Journal of Monetary Economics 60 (2013) 459–473

Contents lists available at SciVerse ScienceDirect

Journal of Monetary Economics

journal homepage: www.elsevier.com/locate/jme

# Factor-eliminating technical change ☆

Pietro F. Peretto [a], John J. Seater [b,*]

[a] Duke University, United States
[b] NC State University, United States

ARTICLE INFO

*Article history:*
Received 18 August 2010
Received in revised form
2 January 2013
Accepted 8 January 2013
Available online 6 April 2013

ABSTRACT

Perpetual growth requires offsetting diminishing returns to reproducible factors of production. In this article we present a theory of factor elimination. For simplicity and clarity, there is no augmentation of non-reproducible factors, thus excluding the standard engine of growth. By spending resources on R&D, agents learn to change the *exponents* of a Cobb–Douglas production function. We obtain the economy's balanced growth path and complete transition dynamics. The theory provides a mechanism for the transition from an initial technology incapable of supporting perpetual growth to one with constant returns to reproducible factors that supports it.



## 1. Introduction

Perpetual economic growth requires that the marginal products of reproducible factors of production (e.g., physical capital) be bounded away from zero. Growth models virtually always satisfy this condition by augmenting the non-reproducible factors (e.g., unskilled labor). There is, however, another way to satisfy the necessary condition: learn to produce without some of the non-reproducible factors. We propose a theory of factor elimination, examine its implications for economic growth, and provide evidence of its relevance.

Non-reproducible factors act as a drag on reproducible factors; as the ratio of reproducible to non-reproducible factors rises, the marginal products of the reproducible factors fall until accumulation stops. Augmenting the non-reproducible factors eliminates this drag by effectively increasing their quantity, thereby raising the marginal products of the reproducible factors and permitting their accumulation to continue. Factor-eliminating technical progress delivers growth through a different mechanism: it relaxes the constraint on growth by reducing the importance of non-reproducible factors in production.

There are two factors of production, one reproducible and one not. For simplicity and clarity, we use a Cobb–Douglas production function with no factor-augmenting technical progress of any kind, thus excluding the standard engine of growth. What is new is the possibility of changing *factor elasticities of output* (that is, the exponents of the Cobb–Douglas) by devoting resources to R&D. The general equilibrium dynamics have two possible outcomes. For high saving rates, production asymptotically becomes *AK*, supporting perpetual growth. For low saving rates, the economy reaches a steady state with no growth and a standard production function with constant factor elasticities between 0 and 1.

☆ We thank the following for comments that improved the paper: Otilia Boldea, Diego Comin, Peter Howitt, Oksana Leukhina, Brad Sturgill, Hernando Zuleta, participants in the University of North Carolina Macroeconomics Workshop and the DEGIT XI Conference in Jerusalem, this Journal's Senior Associate Editor Sergio Rebelo, and an anonymous referee.

* Corresponding author. Tel.: +1 919 513 2697; fax: +1 919 515 7873.
*E-mail address:* john_seater@ncsu.edu (J.J. Seater).

0304-3932/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jmoneco.2013.01.005

The theory resolves the "linearity critique" (Jones, 2005), recently formalized by Growiec (2007), who proves that perpetual growth requires linearity in some part of the economy's dynamical system. The critique implies fragility of growth models because a reason for the linearity is not obvious. With factor-eliminating technical change, however, the required linearity emerges *endogenously* in an economy that initially does not have it.

The theory also offers a fresh insight on how an economy leaves a primitive starting point and develops. Modern theories typically posit that society is endowed from the start with two production functions, one primitive and one advanced, and over time reallocates resources from the former to the latter (e.g., Goodfriend and McDermott, 1995; Hansen and Prescott, 2002; Galor and Weil, 2000). In our theory, by contrast, there is only one initial technology and new technologies appear if and only if people use resources to invent them.

Finally, the theory resolves the controversy about the cross-country behavior of factor shares. Some evidence (Gollin, 2002; Bernanke and Gürkaynak, 2001) suggests that capital and labor shares are unrelated to output per worker, other evidence (Caselli and Feyrer, 2007; Zuleta, 2008a) suggests a systematic relationship. Our theory shows that when the proper distinction between reproducible and non-reproducible factors is maintained, the conflict disappears: output per worker is positively related to the share of reproducible capital (physical and human), negatively related to the share of natural capital, and has no relation with the shares of *total* capital or labor.

## 2. Background, intuition, and related literature

Consider a generic production function $Y=F(K,L)$, where $Y$, $K$ and $L$ are output, reproducible inputs (physical capital and human capital, which hereafter we call "capital" for simplicity) and non-reproducible inputs (land, natural resources, and unskilled labor, which hereafter we call " labor"), respectively, and $F$ satisfies the neoclassical assumptions. For simplicity $L$ is constant. As $K$ grows, the marginal product of capital shrinks until it just equals capital's marginal cost, bringing growth to a halt. What guarantees this outcome is the Inada condition $\lim_{K\to+\infty}F_K(K,L)=0$. To generate perpetual growth the Solow and Cass models introduce augmentation of labor. The production function has the form $F(K,AL)$, where $A$ is labor-augmenting knowledge that grows at the exogenous rate. Perpetual growth is feasible because the endowment of effective labor $AL$ grows over time and drives up the marginal product of capital, sustaining incentives for accumulation. Specifically, the linear homogeneity of $F$ allows us to write $\lim_{K\to+\infty}F_K(K/A,L)$ and the ratio $K/A$ remains finite because in steady state $K$ and $A$ grow at the same rate. This theory is better called a theory *with* growth than a theory *of* growth because growth arises from strictly exogenous forces that the theory makes no attempt to explain. The great advance of endogenous growth theory is precisely to endogenize technical progress. For example, in Romer's (1986) path-breaking model of learning-by-doing, $A$ is equal to $K$, and the technology becomes $F(K,KL)$.

