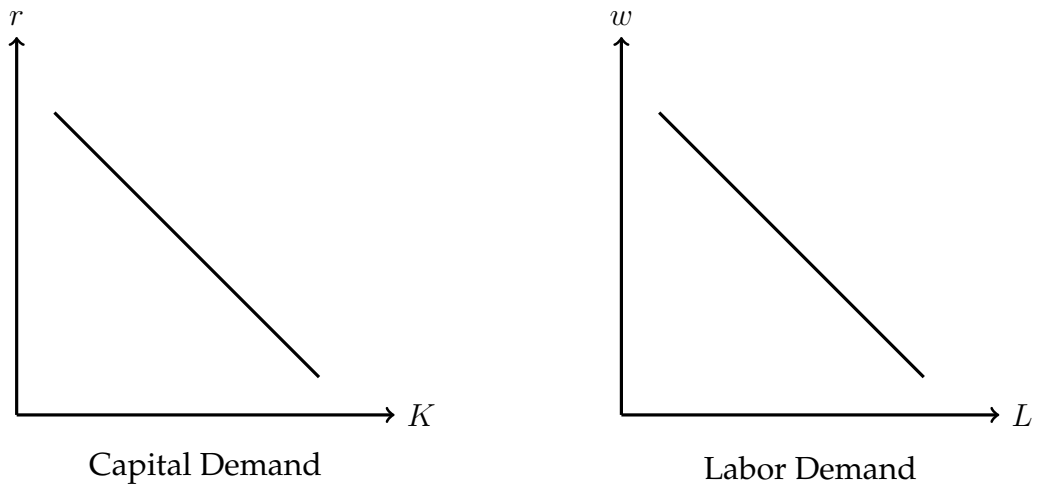

Figure 1: Capital and Labor Demand Curves

the marginal product of labor will be very high. It will therefore pay to hire more labor. As long as the marginal product of labor is above the wage, it pays to continue hiring more labor. The marginal product of labor is the marginal benefit of hiring more labor (the marginal revenue), while the wage is the marginal cost of hiring labor. At some point, the marginal product of labor will have fallen enough to equal the wage. At this point the firm should not hire more workers since the marginal revenue from extra workers will be less than the cost of those workers.

We derived equations (12) and (13) mathematically by solving the firm's profit maximization problem. But the preceding discussion makes clear that we could have derived these conditions from simple economic reasoning: profits are maximized when the marginal revenue of each factor is equal to the marginal cost of that factor.

It is common in economics to represent the optimal choices of households and firms graphically. Figure 1 does this for equations (12) and (13). The left panel of Figure 1 plots equation (12) in (K, r) space. (It is a convention in economics to put the price on the y-axis in graphs like these.) The line in the figure represents the set of points that satisfy equation (12). The equation is plotted for a particular value of the other variables that appear in the equation (A, L, a).

The set of points satisfying equation (12) is plotted in Figure 1 as being downward sloping. We can see why by inspecting equation (12). Suppose we start at a point for which equation (12) is satisfied. Then suppose we raise the rental rate of capital r. How do we need to change the quantity of capital K to make equation

(12) hold for this higher level of r? To restore equality the left hand side of equation (12) must increase. Recall that the left hand side is the marginal product of capital, which is falling in K. This means that K must fall for the left hand side to increase. This implies that the set of points that satisfy equation (12) is downward sloping in (K, r) space. (It is plotted as being linear. But this need not be the case.)

The curve in the left panel of Figure 1 has a name. It is the firm's capital demand curve. The preceding arguments demonstrate mathematically why the firm's capital demand curve is downward sloping. This notion is of course also quite intuitive. As before, our mathematical modeling and simple economic reasoning agree.

The right-hand panel of Figure 1 plots equation (13) in (L, w) space. Again, this is the set of points (L, w) that satisfy equation (13) for a particular value of the other variables appearing in that equation (A, K, a). This curve also has a name: the labor demand curve. We have plotted the labor demand curve as being downward sloping. This can be demonstrated using an analogous argument to the one we used above for the capital demand curve.

3.3 The Power of Competition

We have assumed that the labor market is perfectly competitive in our model. We have concluded that this implies that firms will pay workers their marginal product (equation (13)). In other words, workers will be paid the value of what they produce at the margin (the extra revenue the firm gets from hiring the last worker). Why don't the firms pay the workers less? Why don't they "exploit" the workers?

With perfect competition, each firm is held in check by competition from other firms in the labor market. If one firm tries to pay workers less than their marginal product, another firm will find it profitable to hire the workers away at higher wages. This will be the case at any wage level below the workers' marginal product. In this sense, competition is limiting the ability of the firms to exploit their workers.

If, instead, the firm is a monopsonist in the labor market, it can pay the workers less than their marginal product. The only alternative for the workers is then not to work at all or to move to another location. In many cases, this is not really a feasible alternative. So, the firm in this case has a great deal of power over the workers and can exploit that power. The paradigm example of this is a large employer in a small town (e.g., the mining company in a mining town). Such companies can exploit their workers because they don't face the discipline of competition in the labor market. The power of competition to restrain exploitation is perhaps the most

important lesson economics has to offer.

The labor market in the real world is far from perfectly competitive. There is growing evidence that workers are in many cases paid less than their marginal product because firms can in fact exploit monopsony power over them. Given this, why do we assume that the labor market is competitive? One reason is to illustrate the power of competition. But another reason is that it simplifies the analysis. Models with imperfect competition are usually more complex to analyze than models with perfect competition. This biases our modeling choices towards models with perfect competition.

An important danger to keep in mind in this context is the danger that we forget that we are making an extreme assumption when we assume perfect competition. If we forget this, we may slip into thinking that the implications of our models are true. We may even forget about the very possibility that markets might be imperfectly competitive and wages might not be equal to the marginal product of labor. The Nobel Prize winning economist Daniel Kahneman called this "theory-induced blindness." He said: "Once you have accepted a theory, it is extraordinarily difficult to notice its flaws" (Kahneman, 2011). Using economic theory wisely is a tricky business since this involves absorbing insights from the theory but at the same time guarding against theory-induced blindness.

4 Completing the Model

The firm's problem analyzed in the last section yielded two equations—the firm's capital demand curve (equation (12)) and the firm's labor demand curve (equation (13)). From the firm's perspective, the rental rate of capital r and the wage w are exogenous. In other words, each firm takes the rental rate and the wage as given. But from the perspective of the economy as a whole, these variables are endogenous. They are determined by the interaction of supply and demand in the capital and labor markets. The firm's problem yields the demand curves in these markets. To determine the rental rate and the wage (and the quantity of capital and labor), we need supply curves in these markets.

To keep things simple, in this chapter, we assume that both capital and labor are inelastically supplied. This means that the supply of capital and labor is not affected by the price of capital and labor. For capital, one can think of this assumption as a short run assumption. In the short run, the amount of capital is given. It takes

times to construct more capital. So, the amount of capital supplied is not affected by movements in the price of capital in the short run. It is harder to justify the assumption that labor is inelastically supplied even in the short run. It is best to think about that assumption as a simplifying assumption. Chapter XX [Labor supply chapter] will develop a more sophisticated model of labor supply.

Mathematically, the assumption that capital and labor are inelastically supplied can be stated as

$$K = \bar{K}, \tag{14}$$

and

$$L = \bar{L}, \tag{15}$$

where $\bar{K}$ and $\bar{L}$ are parameters (i.e., given exogenously).

We now have four equations (equations (12), (13), (14), (15)) in four unknown variables: r, K, w, L. These four equations are demand and supply curves in the capital and labor market and they can be solved to determine the price and quantity in these markets.

Before proceeding further, it is important to note that I have slipped in an important assumption without mention. In equation (12), K denotes capital demanded. In equation (14), K denotes capital supplied. By using the same symbol to represent both capital demanded and capital supplied, I have implicitly assumed that capital demanded and capital supplied end up being equal.

A more careful description of the model distinguishes between these two variables. We might denote capital demanded by K^d and capital supplied by K^s. In this case, equation (12) reads $aA(K^d)^{a-1}L^{1-a} = r$ and equation (14) reads $K^s = \bar{K}$. We then need an extra equation since we have an extra variable (K^d and K^s rather than just K).

The extra equation comes from the assumption that the capital market clears. What this assumption means is that the price in the capital market will end up being whatever value is needed to ensure that capital demand is equal to capital supply: $K^d = K^s$. Given this market clearing assumption, we can define $K = K^d = K^s$.

We could carefully distinguish between quantity demanded and quantity supplied in each market in our model. This would imply that we would have three variables for each market (quantity demanded, quantity supplied, and price) and three equations for each market (demand, supply, and market clearing). Since market clearing simply sets the quantity demanded equal to the quantity supplied in every market, we will generally simplify the exposition by implicitly assuming from

the outset that markets clear and using the same variable for quantity in both the demand and supply curve.

The model we are developing has three markets: a capital market, a labor market, and a goods market. The above analysis has yielded demand and supply curves in the capital and labor markets. What about the goods market? Recall that Walras' Law implies that if all but one market in a model clear, then the last market also clears. This means that we need not develop a demand and supply curve for the goods market. In addition, we have chosen the price of goods as the numeraire. This means that we are setting this price equal to one (denominating other prices in terms of goods). As a consequence, there is only one additional endogenous variable in the model: the quantity of goods Y.

To determine the quantity of goods, we need one more equation that relates the quantity of goods to the other endogenous variables (and perhaps also exogenous variables and parameters). One equation that works for this purpose is the production function:

$$Y = AK^a L^{1-a}. \tag{16}$$

This equation completes our model.

5 Equilibrium

In the preceding sections, we have developed a general equilibrium model with a labor market, capital market, and goods market. The model consists of five equations in five unknown endogenous variables. The five equations are

$$aAK^{a-1}L^{1-a} = r, \tag{17}$$

$$(1-a)AK^a L^{-a} = w, \tag{18}$$

$$K = \bar{K}, \tag{19}$$

$$L = \bar{L}, \tag{20}$$

$$Y = AK^a L^{1-a}, \tag{21}$$

while the five endogenous variables are K, L, r, w, and Y.

The solution of a model in economics is usually referred to as an *equilibrium*. An equilibrium refers to the outcome for the endogenous variables in the model when markets clear, i.e., when supply equals demand in all markets. This use of the term

equilibrium is different from in some areas of science where equilibrium is used to refer to a system that is at rest. An equilibrium does not indicate that the economy is in any sort of balance other than that supply equaling demand in all markets. The economy may be in a strong boom or a deep recession and we still refer to the outcome in that state as an equilibrium.

To solve for the equilibrium of a model means to solve for the endogenous variables in terms of only exogenous variables and parameters. In the model we have developed, this means to solve the five equations for the five unknown endogenous variables. In other words, rewrite the five equations such that the endogenous variables appear on one side of the equations and only exogenous variables and parameters appear on the other side.