Elimination of the non-reproducible factors achieves the same end by a different mechanism. The key insight is a straightforward implication of the reason why non-reproducible factors tend to bring growth to a halt. To see it, write

$$\lim_{K\to+\infty}F_K(K,L)=\lim_{K\to+\infty}\frac{F(K,L)}{K}=\lim_{K\to+\infty}F\left(1,\frac{L}{K}\right). \tag{1}$$

This first equality follows from L'Hopital's rule. It shows that the Inada condition is equivalent to the condition that the average product of capital goes to zero. The second equality implies that the average product of capital goes to zero because labor is essential, that is, because $F(1,0)=0$. It follows that theories of perpetual growth are theories of how economic agents overcome scarcity of *essential, non-reproducible* factors of production: augmentation does it by transforming non-reproducible factors into reproducible ones, elimination does it by dispensing with the non-reproducible factors.

Despite early promising theoretical and empirical contributions (Kamien and Schwartz, 1968; Sato and Beckmann, 1968), research on factor-eliminating progress lay dormant for nearly 40 years until Seater (2005) reopened the topic. He studies a centrally planned economy with Cobb–Douglas production and exogenous technical progress that alters the factor exponents. Recently, Zuleta (2008b) has extended Seater's analysis to characterize more fully the growth path. The major limitation of both Seater's and Zuleta's analyses is that they are restricted to the central planning solution.[1]

In this paper, we study factor-eliminating technical progress in the context of market equilibrium rather than central planning. Firms invest resources to learn new technologies that increase production by reducing dependency on non-reproducible factors. We obtain a complete characterization of the economy's dynamics and discuss several implications of the theory, some of which we support with readily-available cross-country data.

## 3. A theory of factor elimination

There are three groups of agents: a representative household, competitive producers of final goods, a fixed mass of monopolistic producers of intermediate goods.

[1] Strictly speaking, they both assume competitive markets with no externalities, which allows them to treat the central planning solution as equivalent to the market solution. Such a framework, however, does not permit technical progress that derives from private research and development.

### 3.1. Households

Households own a fixed endowment of a non-reproducible factor. We call this factor "labor" but it includes all non-reproducible factors: unskilled labor, land, natural resources. Households also own the firms and receive all profits as dividends. Households supply labor services inelastically in a competitive market and save a fixed fraction of their income. We thus ignore both consumption-saving and labor-leisure choice. The more restrictive assumption is the constant saving rate. Allowing for endogenous saving, while worthwhile, is a non-trivial exercise because it adds a state variable to the system. See Zuleta (2008b), whose model includes an endogenous saving rate. As in the passage from Solow to Cass, an endogenous saving rate would introduce interesting generalizations but no fundamental change in the properties that support steady-state growth.

### 3.2. Final good producers

There is one final good, $Y$, produced by competitive firms according to the technology

$$Y=\left[\int_0^1 X_i^{(\epsilon-1)/\epsilon}di\right]^{\epsilon/(\epsilon-1)},\quad \epsilon>1, \tag{2}$$

where $X_i$ is the quantity of intermediate good $i$ and $\epsilon$ is the elasticity of substitution between intermediate goods. The final good is the numeraire, so $P_Y\equiv 1$. Final producers maximize profit $\pi_Y=Y-\int_0^1 P_iX_idi$ subject to (2), where $P_i$ is the price of good $i$. This problem yields the well known demand function

$$X_i=Y\frac{P_i^{-\epsilon}}{\int_0^1 P_i^{1-\epsilon}\,di}. \tag{3}$$

### 3.3. Intermediate goods producers

The intermediate monopolistic firms do three things: produce, invest in capital accumulation, and invest in the development of new production technologies. To fix terminology, we shall refer to the second activity as "investment" and to the third as "R&D" . We begin with a discussion of the technologies available to the typical firm. To keep the notation simple, we suppress time arguments whenever confusion does not arise.

The firm hires labor $L_i$ and combines it with its own capital $K_i$ according to

$$X_i=AK_i^{a_i}L_i^{1-a_i},\quad a_i\in[0,\alpha_i],\quad 0\le\alpha_i\le 1. \tag{4}$$

The firm chooses the capital elasticity $a_i$ from a set of known technologies.[2] The upper boundary $\alpha_i$ is the firm's technology frontier and grows as a result of the firm's R&D.

Total factor productivity (TFP) $A$ is constant and common to all firms.[3] Cobb–Douglas production imposes in the simplest possible way *essentiality* of both capital and labor for $a_i\in(0,1)$. However, $a_i=0$ yields $X_i=AL_i$, so that capital is non-essential, whereas $a_i=1$ yields $X_i=AK_i$, so that labor is non-essential and perpetual growth is feasible. As with labor, "capital" includes all types of reproducible factors of production (physical and human).

The firm's capital stock evolves according to

$$\dot K_i=I_i-\delta K_i, \tag{5}$$

where $I_i$ is gross capital investment in units of the final good and $\delta$ is the depreciation rate.