The model we have developed is sufficiently simple that solving it is trivial. First, equations (19) and (20) already have K and L, respectively, expressed in terms of exogenous variables ($\bar{K}$ and $\bar{L}$). We can then use these two equations to plug in for K and L in the other three equations and arrive at the following solution

$$\begin{aligned} K &= \bar{K}, \\ L &= \bar{L}, \\ r &= aA\bar{K}^{a-1}\bar{L}^{1-a}, \\ w &= (1-a)A\bar{K}^{a}\bar{L}^{-a}, \\ Y &= A\bar{K}^{a}\bar{L}^{1-a}. \end{aligned}$$

The solution to our model is "trivially" simple because we have assumed that both capital and labor are inelastically supplied. The next two chapters (and later chapters in this book) develop models with more interesting solutions.

6 Factor Shares

Some of the revenue firms receive from selling the output they produce is paid to workers as wages. Some is used to purchase intermediate inputs. Some is used to pay taxes. The rest is paid to the owners of capital as interest on loans, dividends, and share buy backs. In the model we have developed, we have ignored intermediate inputs and taxes. This means that all revenue is paid to workers, to the owners of capital, and to the firm's owners as profits. It is instructive to calculate what share of the firm's revenue accrues to each of these three parties in our model.

Let's start with the share of output that workers receive, which we call the *labor share*. Workers receive a wage w. The quantity of labor they supply is L. Their total compensation is therefore wL. Starting from equation (18), we have that

$$w = (1-a)AK^aL^{-a} = (1-a)\frac{AK^aL^{1-a}}{L} = (1-a)\frac{Y}{L},$$

where the last equality uses the production function—equation (21). Multiplying through by L and dividing through by Y in this equation then yields

$$\frac{wL}{Y} = 1-a. \tag{22}$$

In other words, the labor share in our economy is constant and equal to $1-a$.

The owners of capital receive a rental rate of r per unit of capital they supply. They supply K units of capital. Their total compensation is therefore rK. Starting from equation (17), we have that

$$r = aAK^{a-1}L^{1-a} = a\frac{aAK^aL^{1-a}}{K} = a\frac{Y}{K}.$$

Multiplying through by K and dividing through by Y in this equation then yields

$$\frac{rK}{Y} = a. \tag{23}$$

This shows that the capital share is constant and equal to a.

How much income is then left as profits for the owners of the firms? Clearly, none. The labor share is $1-a$ and the capital share is a. These two add up to one. In other words, all of the revenue of the firms is dispersed as payments to the factors of production. Profits are zero. There is nothing left for the owners of the firm.

Why are profits zero in our model? This flows from two assumptions we have made. First, factor markets are perfectly competitive. This implies that labor and capital are paid their marginal product: $w = \partial F(K,L)/\partial L$ and $r = \partial F(K,L)/\partial K$. Second, the production function is constant returns to scale. Mathematically, a production function is constant returns to scale if it is homogeneous of degree one. Euler's theorem then states that

$$F(L,K) = \frac{\partial F(L,K)}{\partial L}L + \frac{\partial F(L,K)}{\partial K}K.$$

Combining these two facts yields

$$Y = wL + rK.$$

In our model, we have assumed (for simplicity) that firms rent all the capital that they use. This means that the owners of the firm don't own any of the capital. Real-world firms usually own most of the capital that they use. This means that the owners of the firm are the owners of much of the capital. However, firms do finance the purchase of their capital partly with borrowed funds. This part of their capital is effectively rented (just like in our model). But much of the capital is supplied by the firm's owners through equity and retained earnings. This means that when the firm pays dividends to its owners it is difficult to assess what fraction of the dividend payments are returns on capital and what fraction are due to pure profits.

An additional practical complication is that standard accounting measures used in the corporate world and for filing taxes divide revenue up in a somewhat different way than the way we do when analyzing our model. Interest expenses are views as costs. (Wages, payment for intermediate inputs, and taxes are also viewed as costs.) However, the returns to equity holders are categorized as accounting profits. Interest expenses are the returns paid to debt holders and are a part of returns to capital. Accounting profits lump together returns to equity capital—the other main part of returns to capital—and pure profits. Pure profits are any profits earned by the firm that are over and above the competitive return on capital. It is important to remember that accounting profits and pure profits as economists think about that concept are not the same concept.

6.1 Why Cobb-Douglas?

The derivations above assume not only that the production function is constant returns to scale, they assume that the production function takes the Cobb-Douglas form. The Cobb-Douglas production function is a popular choice partly because it is mathematically convenient: it is a simple function that lends itself to tractable modeling. However, arguably, a more important reason for its popularity is its prediction regarding factor shares.

We saw above that the labor share in the model we have developed in this chapter is constant. This prediction is not at all general. In fact, it is quite special to the Cobb-Douglas production function. As we discuss in more detail below, this prediction is a consequence of the fact that the Cobb-Douglas production function implies an elasticity of substitution between capital and labor equal to one. Other production functions have different implications about this elasticity of substitution and therefore do not imply a constant labor share.

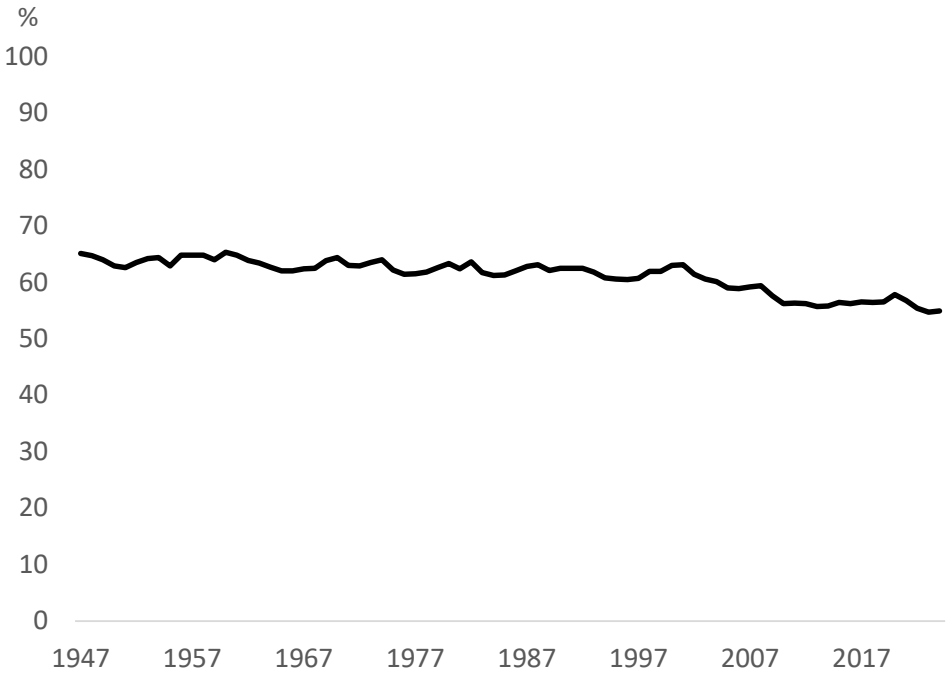

Figure 2: Labor Share in U.S. Non-farm Business Sector

Note: The source of these data is the U.S. Bureau of Labor Statistics.

Figure 2 plots the labor share in the non-farm business sector of the U.S. economy from 1947 to 2024. Over this period, it varied relatively little. Its average value was about 61%. Before 2000, it fluctuated in a very narrow band between 60 and 65%. More recently, it has fallen slightly to around 55%. (More on this below.) The British economist John Maynard Keynes called the high degree of stability of the labor share "one of the most surprising, yet best-established, facts in the whole range of economic statistics" (Keynes, 1939) and Nicholas Kaldor famously argued that it was one of the key stylized facts of economic growth (Kaldor, 1961).

It is not at all obvious or inevitable that the labor share would remain so stable over such a long period of time. Ever since the Industrial Revolution (and before), economic growth has been accompanied by a heavy dose of worry that labor saving technologies would supplant workers. These worries led to machine breaking riots by the Luddites in the early 19th century in England. At the time of this writing, it is artificial intelligence and robotics that are causing such worries. But if machines are taking all the jobs, the share of income accruing to labor should be falling. Figure 2 shows that this has not occurred (yet) to an appreciable degree. In particular, over the course of the second half of the 20th century, despite huge amounts of

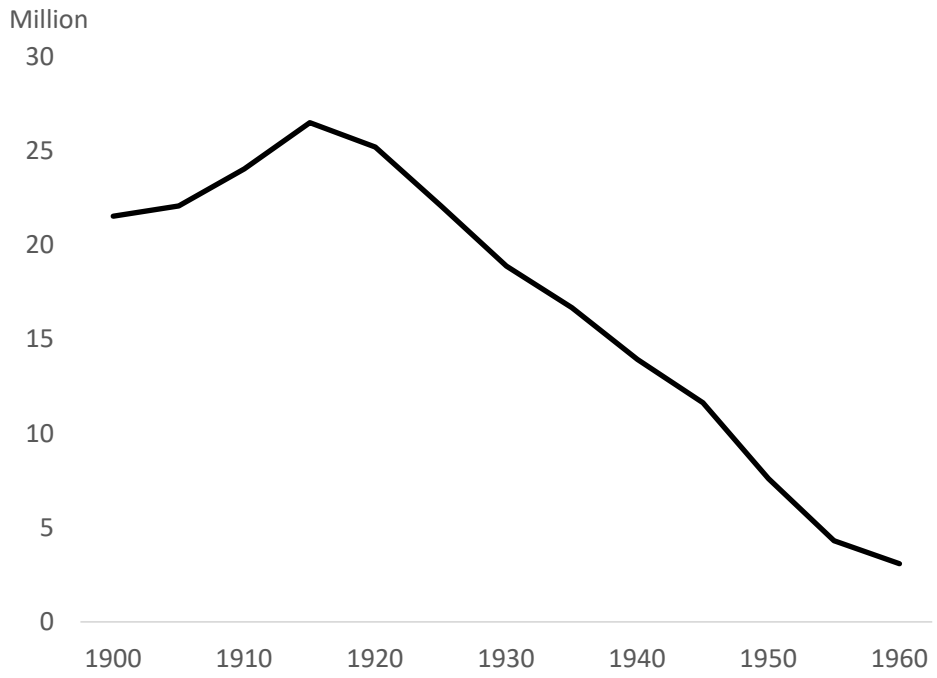

Figure 3: Number of Horses and Mules in the United States, 1900-1960

Note: The source of these data is Ensminger (1969) [XX Look into this Source XX]

technological change, the labor share was largely unchanged.

Contrast this with the fate of horses. Horses were once an essential part of the economy both in agriculture and transportation. But then came the internal combustion engine. Over the course of the first half of the 20th century, machines driven by internal combustion engines replaced horses virtually completely in the economy. Figure 3 illustrates this by plotting the population of horses and mules in the United States from 1900 to 1960. From its peak in 1915, the horse population in the United States fell by almost 90%. Today, horses are used mainly for sport and recreation. The internal combustion engine really did take virtually all the jobs of horses.

Clearly, humans are vastly more versatile workers than horses. But is this going to be enough? Is it perhaps only a matter of time until our usefulness goes the way of horses? This was certainly the prediction of Nobel Prize winning economist Wassily Leontief back in 1983 (Leontief, 1983). He believed that humans had done well because machines had up to that point been dumb and needed humans to carry out more complex "mental" tasks. However, once machines became smart—which he believed was already beginning to occur—he argued that "labor's role as an indispensable "factor of production" [would] progressively diminish" and "techno-

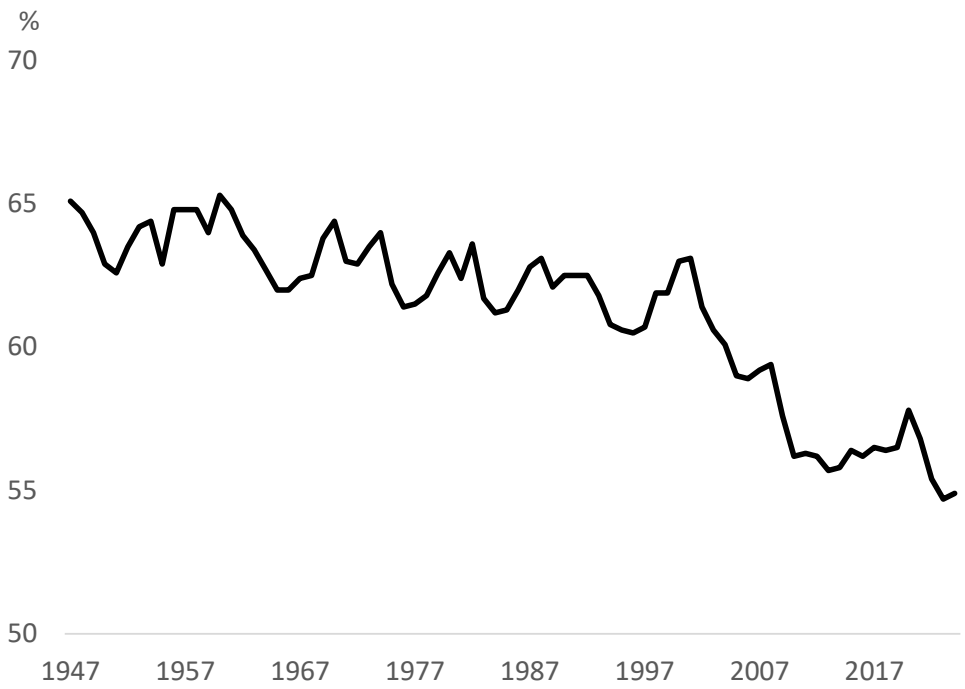

Figure 4: Labor Share in U.S. Nonfarm Business Sector, A Closer Look

Note: The source of these data is the U.S. Bureau of Labor Statistics. The data are annual. The sample period is 1947-2024.

logical unemployment" would beckon.

6.2 Has the Labor Share Started to Fall?

Over forty years have passed since Leontief's pessimistic prognosis for human usefulness and we seem to be doing just fine. Or are we? Figure 4 provides a closer look at the labor share by tightening the vertical axis relative to Figure 2. This close-up reveals that the labor share seems to have been on an ever-so-slight downward trend over the course of the second half of the 20th century which has accelerated substantially since 2000. Between 2000 and 2024, the data in Figure 4 indicate that the labor share fell from 63% to 54%. Perhaps this is the beginning of the long-feared takeover of the machines.

This apparent fall in the labor share since 2000 has led to a substantial amount of new research. One strand of this research has argued that some part or perhaps even all of the apparent fall in the labor share is not actually real, but rather due to mismeasurement. Koh, Santaeulàlia-Llopis, and Zheng (2020) show that accounting changes made by the U.S. Bureau of Economic Analysis (BEA) over the past quar-

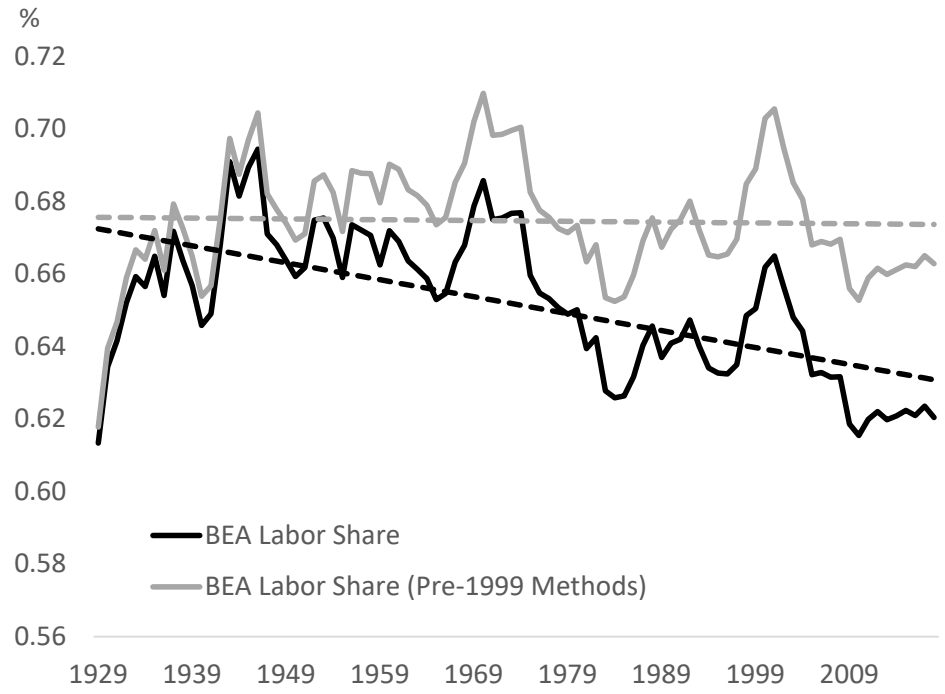

Figure 5: Economy-Wide U.S. Labor Share

Note: This figure is based on data from the U.S. Bureau of Economic Analysis. It reproduces Figure 2 from Koh, Santaeulàlia-Llopis, and Zheng (2020). The data are annual. The sample period is 1929-2018.

ter century which gradually reclassified spending on intellectual property products (chiefly software, research and development, and artistic originals) as investment in intangible capital—as opposed to as intermediate inputs—can explain the entire fall in the labor share since 1950.

Figure 5 plots the BEA's measure of the labor share for the whole economy along with an alternative measure of the labor share constructed by Koh, Santaeulàlia-Llopis, and Zheng using the BEA's pre-1999 methods. The difference is striking. While the official measure using current BEA methods has a substantial downward trend all the way back to 1950, the version using pre-1999 methods has no downward trend at all. Mechanically, this difference in the trends is due to the growing importance of intellectual property in the U.S. economy. It seems that all of the fall in the labor share can be explained by the BEA's changing methodology.

But does this mean that the fall in the labor share is not real? Presumably, the BEA updates its measures to improve measurement. Perhaps the new methodology just reveals a downward trend in the labor share that the old methods erroneously did not capture.

Recent research by economist Robert Barro implies that this is not the case. Barro argues that the standard methodology used to measure output in the economy—Gross Domestic Product (GDP)—double counts investment (Barro, 2021). Investment is counted once when it occurs (the I in $Y = C + I + G + NX$), but then it is effectively counted again when the capital that is built yields a service flow over time, i.e., when it is used as an input to produce goods.

A crucial component of national income accounting is to not double count intermediate inputs. GDP counts the value of final outputs (such as automobiles), but not the intermediate inputs that are used to produce those final outputs (steel, tires, computer chips, batteries, etc.). But why then is investment counted? Isn't a machine used for future production conceptually the same as steel and computer chips, i.e., an intermediate input? Arguably it is.

But output measures are used for many different purposes. One of those purposes is to assess whether the economy is expanding or contracting. If investment were treated as an intermediate input, output would fall when production shifts from consumption to investment. Suppose a country discovers a new natural resource and decides to invest heavily in exploiting this resource. Consumption may fall (due to crowding out) while this investment boom is running its course. But the notion that economic output is contracting doesn't seem to capture what is going on.

This is a tricky issue. In the end, the creators of national income and product accounts chose to include investment in what came to be the main measure of economic output: GDP (actually GNP at the time). This choice reflected the importance placed on using these statistics to assess the state of the business cycle.

When we think about factor shares, our focus is different. We are interested in factor shares as a measure of the relative welfare of workers and owners of capital. For this, GDP is arguably not the correct starting point. We want to know how much workers and owners of capital can afford to consume. Shares of GDP do not answer this question because GDP double counts investment. A substantial part of the problem has to do with depreciation of capital. The owners of capital must use part of the income they receive simply to replace and repair old worn out capital. This part of their income does not support their consumption. The factor shares we are interested in should therefore adjust for depreciation.

The national income and product accounts include a measure of output that subtracts depreciation. This is called net domestic product (NDP) as opposed to gross domestic product (GDP). Factor shares of NDP come closer to measuring the rela-

tive size of income available for consumption for labor and capital. Barro argues, however, that an additional adjustment is needed to take account of the fact that the economy is growing, which means that the capital stock must grow. Some income must be used to build that new capital and that income is not available for consumption.

These adjustments lower the capital share and increase the labor share. Furthermore, if the economy is shifting towards forms of capital with higher depreciation rates, these adjustments become larger over time. As depreciation rates rise, the gross capital share (i.e., capital share of GDP) will rise relative to the net capital share (i.e., capital share of NDP). Equivalently, the gross labor share (which we plot in Figure 2 and 4) will fall relative to the net labor share.

The economy has indeed been shifting towards forms of capital that have high depreciation rates. The share of structures (which have relatively low depreciation rates) in investment has fallen over time, while the share of intellectual property products (which have relatively high depreciation rates) has risen. Software is a significant part of this story. Software has a high depreciation rate and is a growing part of the economy. This implies that a larger and larger part of gross output is needed to replenish obsolete old capital. So, while the gross labor share is falling, it is not as clear that the fraction of output available for consumption that accrues to labor is falling.

6.3 The Falling Relative Price of Investment

The measurement issues discussed above suggest that the labor share (defined appropriately) has not fallen as much as Figure 4 indicates. However, these measurement issues are still debated and controversial. Many scholars retain the view that the labor share has trended downward since 2000 and that this is a break relative to the previous 50 years. There are quite a number of explanations that have been proposed for this change. One prominent explanation is that the root cause has been improvements in machines: rising productivity of machines, falling prices of machines, or both.

This idea seems intuitive: If machines become vastly more productive, surely machines will be used more intensively in production and a larger share of income will accrue to the owners of this larger quantity of machines. Perhaps surprisingly, this is not necessarily true. In fact, it is not true if the economy's production function is Cobb-Douglas.