The firm's highest known capital elasticity $\alpha_i$ evolves according to

$$\dot\alpha_i=\begin{cases} f(\alpha_i)\cdot R_i, & \alpha_i<1\\ 0, & \alpha_i=1\end{cases} \tag{6}$$

where $R_i$ is R&D investment in units of the final good. The productivity of R&D is a function $f(\alpha_i)$. For the time being we do not need to specify its properties. We discuss them below, where they play a key role in the economy's equilibrium dynamics.

Our key assumption is that the firm's discoveries through R&D are fully excludable. It would be more realistic to assume that eventually a firm's knowledge leaks out, but the analysis would be more complicated with no substantial gain of insight. As long as there is some degree of excludability for some period of time, our results hold.

[2] Since firms earn excess profit and thus payments to $K_i$ and $L_i$ do not exhaust their revenues, $a_i$ is not the "capital share" and thus throughout the paper we refer to it as "elasticity of output with respect to capital" or "capital elasticity" for short.

[3] By assuming constant $A$ we do not mean to imply that factor-augmenting technical progress is unimportant. A complete model should include both, but it is useful here to restrict attention to what is new. We plan to develop the complete model in future research.

### 3.4. The typical intermediate firm's decisions

The firm solves

$$\max_{\{a_{it},P_{it},L_{it},I_{it},R_{it}\}_{t=0}^{\infty}} \int_0^{\infty} (P_{it}X_{it} - w_t L_{it} - I_{it} - R_{it}) e^{-\bar{r}_t t}\, dt, \tag{7}$$

where $\bar{r}_t \equiv \frac{1}{t}\int_0^t r_u\, du$ is the average interest rate between time 0 and time $t$, $r_u$ is the instantaneous interest rate at time $u$, and the optimization is subject to (3)–(6) and the restrictions $I_{it} \geq 0$ and $R_{it} \geq 0$. The individual intermediate firm perceives no upper bound on its choices of $I$ or $R$. In the aggregate, of course, firms' choices must satisfy the constraint $\int_0^1 (I_{it} + R_{it})\, di = sY_t$ at every time $t$.

It is convenient to think of the firm as operating two divisions, production and investment, and write the firm's objective function to reflect that internal structure:

$$\max_{\{a_{it},P_{it},L_{it},I_{it},R_{it}\}_{t=0}^{\infty}} \int_0^{\infty} [(P_{it}X_{it} - w_t L_{it} - p_{Ki}K_{it}) + (p_{Ki}K_{it} - I_{it} - R_{it})] e^{-\bar{r}_t t}\, dt. \tag{8}$$

The term inside the first set of parentheses is the instantaneous profit of the production division; the term inside the second set is the instantaneous profit of the investment division. The production division rents capital from the investment division, paying an internal transfer price $p_{Ki}$. The investment division receives that rental income and spends resources on investment and R&D. We then exploit time-separability and solve the firm's maximization problem in two steps. First, the production division chooses the optimal values of $P_i$, $L_i$, and $K_i$ taking $w$, $p_{Ki}$ and $\alpha_i$ as given. Then the investment division chooses $I_i$ and $R_i$.

There is no explicit payment for technology, whose return is included in the rental income accruing to the investment division, $p_{Ki}K_i$. To see this, note that

$$d(p_{Ki}K_i) = \frac{\partial(p_{Ki}K_i)}{\partial K_i} dK_i + \frac{\partial(p_{Ki}K_i)}{\partial \alpha_i} d\alpha = p_{Ki} dK_i + K_i \cdot \frac{\partial p_{Ki}}{\partial \alpha_i} d\alpha. \tag{9}$$

The first term in the last expression is the explicit payment to capital. We assume that the investment division does not exploit monopoly power in supplying capital to the production division because doing so would create inefficient distortions within the firm. Consequently, $\partial p_{Ki}/\partial K_i = 0$. The second term in the last expression is the implicit payment to technology. The intuition is that a more capital intensive technology allows the production division to use its capital and labor more efficiently and thus produce more, inducing the production division to pay more for capital. The increased payment for capital compensates the investment division for the R&D that enabled the increase in capital efficiency.

Henceforth, we omit the $i$ subscript except where clarity requires it. The production division solves

$$\max_{\{a_t,P_t,L_t,K_t\}_{t=0}^{\infty}} \int_0^{\infty} (P_t X_t - w_t L_t - p_{Kt}K_t) e^{-\bar{r}_t t}\, dt \quad \text{s.t.} \quad (3), (4). \tag{10}$$

Splitting the firm into production and investment divisions separates the firm's intratemporal and intertemporal decisions. Because price and quantity setting are equivalent for a monopolist, we can use the demand curve (3) to eliminate $P$ and think of the production division as facing a sequence of independent instantaneous profit maximization problems of the form $\max_{a,L,K} \pi = Y^{1/\epsilon} X^{1-(1/\epsilon)} - wL - p_K K$ s.t. (4).

Given that the firm knows technologies in the interval $a \in [0, \alpha]$, which does it use? The following proposition provides the answer.