Suppose the productivity of machines is augmented by a factor z. We can then write the production function as

$$Y = A(zK)^a L^{1-a}. \tag{24}$$

In this case, the rental rate on capital becomes

$$r = az^a AK^{a-1} L^{1-a} = a\frac{az^a AK^a L^{1-a}}{K} = a\frac{Y}{K}. \tag{25}$$

Multiplying through by K and dividing through by Y in this equation yields

$$\frac{rK}{Y} = a.$$

Notice that this is the same expression for the capital share as we had derived before. The capital share turns out to be a as before. In other words, it is independent of the capital-augmenting productivity factor z. While the marginal product of capital rises with z (the second expression in equation (25)), output rises by the same factor (the numerator of the third expression in equation (25)). This means that the share of income accruing to capital remains unchanged. As it turns out, labor and capital benefit equally (in proportional terms) from the increase in the productivity of the machines when the production function is Cobb-Douglas.

This invariance result turns out to be a "knife-edge" case. More generally, improvements in machines do affect the labor share. But the direction in which they do depends on the shape of the production function. To see this, we must introduce a more general class of production functions.

We also take this opportunity to introduce notation for the time at which a variable is defined. This will be helpful when we talk about how the economy evolves over time. For simplicity, we assume that time is discrete, i.e., the economy moves from one discrete period (e.g., a year) to the next. We use t subscripts on variables to denote the time period of that variable. For example, Y_t will denote output in period t.

Using this notation, consider the following production function

$$Y_t = \left[a(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}. \tag{26}$$

where $A_{K,t}$ and $A_{L,t}$ denote two forms of productivity. $A_{K,t}$ is capital-augmenting productivity. It represents technologies that make capital more productive. $A_{L,t}$ is labor-augmenting productivity. It represents technologies that make labor more productive.

Production function (26) may look a bit overwhelming and perhaps a bit odd at first sight. But there is a method to the madness. This production function is actually quite carefully crafted. To appreciate this, it is useful to consider a few of its basic properties. First, it is easy to see that this production function is constant returns to scale. Suppose we multiply both inputs (K_t and L_t) by a factor γ (Greek letter gamma). Then we have

$$\begin{aligned}
&\left[a(A_{K,t}\gamma K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}\gamma L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \\
=&\left[a\gamma^{\frac{\sigma-1}{\sigma}}(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)\gamma^{\frac{\sigma-1}{\sigma}}(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \\
=&\left[\gamma^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}\left[a(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \\
=&\gamma\left[a(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}.
\end{aligned}$$

This shows that multiplying all inputs by a factor γ increases output by that same factor, which implies that the production function is constant returns to scale. Notice, how the exponent $\frac{\sigma}{\sigma-1}$ outside the square bracket, cancels the exponent $\frac{\sigma-1}{\sigma}$ inside the square bracket once we factor γ out of the main square bracket in this calculation. The role of the outermost exponent $\frac{\sigma}{\sigma-1}$ in this production function is to make sure that the production function is constant returns to scale.

With production function (26), the profit maximization problem of the firm is

$$\max_{K_t,L_t}\left[a(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} - r_tK_t - w_tL_t. \tag{27}$$

Differentiating the firm's profit function with respect to K_t and setting the resulting expression equal to zero yields

$$\left(\frac{\sigma}{\sigma-1}\right)\left[a(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}} + (1-a)(A_{L,t}L_t)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}-1} a\left(\frac{\sigma-1}{\sigma}\right)(A_{K,t}K_t)^{\frac{\sigma-1}{\sigma}-1}A_{K,t} = r_t.$$

Canceling terms and using the fact that

$$\frac{\sigma}{\sigma-1} - 1 = \frac{1}{\sigma-1} = \left(\frac{\sigma}{\sigma-1}\right)^{\frac{1}{\sigma}}$$

allows us to rewrite this equation as

$$aY_t^{\frac{1}{\sigma}}A_{K,t}^{\frac{\sigma-1}{\sigma}}K_t^{-\frac{1}{\sigma}} = r_t.$$

Further manipulation yields

$$aA_{K,t}^{\frac{\sigma-1}{\sigma}}\left(\frac{Y_t}{K_t}\right)^{\frac{1}{\sigma}} = r_t. \tag{28}$$

This is the firm's capital demand curve when the production function is given by equation (26).

Differentiating the profit function with respect to L_t, setting the resulting expression equal to zero, and performing a similar set of manipulations as above yields the firm's labor demand curve

$$(1-a)A_{L,t}^{\frac{\sigma-1}{\sigma}}\left(\frac{Y_t}{L_t}\right)^{\frac{1}{\sigma}} = w_t. \tag{29}$$

The parameter σ determines the elasticity of both labor demand and capital demand with respect to the prices of labor and capital, respectively. To see this, take a natural logarithm on both sides of equation (28) and rearrange to get

$$\ln K_t = -\sigma \ln r_t + \ln Y_t + (\sigma - 1)\ln A_{K,t} + \sigma \ln a. \tag{30}$$

Differentiating this with respect to $\ln r_t$ then yields the elasticity of capital demand:

$$\frac{\partial \ln K_t}{\partial \ln r_t} = -\sigma.$$

Similar manipulation of equation (29) yields that the elasticity of labor demand is also equal to $-\sigma$.

The parameter σ also determines the ease with which capital and labor can be substituted one for the other in production. To see this, divide equation (28) by equation (29) to get

$$\frac{a}{1-a}\left(\frac{A_{K,t}}{A_{L,t}}\right)^{\frac{\sigma-1}{\sigma}}\left(\frac{L_t}{K_t}\right)^{\frac{1}{\sigma}} = \frac{r_t}{w_t}. \tag{31}$$

Take a natural logarithm of both sides of this equation and rearrange to get

$$\ln\left(\frac{L_t}{K_t}\right) = \sigma \ln\left(\frac{r_t}{w_t}\right) - (\sigma-1)\ln\left(\frac{A_{K,t}}{A_{L,t}}\right) - \ln\left(\frac{a}{1-a}\right). \tag{32}$$

This equation shows that a one-percent increase in the price of capital relative to the price of labor results in a σ-percent decrease in the amount of capital used in production relative to the amount of labor used. Here we equate log changes with percentage changes. These are equal up to a first order approximation for small changes. (You can see this by taking a first order Taylor approximation of $\ln x$ around x_0.)

When σ is small, a one-percent fall in the relative price of capital versus labor results in a small increase in capital used in production relative to labor. This means that capital and labor are not easily substituted in production. It takes a large change in their relative price to induce an appreciable shift in their use in production.

Conversely, when σ is large, a one-percent fall in the relative price of capital versus labor results in a large shift towards capital and away from labor in production. In this case, capital and labor are easily substituted, and even a small change in their relative price is enough to appreciably shift their use in production.

The coefficient in front of $\ln(r_t/w_t)$ in equation (32) is called the *elasticity of substitution* between capital and labor. Formally, the elasticity of substitution is defined to be the curvature of the isoquant of the production function:

$$\frac{\partial \ln(L/K)}{\partial \ln(\text{Slope})} \quad \text{where} \quad \text{Slope} = \frac{\partial F(\cdot)/\partial K}{\partial F(\cdot)/\partial L}$$

With competitive factor markets, the marginal products of capital and labor are equal to their prices. So, we have that $\partial F(\cdot)/\partial K = r_t$ and $\partial F(\cdot)/\partial L = w_t$, which means that the elasticity of substitution is the coefficient in front of $\ln(r_t/w_t)$ in equation (32).

The production function (26) derives its name from the fact that it yields a constant elasticity of substitution between capital and labor: it is usually referred to as the *constant elasticity of substitution* (CES) production function. The fact that the elasticity of substitution as well as the elasticity of demand for both capital and labor are all constant makes the CES production function a very convenient production function to work with.

The Cobb-Douglas production function is actually a special case of the CES production function. When $\sigma \to 1$, the CES production function converges to the Cobb-Douglas production function. In other words, the Cobb-Douglas production function is a CES production function with an elasticity of substitution between capital and labor equal to one. I leave the proof of this as an exercise for interested readers. (Hint: Take natural logs and then a Taylor series approximation.) It is easier to see that the capital and labor demand curves—equations (28) and (29)—converge to the corresponding curves for a Cobb-Douglas production function—equations (12) and (13)—when $\sigma \to 1$, and that the elasticity of substitution in equation (32) becomes one.

Deriving an expression for the labor share when the production function is CES involves a little bit more manipulation than in the Cobb-Douglas case. We start with the capital demand equation (28). First, we raise both sides of this equation to the power σ. Then we multiply both sides by K_t/Y_t and divide both sides by $r_t^{\sigma-1}$. This yields

$$\frac{r_t K_t}{Y_t} = a^\sigma \left(\frac{A_{K,t}}{r_t}\right)^{\sigma-1}. \tag{33}$$

The left-hand side of this equation is the capital share. Since profits are zero, the labor share is one minus the capital share. This implies that the labor share is

$$s_{L,t} = 1 - a^{\sigma} \left(\frac{A_{K,t}}{r_t} \right)^{\sigma - 1}. \tag{34}$$

When $\sigma = 1$ (the Cobb-Douglas case), this expression simplifies to a constant $s_L = 1 - a$. However, when $\sigma \neq 1$, the labor share is no longer constant. It is a function of two variables: capital-augmenting productivity $A_{K,t}$ and the price of capital r_t. Furthermore, the direction in which these variables affect the labor share depends on whether σ is larger than one or smaller than one.

If the elasticity of substitution between capital and labor σ is larger than one, an increase in capital-augmenting productivity $A_{K,t}$ or a reduction in the price of capital r_t increases the capital share and reduces the labor share. In this case, capital and labor are said to be *gross substitutes*.

When $\sigma > 1$, the capital demand curve is relatively "elastic," i.e., its slope is small. (To see this, solve equation (30) for $\ln r_t$ and plot the result with $\ln r_t$ on the y-axis and $\ln K_t$ on the x-axis.) This means that an increase in capital (holding output fixed) will result in a relatively small decrease in the price of capital. More precisely, a 1% increase in capital will result in a decrease in the price of capital that is smaller than 1%. Since the price of capital falls less (in proportional terms) than the quantity of capital increases (per unit of output produced) in this case, the product $r_t K_t$ increases. In other words, the share of each unit of output that accrues as a return to capital ($r_t K_t$) increases in this case. This means that labor share must fall.

The converse is true when $\sigma < 1$. In this case, the capital demand curve is relatively inelastic and the price of capital falls more than one-for-one when the quantity of capital increases (holding output fixed). This implies that $r_t K_t$ decreases when the quantity of capital increases, which decreases the capital share and increases the labor share. In this case, capital and labor are said to be *gross complements*.

There is some debate in the academic literature regarding whether capital is becoming cheaper or more expensive over time. The price of housing has risen considerably in the United States between 1980 and 2020. However, the price of investment goods has fallen. Figure 6 plots the price of investment goods relative to the price of output in the United States since 1947. Starting in the early 1980s, the relative price of investment goods has fallen steadily. By 2024, it had fallen by almost 40%.