**Proposition 1** (*Technology choice*). *A firm that has a positive capital stock and that has available Cobb–Douglas technologies with constant returns to scale and capital elasticities in the range* $a \in [0, \alpha]$ *uses the lowest and the highest capital elasticities.*

**Proof.** See the Appendix.

This result says that the firm considers only the extreme technologies corresponding to the highest and lowest values of $a$. To see why, consider a firm with capital $K$ and labor $L$ and consider two values $a_1 < a_2$. If $K/L < 1$, then $(K/L)^{a_1} > (K/L)^{a_2}$ and the firm benefits from assigning all of its capital and labor to the technology with lower capital elasticity. The opposite holds if $K/L > 1$. The key is that the firm can split $K$ and $L$ so to create two capital/labor ratios, one larger and one smaller than 1. It then benefits from assigning the capital and labor in ratio less than 1 to the technology with the lowest capital elasticity and the capital and labor in ratio larger than 1 to the technology with the highest capital elasticity. The insight is that operating the two technologies with extreme values of the capital elasticity yields more output than operating only one because it allows the firm to maximize the marginal product of labor in one—the lowest capital elasticity is in fact the highest labor elasticity—and the marginal product of capital in the other.

The firm's output is thus $X = A[(L-l) + (K^{\alpha} l^{1-\alpha})]$, where $L$ is the firm's total employment, $l \in [0, L]$ is labor allocated to the advanced plant and $L-l$ is labor allocated to the primitive plant. Capital is inessential here: an economy that knows only the primitive technology with $\alpha = 0$ builds no capital and produces with labor only. A more advanced economy with $\alpha > 0$ and some capital can still use the primitive technology and produce output even if the capital disappears. Thinking of capital elasticity as a *choice* rather than an exogenous parameter changes radically the properties of an otherwise conventional economy.

The production division's problem now is

$$\max_{l,L,K}\pi = Y^{1/\epsilon}X^{1-(1/\epsilon)} - wL - p_K K \quad \text{s.t. } X = A[(L-l) + (K^{\alpha}l^{1-\alpha})]. \tag{11}$$

The first-order conditions $\partial\pi/\partial K = 0$ and $\partial\pi/\partial L = 0$ yield the firm's demand for capital and labor. The condition $\partial\pi/\partial l = 0$ gives the allocation of labor across the two plants

$$\frac{\partial\pi}{\partial l} = A\left[-1 + (1-\alpha)\left(\frac{K}{l}\right)^{\alpha}\right] = 0 \Rightarrow l = K(1-\alpha)^{1/\alpha}. \tag{12}$$

The right-hand side of (12) may exceed the firm's total employment $L$. To account for the constraint $l \le L$ we write

$$l = \begin{cases} K(1-\alpha)^{1/\alpha}, & K(1-\alpha)^{1/\alpha} < L \\ L, & K(1-\alpha)^{1/\alpha} \ge L. \end{cases} \tag{13}$$

Similarly, we write the demands for capital and labor as

$$p_K = \begin{cases} Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)A\alpha\left(\frac{K}{l}\right)^{\alpha-1}, & K(1-\alpha)^{1/\alpha} < L \\ Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)A\alpha\left(\frac{K}{L}\right)^{\alpha-1}, & K(1-\alpha)^{1/\alpha} \ge L, \end{cases} \tag{14}$$

$$w = \begin{cases} Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)A, & K(1-\alpha)^{1/\alpha} < L \\ Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)A(1-\alpha)\left(\frac{K}{L}\right)^{\alpha}, & K(1-\alpha)^{1/\alpha} \ge L. \end{cases} \tag{15}$$

The first line of (14) contains $K/l$, whose denominator is *not* the firm's total labor employment $L$ (and thus in general equilibrium is *not* the economy's labor endowment) because the firm splits total labor across the two plants. The first line of (15) says that the value marginal product of labor in the primitive plant determines the wage in the unconstrained case because the firm equalizes the marginal products of labor across the two plants.

To see the advantage of keeping the primitive plant operational in the unconstrained case, define $k \equiv K/l$. Rewrite the unconstrained part of (13) as $k = (1-\alpha)^{-1/\alpha} > 1$ and the two-plant production function as $X = A[L + K(k^{\alpha}-1)/k]$. Note that $(1-\alpha)^{-1/\alpha} = \arg\max_k\{(k^{\alpha}-1)/k\}$, which says that the firm's unconstrained allocation of labor between plants is that which maximizes the marginal product of capital in the advanced plant. Any remaining labor is employed in the primitive technology. This maximizing value of $k$ depends on $\alpha$ alone and so does not diminish with $K$. The reason is that the primitive plant provides the firm with a "labor pool" that allows it to keep the marginal product of capital at its maximum value, determined solely by technology, as it uses more capital. For later discussion, define

$$m(\alpha) \equiv \max_k \frac{k^{\alpha}-1}{k} = \alpha(1-\alpha)^{(1-\alpha)/\alpha}. \tag{16}$$

The function $m(\alpha)$ has the following properties: $m' > 0$ and $m'' > 0$ for all $\alpha \in [0,1]$, $m(0) = 0$, $m(1) = 1$, and $m'(1) = \infty$ (see the Appendix for details). We call $Am(\alpha)$ the "maximized discretionary marginal product," i.e., the part of the maximized marginal product that the firm controls.

The consequences of this static technology choice for dynamics are striking. The firm splits labor across the two plants so that the (maximized) capital/labor ratio in the advanced plant is (a) independent of factor prices, (b) always larger than 1, and (c) increasing in capital elasticity. Property (a) implies that the capital/labor ratio $k$ is different from and independent of the economy's endowment ratio $K/L$. Property (b) implies that the plant's output is increasing in $\alpha$, which provides the rationale both for using only the $a = \alpha$ technology out of the set $(0, \alpha]$ and for pursuing $\alpha-$increasing innovations. Property (c) implies that the development of more capital intensive technologies drives up the optimal capital/labor ratio and thus the incentive to accumulate capital. Together, (b) and (c) yield a positive feedback between capital accumulation and the pursuit of higher capital elasticity.

The investment division, taking the production division's decisions as given, chooses $I$ and $R$ to maximize its present value

$$\max_{\{I_t,R_t\}_{t=0}^{\infty}} \int_0^{\infty} (p_K K_t - I_t - R_t)e^{-\bar{r}_t t}\,dt, \quad \text{s.t. (5), (6), (14), } K_0,\ \alpha_0. \tag{17}$$

The first-order conditions, shown in the Appendix, yield the following expressions for the returns to investment (net of depreciation) and R&D:

$$r_K \equiv \begin{cases} Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)Am(\alpha) - \delta, & K(1-\alpha)^{1/\alpha} < L \\ Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)\alpha A\left(\frac{K}{L}\right)^{\alpha-1} - \delta, & K(1-\alpha)^{1/\alpha} \ge L; \end{cases} \tag{18}$$

$$r_\alpha \equiv \begin{cases} Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)AKm'(\alpha)f(\alpha), & K(1-\alpha)^{1/\alpha} < L \\ Y^{1/\epsilon}X^{-1/\epsilon}\left(1-\frac{1}{\epsilon}\right)AL\left(1+\ln\frac{K}{L}\right)\left(\frac{K}{L}\right)^{\alpha}f(\alpha), & K(1-\alpha)^{1/\alpha} \geq L. \end{cases} \tag{19}$$

These expressions show that the production division's ability to allocate resources between two plants/technologies creates very strong increasing returns.

To characterize the economy's dynamics it is convenient to work with a symmetric equilibrium where firms produce identical quantities and follow identical R&D paths. To do so, we need to introduce a restriction that weakens increasing returns at the firm level so that no firm has an incentive to take over the whole economy. Specifically, we assume

$$\frac{d}{d\alpha}[m'(\alpha)f(\alpha)] < 0 \quad \forall \alpha \in [0,1], \tag{20}$$

so that the firm's return to innovation in (19) is decreasing in $\alpha$. Because $m''(\alpha) > 0$, this assumption can hold only if the function $f(\alpha)$, the productivity of $R$ in the R&D technology (6), is decreasing in $\alpha$. To streamline the presentation, in the text we just assume $f'(\alpha) < 0$ and that (20) holds. In the Appendix we discuss the microfoundations of $f'(\alpha) < 0$ in terms of the process of research and knowledge accumulation by the firm.

### 3.5. Taking stock: the "hybrid" AK model

Let us assess what we have so far. The main innovations of our analysis are that: (a) all prices and quantities are determined in market equilibrium; (b) the firm chooses its capital elasticity $a$ from a set of known technologies $[0,\alpha]$; (c) the firm invests in R&D to expand the technology frontier $\alpha$. The first innovation distinguishes our analysis from that of Seater (2005) and Zuleta (2008b), who consider only the central planning solution.[4] The second innovation changes radically the production structure of the economy and gives rise to what we call the "hybrid AK" model. The third innovation has important implications for long-run growth. In this section we focus on the main features of the hybrid AK model. We discuss its dynamic implications in the next section.

Condition (13) implies that there exists a region of $(\alpha, K)$ space where output is *linear* in labor and capital *separately*. To see this, use (13) to write

$$X = \begin{cases} A[L + m(\alpha)K], & K < L\left(\frac{1}{1-\alpha}\right)^{1/\alpha} \\ AK^{\alpha}L^{1-\alpha}, & K \geq L\left(\frac{1}{1-\alpha}\right)^{1/\alpha} \end{cases} \tag{21}$$

and note two characteristics that are important for the economy's dynamics. (1) In the unconstrained case, for given $\alpha$, capital has *constant* rather than diminishing returns. Given $K$, the firm uses the labor-intensive technology to absorb any excess labor beyond that necessary to yield the output-maximizing capital/labor ratio in the capital-intensive plant. As a result, the firm is always willing to invest in the marginal unit of $K$. (2) The presence of the labor-intensive plant allows the firm to keep the capital/labor ratio $k$ in the capital-intensive plant above 1. An increase in $\alpha$ then increases output, and the firm always is willing to invest in the marginal unit of $\alpha$. This self-reinforcing "virtuous circle" tends to generate perpetual growth.

When the economy operates the labor-only technology, the condition $MPL = A$ holds for all firms. Consequently, all firms price goods according to $P_i(1-1/\epsilon)A = w$. This result is important. Irrespective of the firm's technology level (i.e, the pair $\alpha_i, K_i$), the firm sells the same quantity as all the other firms and thus, according to Eq. (3), the value of its sales is $P_iX_i = Y \cdot (P_i/P_X)^{1-\epsilon}$. Using the pricing equation, we thus have $P_iX_i = Y(1-1/\epsilon)A/w$. In other words, as long as the economy operates the labor-only technology, there is a force that equalizes wages across firms even when they have different technological levels. This result is crucial for the interpretation of our equilibrium: the force equalizing wages is a force equalizing revenues and thus makes factor-eliminating technological change fundamentally different from the traditional cost-reducing technological change studied in, e.g., Peretto (1998, 1999). That type of cost-reducing innovation operates through the chain: cost reduction → price reduction → larger market share → higher profit. That chain is broken here because price reductions (relative to the industry price index) are not feasible.