Karabarbounis and Neiman (2014) argue that this large fall in the price of in-

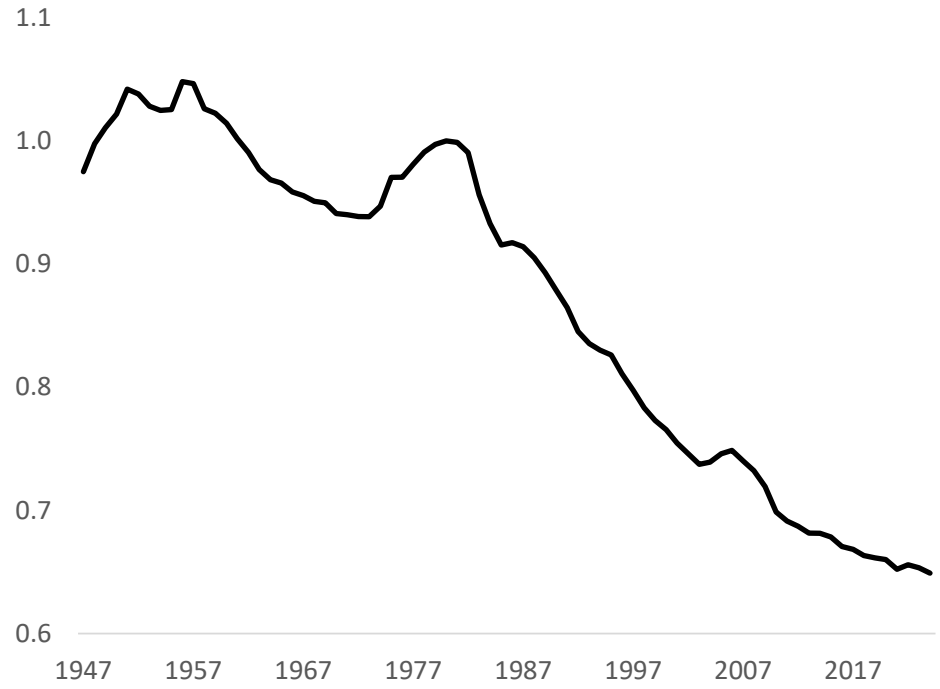

Figure 6: The Price of Investment Goods Relative to the Price of Output

Note: The source of these data is the U.S. Bureau of Economic Analysis. The sample period is 1947-2024.

vestment goods has been a major contributor to the fall in the labor share over the past few decades. They present a model with two types of goods: investment goods and consumer goods. Technological progress in the production of investment goods lowers the price of these goods. This results in a reduction in the rental price of capital r_t. From equation (34), we see that this will lower the labor share if $\sigma > 1$.

6.4 Estimating the Elasticity of Substitution σ

But is it reasonable to think that $\sigma > 1$? When we casually observe the production process of a particular firm it often seems as though there is little scope for substitution of capital for labor. Each machine needs a certain number of workers to operate it. Perhaps output can be increased some by adding more workers for a given number of machines or by adding more machines for a given number of workers. But it often seems like this would quickly run into severely diminishing returns. These types of casual observations suggests that σ is quite small when it comes to the technology used at a given firm or plant.

However, even if each and every technology used in the economy allows for

little substitution between labor and capital, this does not mean that the elasticity of substitution at the level of an industry or at the level of the economy as a whole is small. The reason for this is that when the price of capital changes relative to the price of labor, firms can change from using one technology to using another.

In places were labor is very cheap, the technologies used are very labor intensive. A simple example is the washing of dishes. This can be done by hand or using a dishwasher. Hand-washing is labor intensive, while using a dishwasher is more capital intensive. The use of dishwashers is much more prevalent in countries where labor is expensive. Conversely, in countries were labor is cheap, it is considered absurd by many to buy a dishwasher. Hiring a housekeeper is more economical. As wages rise, people switch from the one technology to the other (both at home and in restaurants). The same logic applies to most other production processes.

The economist Henry Houthakker made this point in a particularly stark way in a famous 1955 article (Houthakker, 1955). Houthakker supposed that each technology available for production involved using labor and capital in fixed proportions. This is the limit case were $\sigma = 0$, i.e., labor and capital are not substitutable at all. Economists refer to this production function as the Leontief production function (in honor of Wassily Leontief). While each technology was Leontief in Houthakker's model, he supposed that there existed many such technologies that used labor and capital in different proportions. He, furthermore, assumed that the capacity of the economy to use each technology was limited and given by a Pareto distribution over the different technologies. In this economy, he showed that the aggregate production function was Cobb-Douglas.

In other words, in Houthakker's model, even though the elasticity of substitution at each firm in the economy is zero, the economy's aggregate production function has an elasticity of substitution between capital and labor equal to one! All substitution in his economy occurs by production shifting between technologies. The assumptions Houthakker made to go from a micro elasticity of substitution of zero to a macro elasticity of substitution of exactly one are special. But, the logic of Houthakker's example has two more general implications: 1) the elasticity of substitution at the industry and aggregate level is higher than the elasticity of substitution at the level of an individual firm or plant; 2) this difference can be arbitrarily large, i.e., the elasticity of substitution between capital and labor can be arbitrarily large at the aggregate level no matter how low it is at each establishment. Aggregation really matters!

Houthakker's idea implies that we can't use our intuition about individual pro-

duction technologies to think about the elasticity of substitution between labor and capital at the level of the aggregate economy. More importantly, it means that we can't use formal empirical evidence on this elasticity of substitution at the firm level to make inference about the elasticity of substitution at the aggregate level. We need empirical evidence from aggregate data to estimate the aggregate σ.

Unfortunately, estimating σ at the aggregate level is notoriously difficult. The reason for this is an instance of the most classic problem in empirical economics: *the simultaneous equations problem*. The elasticity of substitution σ is the slope of a relative demand curve—equation (32). (You can flip the L_t/K_t on the left-hand-side and multiply through by minus one to see that this relative demand curve is downward sloping.) But the equilibrium relative price r_t/w_t and relative quantity K_t/L_t are determined not only by the relative demand curve but also by the relative supply curve. These two equations, form a system of two equation in two unknown variables. Solving this system gives the equilibrium values of r_t/w_t and K_t/L_t. Clearly, these variables are determined jointly (*simultaneously*) by the relative demand curve and the relative supply curve.

Suppose an empirical economist has gathered data on r_t/w_t and K_t/L_t either over time or across countries. They would like to use these data to estimate σ. A simple-minded approach would be to run a regression analogous to equation (32):

$$\ln\left(\frac{K_t}{L_t}\right) = -\sigma \ln\left(\frac{r_t}{w_t}\right) + \epsilon_t. \tag{35}$$

This approach will typically not work. The reason is that the simple relationship between r_t/w_t and K_t/L_t (the "best fitting" line through the $(K_t/L_t, r_t/w_t)$ points) is influenced both by movements in the relative demand curve and also be movements in the relative supply curve. It therefore does not yield the slope of the relative demand curve σ.

To understand this better, consider Figure 7. The left panel depicts an idealized situation where the variation in $(K_t/L_t, r_t/w_t)$ arises only from variation in the relative supply curve. Notice that in this panel the relative supply curve is shifting and this is giving rise to variation in $(K_t/L_t, r_t/w_t)$ that traces out the relative demand curve: Since the relative demand curve is stable (doesn't shift) all the points in this panel are on the same relative demand curve. One can then draw a line through these points and the slope of this line will be the slope of the relative demand curve, i.e., it will be σ.

The left-hand-side panel in Figure 7, therefore, shows that, if one can isolate

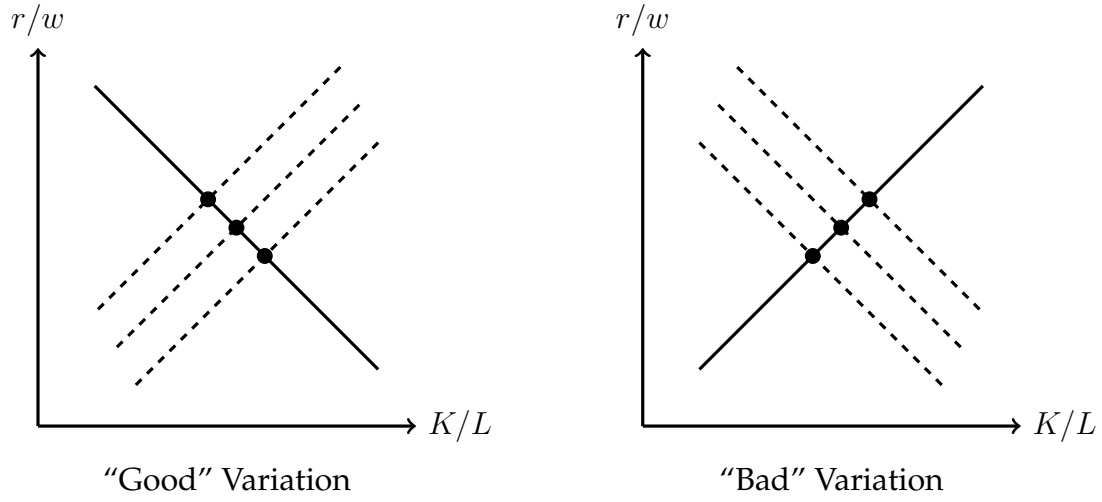

Figure 7: Estimating Elasticity of Substitution

variation $(K_t/L_t, r_t/w_t)$ that arises *only* from variation in the relative supply curve, one can use this variation to estimate σ. This is why we label the left-hand-side panel as "good" variation.

Contrast this with the right-hand-side panel of Figure 7. In this panel, the variation in $(K_t/L_t, r_t/w_t)$ arises from shifts in the relative demand curve. In this case, the variation in $(K_t/L_t, r_t/w_t)$ doesn't trace out the relative demand curve, rather it traces out the relative supply curve: Since the relative supply curve is stable (doesn't shift) all the points in this panel are on the same relative supply curve. One can therefore use these points to estimate the slope of the relative supply curve. But one cannot use these points to estimate the slope of the relative demand curve. For the purposes of estimating the relative demand curve, this is "bad" variation.

In most real world situations, both the relative demand curve and the relative supply curve will shift around. This will give rise to a cloud of points in $(K_t/L_t, r_t/w_t)$ space. Running regression (35) will then recover neither the slope of the relative demand curve nor the slope of the relative supply curve. This is a major challenge for empirical economics, arguably *the* central empirical challenge in the field.

The primary approach economists take to solving this challenge is to look for "natural experiments." In this context, a natural experiment is a situation where the researcher can argue that some slice of the variation $(K_t/L_t, r_t/w_t)$ arises only from variation in relative supply. One approach to this is to identify an "instrumental variable". An instrumental variable is a variable that proxies for pure variation in

relative supply.

Suppose one's goal was to estimate the slope of the demand curve for coffee in the United States. Rainfall and temperature in Brazil (a major coffee growing region) would then be plausible instrumental variables. Weather in Brazil affects the world supply of coffee beans, but probably has a negligible effect on the U.S. demand curve for coffee. Variation in the U.S. price and quantity of coffee associated with weather in Brazil is therefore good variation when the goal is to estimate the slope of the U.S. demand curve for coffee.

When it comes to estimating σ, the problem is that it has proven difficult to identify plausible instrumental variables. Most studies that attempt to estimate σ instead make relatively strong theoretical assumptions (for example that $\ln(A_{K,t}/A_{L,t})$ are constant across countries or are equal to a time trend over time). Different assumptions yield different estimates of σ. Some are larger than one, while others are smaller than one. Unfortunately, none of these estimates is particularly convincing. As a consequence, there is no consensus in the field when it comes to the value of the aggregate elasticity of substitution between capital and labor. More research is needed.

7 Man Versus Machine

Traditional formulations of the production function—such as the Cobb-Douglas and CES productions functions—are very useful modeling devices for many purposes. However, they also have important drawbacks. One such drawback is that they are a "black box", i.e, they lack descriptive realism regarding the actual process of production in the economy. Consider Adam Smith's famous description of a pin factory in Wealth of Nations:

> One man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on, is a peculiar business, to whiten the pins is another; it is even a trade by itself to put them into the paper. (Smith, 1776/2000, p. 4)

The Cobb-Douglas and CES production functions have no such descriptive realism. The only aspect of production that these functions capture is the fact that production involves combining labor and capital. The manner in which this occurs is left completely opaque.

This lack of descriptive realism would be forgivable if these production functions captured the essence of production. Arguably, however, they miss a considerable amount of the essence of how technological change affects the production process. Traditional production functions model technological progress as being *factor augmenting*. For example, the CES production function we discuss in section 6.3 allows for two types of technological progress: labor augmenting and capital augmenting. This turns out to be a very restrictive way to model technical progress.

7.1 Can Innovation Hurt Workers?

To see how restrictive it is to model productivity as being factor augmenting, let's return to the persistent concern that technological progress destroys jobs by replacing workers with machines. Can the Cobb-Douglas or CES production functions capture this notion? This is not immediately obvious (given their lack of descriptive realism). However, we can ask a very related question: Can an increase in the quantity (or productivity) of capital make workers worse off in the sense of lowering their wages? If machines destroy jobs and replace workers, they may reduce labor demand and thereby reduce the wages of workers. Can this occur in an economy with a Cobb-Douglas or CES production function?

Consider first the case of an economy with a Cobb-Douglas production function and a competitive labor market. In this case, worker wages are equal to the marginal product of labor

$$w = \frac{\partial F(K,L)}{\partial L} = (1-a)AK^aL^{-a}.$$

The effect of an increase in capital K or technology A on wages can then be calculated as

$$\frac{\partial w}{\partial K} = a(1-a)AK^{a-1}L^{-a} > 0 \quad \text{and} \quad \frac{\partial w}{\partial A} = (1-a)K^aL^{-a} > 0.$$

We see that both an increase in capital and technical progress unambiguously increase the marginal product of labor and therefore wages. Capital and technical progress complements workers when the production function is Cobb-Douglas. An economy with this production function, thus, cannot capture the notion that technology or machines may hurt workers by destroying jobs.

How about an economy with a CES production function and a competitive labor

market. In this case the wage is

$$w = (1-a)A_L^{\frac{\sigma-1}{\sigma}}\left(\frac{Y}{L}\right)^{\frac{1}{\sigma}} = (1-a)A_L^{\frac{\sigma-1}{\sigma}}\left[aA_K^{\frac{\sigma-1}{\sigma}}\left(\frac{K}{L}\right)^{\frac{\sigma-1}{\sigma}} + (1-a)A_L^{\frac{\sigma-1}{\sigma}}\right]^{\frac{1}{\sigma-1}}$$

The effect of an increase in capital K or capital-augmenting technology A_K on wages can then be calculated as

$$\frac{\partial w}{\partial K} = \frac{a(1-a)}{\sigma}A_L^{\frac{\sigma-1}{\sigma}}A_K^{\frac{\sigma-1}{\sigma}}Y^{-1}L^{\frac{-1}{\sigma}}K^{\frac{-1}{\sigma}} > 0,$$

$$\frac{\partial w}{\partial A_K} = \frac{a(1-a)}{\sigma}A_L^{\frac{\sigma-1}{\sigma}}A_K^{\frac{-1}{\sigma}}Y^{-1}L^{\frac{-1}{\sigma}}K^{\frac{\sigma-1}{\sigma}} > 0.$$

Again, an increase in capital and (capital-augmenting) technical progress unambiguously increase the marginal product of labor and therefore wages. This is true for any value of the elasticity of substitution between capital and labor.

7.2 A Task-Based Production Function

To capture the notion that technology (in the form of machines) may replace workers, destroy jobs, and thereby make workers worse off, we must consider a more descriptive model of production. Acemoglu and Restrepo (2018) present such a model. In their model, the production of goods involves completing a range of tasks (just as in Smith's pin factory example). For concreteness, suppose the production of a particular good involves J tasks. We give each of these tasks a label j from 1 to J. We denote by $y(j)$ the number of times the jth task is performed and we denote by Y the amount of the good produced.

Suppose that the tasks are somewhat substitutable and output of the good is given by the production function

$$Y = \left[\sum_{j=1}^{J} y(j)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}. \tag{36}$$

This production function captures the intuitive notion that the good is produced by performing the tasks. The more tasks are performed, the more goods are produced. This type of production function is called a *task-based production function*.

In equation (36), the parameter σ is the elasticity of substitution of the different tasks. It plays a similar role to σ in equation (26). The difference is that in equation (36) there are J tasks that are combined to produce the good, while in equation

(26) there are two factors of production (labor and capital) that are combined in the production process.

A useful limiting case to consider is $\sigma = 0$. In this case, the production function becomes $Y = \min(y(j))$. Each task must be performed once to produce an extra unit of the output good. Smith's description of the pin factory has this characteristic. But more generally, we suppose that there is more than one way to produce the good and these different methods involve some substitution of different tasks.

How are the tasks performed? We assume that they can be performed either by workers or by machines. Specifically, we assume that

$$y(j) = \psi_L(j)L(j) + \psi_K(j)K(j). \tag{37}$$

Here $L(j)$ denotes the amount of labor devoted to task j, $K(j)$ denotes the amount of capital devoted to task j, and $\psi_L(j)$ and $\psi_K(j)$ (ψ is the Greek letter psi) denote the productivity of labor and capital, respectively, at performing task j.

Equation (37) is a very particular choice of production function for task j in that it assumes that labor and capital are perfectly substitutable in performing task j. In other words, this is the limit of a CES production function over labor and capital for task j when the elasticity of substitution between labor and capital goes to infinity. This assumption has the (stark) implication that task j is either produced by labor or by capital, not by a mix of the two. (Except in the case when it is equally costly to use labor and capital.)

Notice that the marginal product of labor in task j is constant at $\psi_L(j)$ independent of the amount of labor devoted to task j. Likewise, the marginal product of capital in task j is also constant at $\psi_K(j)$ independent of the amount of capital devoted to task j. Suppose that the wage in the economy is w and the rental rate of capital is r. The cost of producing one unit of task j with labor is then $w/\psi_L(j)$, while the cost of producing a unit of task j with capital is $r/\psi_K(j)$.

Given these costs, the firm will choose to produce task j with labor if

$$\frac{w}{\psi_L(j)} < \frac{r}{\psi_K(j)}. \tag{38}$$

Otherwise, the firm will produce task j with capital. (The firm is indifferent if equation (38) holds with equality.) The perfect substitutability assumption in equation (37) is what yields this "either or" implication for labor versus capital in task j. It simplifies the analysis greatly.

When analyzing this model below, we assume, for simplicity, that the supply of capital is exogenously given at $\bar{K}$ and also that labor supply is exogenously given at $\bar{L}$. This is the same assumption we made earlier in the chapter.

7.3 Innovation on Tasks

In the task-based framework, technical progress increases $\psi_L(j)$ and $\psi_K(j)$. Some innovations will increase $\psi_L(j)$, others will increase $\psi_K(j)$, and some will increase both. For concreteness, consider innovations that increase $\psi_K(j)$. Importantly, we can consider innovations that make machines better at some tasks but not others, i.e., increase $\psi_K(j)$ for some j but not other j. Arguably, most innovations take this form: someone invents a (better) machine that performs a particular task (or perhaps some collection of tasks).

Several hundred years ago, few machines had been invented. Most tasks could therefore only be performed by labor. In this case $\psi_K(j) = 0$ for many tasks. Typically, the first machines invented to perform a task were extremely inefficient (e.g., the first steam engine). The $\psi_K(j)$ of these machines was positive, but very small. This meant that these machines were not cost competitive for many tasks. (The first steam engine was used to pump water out of coal mines where the fuel to power it was effectively free.) Much innovation then involved improving these machines (raising their $\psi_K(j)$). As their productivity rose, they became cost competitive for more and more tasks. This meant that the machines replaced labor in more and more tasks.

To understand this process better, consider, as an example, a good that is produced with seven tasks. The cost of performing these tasks with labor ($w/\psi_L(j)$) and the cost of performing them with capital ($r/\psi_K(j)$) are plotted in Figure 8. The white circles denote the cost of performing the various tasks with labor, while the black stars denote the cost of performing the tasks with machines.

I have ordered the tasks such that the ratio of these two costs is falling from left to right. Task 1 is the task for which the cost of performing that task with labor is highest relative to the cost of performing that task with a machine. This is therefore the task for which machines have the strongest comparative advantage. Task 7, on the other hand, is the task for which the cost of performing that task with a machine is highest relative to the cost of performing it with labor. This is the task for which labor has the strongest comparative advantage.