The question then is why firms do R&D in this model. The answer is that, because the firm earns $P_iX_i = Y(1-1/\epsilon)A/w$ whatever the pair $\alpha_i, K_i$, the reward to R&D is that the firm reduces the total wage bill without scaling down production and revenues. To see how, invert the production function in Eq. (21) to get $L_i = Y(1-1/\epsilon)/w - m(\alpha_i)K_i$, which says that an increase in the firm's capital intensity $\alpha_i$ and use of $K_i$ results in the elimination of labor. Evaluating the firm's profit at this equilibrium in prices and quantities yields $\pi(\alpha_i, K_i) = P_iX_i - wL_i = Y/\epsilon + m(\alpha_i)K_i \cdot w$, which says that the profit of the firm comes in two components: a traditional Dixit–Stiglitz part, given by $Y/\epsilon$, and a novel part capturing the elimination of labor

[4] Our analysis also differs from Zuleta's in that we weaken some of the underlying assumptions. To guarantee existence of a solution, Zuleta restricts total factor productivity to be less than twice the rate of time preference. To study the transition dynamics, he assumes that $K$ can be transformed costlessly into $\alpha$ and vice versa. Our analysis requires neither assumption.

from the payroll in direct proportion to capital use, given by the term $m(\alpha_i)K_i \cdot w$. This property of firm profit conveys more than any other the novelty of our approach. Our mechanism does not rely on the traditional idea that cost-reducing innovation allows the firm to steal business (i.e., market share) from other firms because no such thing occurs in our equilibrium. Rather, what happens is that the firm sells the same amount at the same price and makes more profit because it replaces labor services, whose price is constant at $w=(1-1/\epsilon)A$, with cheaper capital services.

The property has a striking implication: as $\epsilon\to\infty$ only the Dixit–Stiglitz component of the firm profit vanishes, not the factor-elimination component. The reason is that the factor-elimination component does not depend on the traditional notion of profit through product differentiation because it is internal to the firm and fully independent of the firm's interaction with other firms in the marketplace. Consequently, our model does not require monopoly power in the product market, only appropriability of firm-specific technical knowledge.[5]

We can describe the firm's actions as "localized elimination" and "separation" of the factors of production. The firm invests in R&D to learn to produce with less labor. It thus gradually eliminates labor from the advanced plant and so creates endogenously the engine of perpetual growth—a technology that in the limit uses reproducible inputs only. This elimination of non-reproducible inputs is "local" in that it affects only the advanced plant, not the overall production function of the firm. Labor is a productive resource that the firm knows how to use, with or without capital. Leaving idle the labor rendered "surplus" by technical progress would be suboptimal, so the firm "separates" the labor not needed in the advanced plant from capital by employing it in the primitive technology. In doing so, the firm removes diminishing returns from its total technology.

We have discussed a "firm" allocating labor to two different "plants" because it is convenient to set up the allocation problem with only one decision maker. We could decentralize labor allocation and talk about an "economy" that allocates labor to two different "sectors" that, respectively, *specialize* in the primitive and advanced technologies and that are linked by the arbitrage condition that wages be equalized. The exposition would be somewhat longer, but the results would be the same: the economy would bring into existence a technology that eventually uses only capital. As labor demand falls in the capital-using sector, labor would be released and then absorbed by the primitive sector. Firms in the advanced sector would not use the primitive technology, though the economy as a whole would. That is the situation in modern economies, where some activities continue to rely on quite primitive technologies that use unskilled labor and little physical or human capital, such as janitorial services and some forms of agricultural harvesting. Irrespective of which approach is taken, the primitive technology operates even if $\alpha=1$ because otherwise the resource $L$ whose marginal product is strictly positive would be left idle.

## 4. General equilibrium

Given the assumption that capital and technology accumulate internally to the firm, our economy has three markets: final good, intermediate goods, and labor. Because of our assumption of a constant saving rate, the market-clearing condition is $I+R=sY$. The market for intermediate goods is a continuum of monopolistic markets wherein each producer sets his price and thereby chooses a point on the demand curve he faces. Since the mass of products is 1, in symmetric equilibrium $Y=X$.

The labor market is competitive, and households provide labor services inelastically. Equilibrium aggregate employment therefore is the economy's labor endowment $L$, and the wage rate is the value marginal product of that quantity of labor

$$w=\begin{cases}\left(1-\dfrac{1}{\epsilon}\right)A, & K<L\left(\dfrac{1}{1-\alpha}\right)^{1/\alpha}\\ \left(1-\dfrac{1}{\epsilon}\right)A(1-\alpha)\left(\dfrac{K}{L}\right)^{\alpha}, & K\geq L\left(\dfrac{1}{1-\alpha}\right)^{1/\alpha}.\end{cases}\tag{22}$$

To determine how firms split the total flow of savings across investment and R&D, we use the equilibrium wage in (22) to write the returns to $K$ and $\alpha$ in (18) and (19) as

$$r_K=\begin{cases}\left(1-\dfrac{1}{\epsilon}\right)Am(\alpha)-\delta, & K(1-\alpha)^{1/\alpha}<L\\ \left(1-\dfrac{1}{\epsilon}\right)A\alpha\left(\dfrac{K}{L}\right)^{\alpha-1}-\delta, & K(1-\alpha)^{1/\alpha}\geq L;\end{cases}\tag{23}$$