Given the costs depicted in Figure 8, tasks 1, 2 and 3 will be performed by ma-

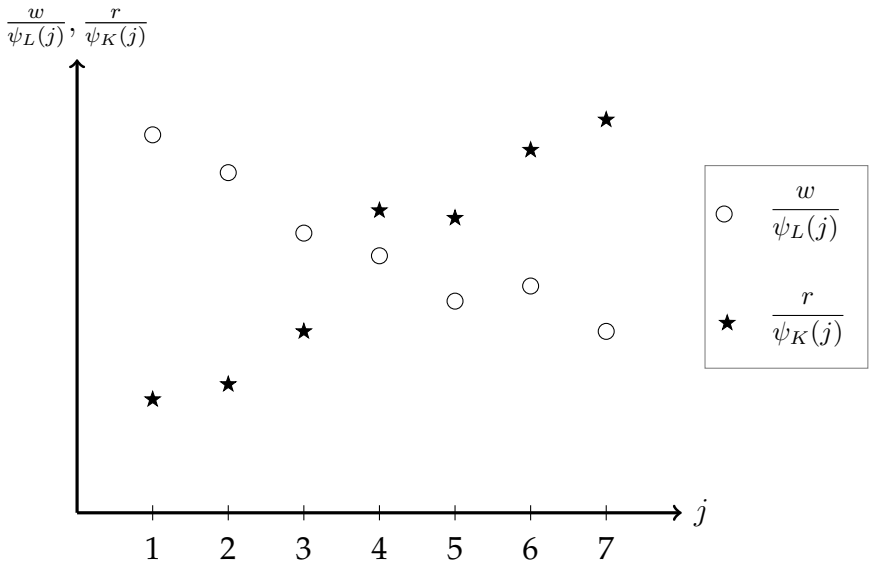

Figure 8: Cost of Producing Tasks with Labor and Capital

Note: The figure considers the case of a good that is produced from seven tasks. For each of these tasks, the cost of performing that task with labor ($w/\psi_L(j)$) is denoted by a white circle, while the cost of performing that task with a machine ($r/\psi_K(j)$) is denoted by a black star.

chines, while tasks 4, 5, 6, and 7 will be performed by labor. These are the lowest cost methods for each task.

Let's now consider innovations that improve the productivity of machines in performing these tasks. Figure 9 considers three such innovations. The first of these—labeled A in the figure—improves the efficiency of machines at performing task 2, i.e., it raises $\psi_K(2)$. This increase in $\psi_K(2)$ lowers the cost of producing task 2 with machines—$r/\psi_K(2)$—from the black star at $j = 2$ to the gray star at $j = 2$.

Notice that even before this innovation, machines were able to perform task 2 more cheaply than labor. The task had, therefore, already been automated. Innovation A simply makes machines even better at performing this task that they already perform. Since no worker performs task 2 before or after the innovation, workers are not *directly* affected by the innovation.

However, workers are affected indirectly. Innovation A makes it less costly to perform task 2. This implies that the overall cost of producing the good (that is made from the seven tasks) also falls. Suppose for simplicity, that the producers of the good sell it in a competitive output market. Competition among producers of this good will then drive down the price of the good in line with the fall in its costs. The fall in the price of the good raises the purchasing power of the wages workers

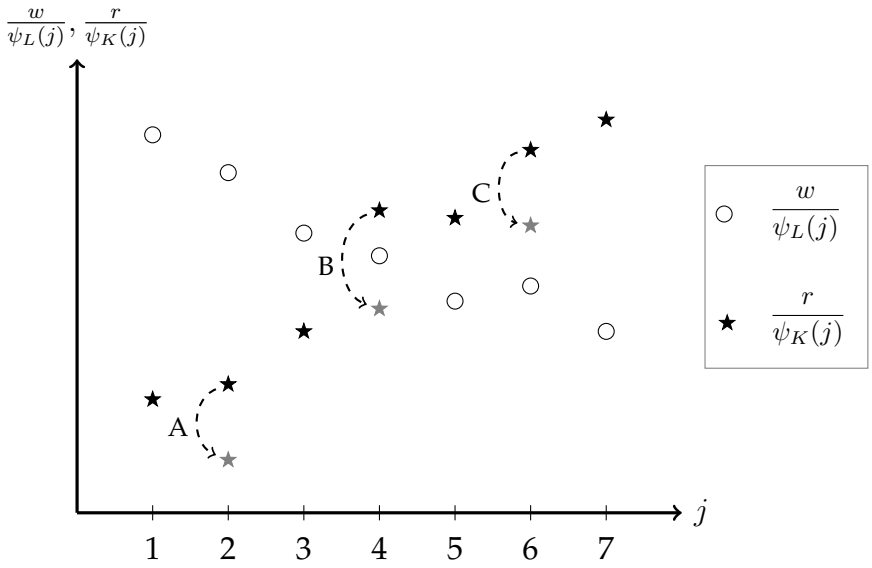

Figure 9: Technological Change that Improves Machines

Note: The three gray stars represent the cost of performing tasks 2, 4, and 6 with machines after innovations have improved the productivity of performing these tasks with machines.

earn in the economy: their "real" wage. The innovation, thus, benefits workers through raising their real wage. This effect is called the *productivity effect*.

Consider next an innovation that modestly improves the productivity of machines in performing task 6. This innovation is labeled C in Figure 9. It raises $\psi_K(6)$ and therefore lowers the cost of producing task 6 with machines—$r/\psi_K(6)$. This fall is the shift from the black star at $j = 6$ to the gray star at $j = 6$.

In this case, the cost of performing task 6 with labor is lower than with machines both before and after the innovation. The innovation, therefore, does not affect the lowest cost method for performing task 6. As a consequence, the innovation has no effect on how task 6 is performed. It was performed with labor before the innovation; it is performed with labor after the innovation; and the productivity of performing the task with labor has not changed. In other words, this innovation has no effect on the economy at all.

Finally, consider the innovation labeled B in Figure 9 that improves the productivity of machines in performing task 4. Before this innovation, the lowest cost method for performing task 4 was hiring labor: the white circle at $j = 4$ is below the black star at $j = 4$. After innovation B takes place, however, machines have become the lowest cost method for performing task 4. The gray star at $j = 4$ depicts the cost of performing task 4 with machines after the innovation. The fact that the gray start

at $j = 4$ is below the white circle at $j = 4$ indicates that the innovation has made machines the lowest cost method for performing task 4.

How does this innovation affect labor? This is a more complicated case to analyze than the prior two cases. To understand this case, it is instructive to imagine that the innovation gradually raises $\psi_K(4)$ and thus gradually reduces $r/\psi_K(4)$. It is furthermore instructive to break this gradual reduction in $r/\psi_K(4)$ into three parts: 1) the part at the beginning when $r/\psi_K(4) > w/\psi_L(4)$, 2) the point at which $r/\psi_K(4) = w/\psi_L(4)$, and 3) the part at the end when $r/\psi_K(4) < w/\psi_L(4)$.

While $r/\psi_K(4) > w/\psi_L(4)$, the innovation has no effect on the economy for the same reason as innovation C that we discussed above. In this range, labor is still the lowest cost method for performing task 4, and exactly how much more it costs to perform the task with machines is not material. When the sign of the relative cost has flipped—$r/\psi_K(4) < w/\psi_L(4)$—and machines have become the lowest cost method for performing the task, any further reductions in the cost of performing the task with machines make labor better off for the same reason as with the innovation to task 2 we considered above. In this range, further reductions in the cost of performing the task with machines lower the overall cost of producing the good and this raises the real wages of workers. Thus, when $r/\psi_K(4) < w/\psi_L(4)$, further improvements in $\psi_K(4)$ raise worker real wages through a productivity effect.

This leaves the point at which $r/\psi_K(4) = w/\psi_L(4)$. As $r/\psi_K(4)$ passes this point, the lowest cost method for performing task 4 switches from being labor to being capital. When this happens, the firms producing the good will fire the workers that were performing task 4 and switch to renting (or buying) machines to perform this task. In other words, at this point the demand for labor in the economy falls and the demand for capital in the economy increases. This effect is called the *displacement effect*.

In our simple model with competitive factor markets and competitive product markets, the effect of the shift in factor demand described above is to reduce wages in the economy and increase the rental rate of capital. This is depicted in Figure 10. When task 4 gets automated, the labor demand curve in the economy shifts down, while the capital demand curve shifts up. Since we have assumed that labor supply and capital supply are given, the labor and capital supply curves are vertical. The shift in labor and capital demand therefore translates into a fall in the wage and an increase in the rental rate on capital. In the figure, the economy moves from point A to point B.

In the real world, the effects of the automation of a task are likely more complex.

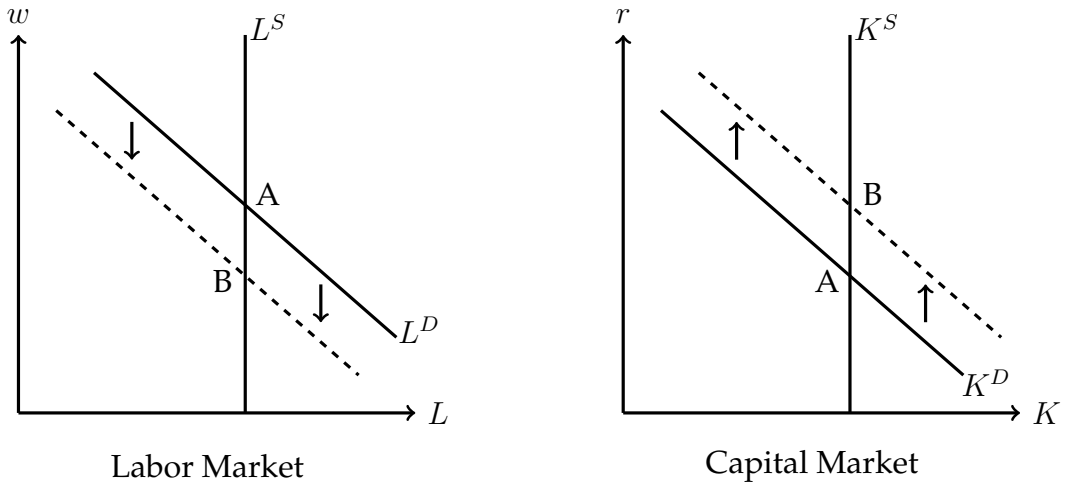

Figure 10: Effect of Automation on Factor Demand

The workers that lose their jobs may find it difficult to find new work. Automation may thus increase unemployment for some time. But as these workers seek alternative work, they will compete with the other workers in the economy for the remaining jobs. This will put downward pressure on wages as in Figure 10. Also, the increase in the rental rate on capital will encourage more saving in the economy and the accumulation of more capital. Finally, the reduction in wages and increase in the rental rate on capital will shift research away from ideas that seek to economize labor towards other types of research.

If the workers that lost their jobs due to innovation B had accumulated specific skills related to these jobs or were earning rents for some reason, they will permanently lose these skill premia or rents. This will mean that innovation B is worse for them than for other workers. Innovations such as innovation B may therefore hurt some workers while making other workers better off. This will happen if the productivity effect of the innovation is large enough to overcome the displacement effect on overall wages in the economy. In this case, the other workers in the economy will see their real wages rise due to the innovation. But the workers that lost their jobs will be worse off due to facing an unemployment spell, and also due to losing their job-specific skill premia and the rents they were earning.

The three examples of innovation depicted in Figure 9 illustrate that the task-based model of production is able to capture a much richer range of effects of innovation on workers than the Cobb-Douglas and CES models we considered earlier in the chapter. Importantly, the task-based model captures both the positive pro-

ductivity effect that such innovations can have and also the negative displacement effect. The earlier models we analyzed were missing the displacement effect and therefore implied that improvements in machines necessarily made workers better off. The task-based model shows how this may not necessarily be the case.

Whether technological progress that improves machines makes workers better or worse off depends on the relative strength of the productivity effects these innovations have and the displacement effects that they have. Intuitively, for any given innovation, this depends on the size of the productivity effect relative the amount of displacement of labor in induces. Many innovations only have productivity effects (such as innovation A in Figure 9). These unambiguously make workers better off. But others have a mix of effects. Worst are innovations that do not improve productivity much but happen to shift $r/\psi_K(j)$ from above to below $w/\psi_L(j)$ for tasks that employ many of workers. These innovations cause large displacement effects but small productivity effects and will harm workers.

But technical progress that makes machines more efficient and displaces workers in more and more of the traditional tasks in the economy also causes the economy to grow. This growth will naturally result in additional division of labor, i.e., the subdivision of tasks performed by labor, that previously were bundled together, into several distinct tasks. In addition, the invention of more and more machines will in-and-of itself create tasks related to the design, construction, and maintenance of the machines. Finally, as people's income rises, they will be able to afford goods and services that they were not able to afford before, and the production of these goods and services will give rise to a multitude of new tasks. All of this will increase labor demand and therefore benefit workers. Acemoglu and Restrepo (2018) refer to all of these effects as the *reinstatement effect* of technical progress.

7.4 The Evolution of Work Since the Industrial Revolution

If one thinks about the broad sweep of economic history since the onset of the Industrial Revolution, it is clear that almost every task that labor performed in the 18th century has been automated by now. In the 18th century, most workers worked in agriculture. As the economy industrialized, a larger and larger fraction of workers moved into other sectors. Figure 11 plots the share of workers in agriculture (as well as manufacturing and professional services) from 1850 to 2018. The share in agriculture fell steadily from about 60% in 1850 to about 3% in 1970, and was barely above 2% in 2018. Most tasks farm workers performed in 1750 are now performed by ma-

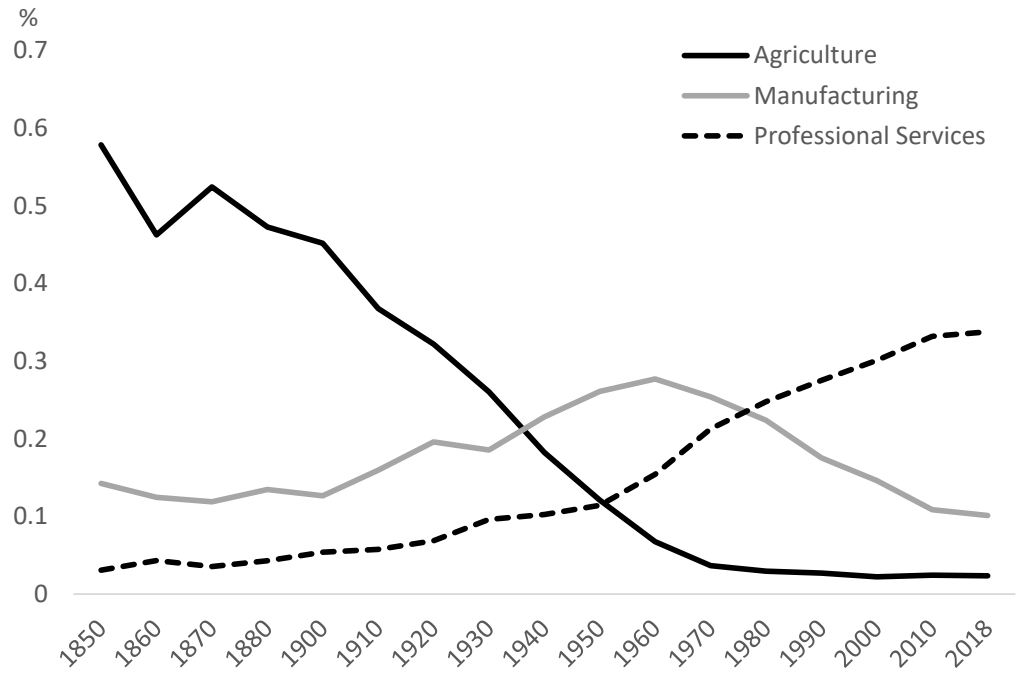

Figure 11: Evolution of Select Industry Shares in the United States

Note: The figure plots the fraction of workers employed in each of three industries over time. I have included Finance, Insurance, and Real Estate in Professional Services. The sources of these data are the U.S. Census and American Community Survey.

chines (tractors, harvesters, milking machines, etc.) and most of the tasks today's farmers perform did not exist in any appreciable form in 1750.

In the 19th and 20th centuries, many workers moved from agriculture to manufacturing. The share of workers in manufacturing rose from about 12% in the late 19th century to about 28% in 1960. In manufacturing, these workers were performing tasks that mostly did not exist 100 years prior. But again, most of these tasks have by now been mechanized and are performed by machines. The share of the population employed in manufacturing has been falling since 1960. In 2018, it was down to about 10%.

Over the past 75 years, workers have increasingly been moving into various forms of services. The employment share in retail trade has, for example, more than doubled since the late 19th century. Much more dramatically, the employment share of professional services (lawyers, doctors, engineers, software developers, consultants, etc.) has risen by a factor of about ten from about 3% to almost 30% over the past 150 years. Some of the tasks performed by these workers did exist in some form back in 1850, while others are entirely new. Regardless, as society has grown

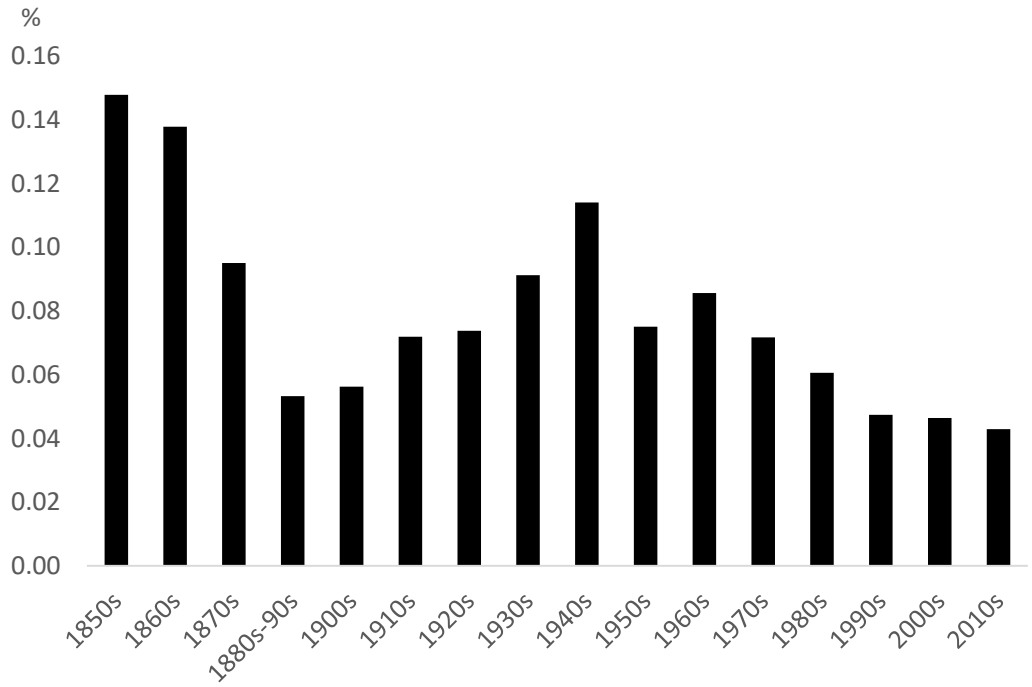

Figure 12: Job Shifts Across Broad Occupations in the United States

Note: The figure plots shifts in jobs between broad occupation groupings in the United States at a decadal frequency. Jobs are divided into 11 categories: Farmers; Farm Laborers; Laborers; Craftsmen; Clerical Workers; Operatives; Sales Workers; Household Services; Non-Household Services; Professional and Technical Workers; Managers, Officials, and Proprietors. The figure shows the fraction of jobs that shift between these categories each decade. The sources of these data are the U.S. Census and American Community Survey. The data for the 2010s are from 2010 to 2018.

richer, the demand for these tasks has skyrocketed. Now artificial intelligence is threatening to automate a good number of these tasks.

At any given point in time, it can be difficult to imagine what new tasks will be able to absorb all the workers that are losing their current jobs. In the 2020s (when this is written), anxiety about the future of work centers on to artificial intelligence and robotics. These new technologies are likely to transform large parts of the economy and destroy a great many jobs in the process. Anxiety about this is understandable. It is hard to predict what new jobs will take the place of the old ones that are destroyed. Many of these tasks and jobs do not even exist today.

When thinking about the future of work, it is useful to have some historical perspective. The current moment feels extremely disruptive in terms of technological change. But such change has been going on for over 200 years. Actually, the current moment is arguably not as disruptive as some earlier periods. Figure 12 attempts to measure the degree of disruption in the labor market as the share of jobs that shift

broad occupations from decade to decade. On this metric, the last few decades were less disruptive than the mid-20th century. The last few decades saw an IT revolution, greatly expanding globalization and offshoring of jobs, and a massive increase in the use of robots. But the mid-20th century saw the mechanization of agriculture, the assembly line, household appliances, the forklift, and shipping containers to name just a few technologies. Despite all this change, we have avoided—at least up until now—persistent predictions of mass technological unemployment.

References

ACEMOGLU, D. AND P. RESTREPO (2018): "The Race between Man and Machine: Implications of Technology for Growth, Factor Share, and Employment," *American Economic Review*, 108, 1488–1542.

BARRO, R. (2021): "Double Counting of Investment," *Economic Journal*, 131, 2333–2356.

HOUTHAKKER, H. S. (1955): "The Pareto Distribution and the Cobb-Douglas Production Function in Activity Analysis," *Review of Economic Studies*, 23, 27–31.

KAHNEMAN, D. (2011): *Thinking Fast and Slow*, New York, NY: Farrar, Straus and Giroux.

KALDOR, N. (1961): "Capital Accumulation and Economic Growth," in *The Theory of Capital*, ed. by D. C. Hague, London, UK: Palgrave Macmillan, 177–222.

KARABARBOUNIS, L. AND B. NEIMAN (2014): "The Global Decline of the Labor Share," *Quarterly Journal of Economics*, 129, 61–103.

KEYNES, J. M. (1939): "Relative Movements of Real Wages and Output," *Economic Journal*, 49, 34–51.

KOH, D., R. SANTAEULÀLIA-LLOPIS, AND Y. ZHENG (2020): "Labor Share Decline and Intellectual Property Products Capital," *Econometrica*, 88, 2609–2628.

LEONTIEF, W. (1983): "Technological Advance, Economic Growth, and the Distribution of Income," *Population and Development Review*, 9, 403–410.

SMITH, A. (1776/2000): *An Inquiry into the Nature and Causes of the Wealth of Nations*, New York, NY: The Modern Library.