[5] Different products are produced with different technologies, so achieving higher capital intensity in one industry does not draw on the same knowledge base as achieving it in another industry. There are two sides to technological differentiation. On the demand side, differentiation applies to products and makes them different in the eye of the user. On the supply side, differentiation applies to production processes. When $\epsilon\to\infty$ intermediate goods become perfect substitutes for the user, but that does not mean that they are produced with identical technologies.

$$r_\alpha = \begin{cases} \left(1-\frac{1}{\epsilon}\right) AKm'(\alpha)f(\alpha), & K(1-\alpha)^{1/\alpha} < L \\ \left(1-\frac{1}{\epsilon}\right) AL\left(1+\ln\frac{K}{L}\right)\left(\frac{K}{L}\right)^{\alpha} f(\alpha), & K(1-\alpha)^{1/\alpha} \geq L. \end{cases} \tag{24}$$

Because we have bang–bang control (see the Appendix), there are three possible configurations for the pair $(R, I)$: $I>0, R=0$ for $r_K > r_\alpha$; $I>0, R>0$ for $r_K = r_\alpha$; $I=0, R>0$ for $r_K < r_\alpha$. In the first and third cases, the larger rate of return equals the value that makes the positive type of investment (either $R$ or $I$) equal the given amount of saving. In the second case, the two rates of return are equal and at a level that makes the sum of the two types of investment equal saving.[6]

To analyze dynamics we construct two loci. The first is the *arbitrage locus*, along which agents are indifferent between allocating resources to investment or R&D, that is, where $r_\alpha = r_K$. The second is the *stationarity locus*, along which total net investment $\dot{K} + \dot{Z} = \dot{K} + \dot{\alpha}/f(\alpha)$ is zero and the "total capital stock" $K+Z$ is constant.

To construct the arbitrage locus we substitute (23) and (24) into the no-arbitrage condition $r_\alpha = r_K$ and manipulate terms to obtain the following result.

**Proposition 2** (*Arbitrage locus*). *Assume* $L \geq 1 \geq \alpha$.[7] *Then the arbitrage locus in* $(\alpha, K)$ *space is*

$$K = \begin{cases} 0, & 0 \leq \alpha \leq \overline{\alpha} \\ \frac{1}{m'(\alpha)f(\alpha)}\left[m(\alpha) - \frac{\delta\epsilon}{A(\epsilon-1)}\right], & \overline{\alpha} < \alpha \leq 1 \end{cases} \tag{25}$$

*where*

$$\overline{\alpha} = \begin{cases} 1 & \frac{\delta\epsilon}{A(\epsilon-1)} > 1 \\ \arg\text{solve}\left\{m(\alpha) = \frac{\delta\epsilon}{A(\epsilon-1)}\right\} & \frac{\delta\epsilon}{A(\epsilon-1)} \leq 1. \end{cases}$$

*The locus is increasing in the interval* $(\overline{\alpha}, 1)$. *If* $\lim_{\alpha\to 1} m'(\alpha)f(\alpha) = 0$, *the locus converges asymptotically to the vertical line* $\alpha = 1$. *If* $\lim_{\alpha\to 1} m'(\alpha)f(\alpha) > 0$, *the locus converges with smooth pasting to the vertical line* $\alpha = 1$ *at a positive, finite value of* $\alpha$. *The locus lies in the region* $K < L(1-\alpha)^{-1/\alpha}$ *where the interior equilibrium* $l < L$ *holds.*

**Proof.** See the Appendix.

To construct the stationarity locus we use (5), (6) and (21) to write the equilibrium condition of the final goods market as

$$R + I = \frac{\dot{\alpha}}{f(\alpha)} + \dot{K} + \delta K = \begin{cases} sA[L + m(\alpha)K] & K < L\left(\frac{1}{1-\alpha}\right)^{1/\alpha} \\ sAK^{\alpha}L^{1-\alpha} & K \geq L\left(\frac{1}{1-\alpha}\right)^{1/\alpha}. \end{cases} \tag{26}$$

Upon imposing $\dot{\alpha}/f(\alpha) + \dot{K} = 0$ and rearranging terms, we obtain the following result.

**Proposition 3** (*Stationarity locus*). *The stationarity locus in* $(\alpha, K)$ *space is*

$$K = \begin{cases} L\left(\frac{sA}{\delta}\right)^{1/(1-\alpha)}, & \frac{sA}{\delta} \geq e \text{ or } e > \frac{sA}{\delta} \geq 1, \ 0 \leq \alpha \leq \hat{\alpha} \\ L\frac{sA}{\delta - sAm(\alpha)}, & 1 > \frac{sA}{\delta} \text{ or } e > \frac{sA}{\delta} \geq 1, \ \hat{\alpha} \leq \alpha \leq 1, \end{cases} \tag{27}$$

*where*

$$\hat{\alpha} = \arg\text{solve}\left\{\frac{sA}{\delta} = \left(\frac{1}{1-\alpha}\right)^{(1-\alpha)/\alpha}\right\}$$

*and* $\delta/sA - m(\alpha) > 0 \ \forall\alpha$. *The locus has two branches. One branch lies in the positive quadrant, starts at* $LsA/\delta$, *is increasing, and approaches the asymptote* $\alpha = \tilde{\alpha}$, *where* $\tilde{\alpha}$ *satisfies* $m(\tilde{\alpha}) = \delta/sA$. *The other branch lies to the right of* $\alpha = \tilde{\alpha}$ *in the negative quadrant.*

[6] We also have used an interest rate $r$ in the formal definition of the firm's problem, but $r$ plays no role in the solution (because the saving ratio is fixed) and is determined after the fact by the paths of $r_K$ and $r_\alpha$, just as in the Solow model.

[7] If we take $L$ to be unskilled labor only and use natural units to count population (i.e., use the census count), then $L \geq 1$ because the smallest possible economy comprises one person. With this interpretation, the initial assumption in Proposition 7 would be satisfied trivially because $\alpha \in [0, 1]$. The situation is less clear when $L$ is taken to be all non-reproducible factors of production, for which there is no obvious reason why $L \geq \alpha$ need hold.

**Proof.** See the Appendix.

Henceforth, for simplicity, we restrict attention to the unconstrained case; see the Appendix for the constrained case.[8]

The economy's dynamics depend crucially on whether the two loci intersect, which depends on whether $sA/\delta > 1$ or $sA/\delta < 1$, which we call the high and low saving cases, respectively. Figs. 1 and 4 show the phase diagrams for the two cases.

In both figures, points below the arbitrage locus yield $r_\alpha < r_K$, so that $R = \dot{\alpha}/f(\alpha) = 0$ (which implies that $\dot{\alpha} = 0$ because $\dot{\alpha} \gtreqless 0$ as $\dot{\alpha}/f(\alpha) \gtreqless 0$) and $I = \dot{K} - \delta K > 0$, whereas all points above the locus yield $r_\alpha > r_K$, so that $R > 0$ and $I = 0$. Points on the locus yield $r_\alpha = r_K$ so that any combination of $R$ and $I$ for which $\dot{\alpha}/f(\alpha) + \dot{K} = 0$ is possible. In summary, points below, above, or on the stationarity locus yield $\dot{\alpha}/f(\alpha) + \dot{K} > 0$, $\dot{\alpha}/f(\alpha) + \dot{K} < 0$, and $\dot{\alpha}/f(\alpha) + \dot{K} = 0$, respectively. We now discuss the dynamics in each case.

Fig. 1 shows the phase diagram when $sA/\delta > 1$. The dotted line is the labor constraint boundary. The hyperbolic curves in the upper left and lower right corners together are the stationarity locus. The remaining curve is the arbitrage locus for the case where $\alpha = 1$ is reached in finite time. When $\alpha = 1$ is reached only asymptotically, the arbitrage locus is asymptotic to the vertical line $\alpha = 1$. In either case, the arbitrage locus lies everywhere below the positive branch of the stationarity locus (see the Appendix). The arbitrage stationarity loci divide the phase plane into three regions, labelled I, II, and III. In region I, total accumulation is positive, $\dot{\alpha}/f(\alpha) + \dot{K} > 0$, and the rate of return to R&D is less than the rate of return to capital, $r_\alpha < r_K$. The latter fact implies, by the bang–bang nature of the problem, that gross capital investment $I$ is positive and R&D is zero. Because R&D is zero, technology $\alpha$ stays constant. Total accumulation is positive, however, so gross investment $I$ exceeds depreciation $\delta K$, and the capital stock $K$ grows. The resulting equilibrium paths are vertical lines pointing north. In region II, we still have $\dot{\alpha}/f(\alpha) + \dot{K} > 0$, but $r_\alpha > r_K$ so that R&D is positive and gross capital investment is zero. Hence, $\alpha$ grows, and $K$ falls because of depreciation. The resulting paths point southeast. Region III is like region II except that total accumulation is negative, $\dot{K} + \dot{\alpha}/f(\alpha) < 0$, meaning that although $\dot{\alpha}/f(\alpha)$ is positive, it is smaller than depreciation. As in region II, the paths point southeast.

Consider the case where the economy reaches $\alpha = 1$ in finite time. From that time forward the aggregate production function is $Y = X = A(L + K)$, a function that is never exactly $AK$ because the primitive sector always operates but that becomes $AK$ as the primitive sector becomes a negligible part of the aggregate economy. The growth rate is

$$\frac{\dot{X}}{X} = \frac{Am(\alpha)K}{A[L + m(\alpha)K]}\left(\frac{\dot{\alpha}}{\alpha}\ln\left[\left(\frac{1}{1-\alpha}\right)^{1/\alpha}\right] + \frac{\dot{K}}{K}\right). \tag{28}$$

Evaluating this expression along the arbitrage locus and imposing the resources constraint, the paths of the rates of return and the associated growth rate can be characterized analytically as in Figs. 2 and 3 respectively (see the Appendix for details). Interestingly, once $\alpha = 1$ the expression for the growth rate simplifies to $\dot{X}/X = sA - \delta K/(L + K)$, which says that after the economy reaches $\alpha = 1$ the growth rate converges to the limit $As - \delta$ from above.

If the economy does not reach $\alpha = 1$ in finite time, the advanced sector becomes $AK$ only as $t \to \infty$ and thus is never exactly $AK$. Qualitatively, however, the economy's dynamics are identical to the previous case: the aggregate technology $A[L + m(\alpha)K]$ becomes $AK$ in the limit as $K$ goes to infinity *and* $\alpha$ goes to 1.

The intuition behind these results is that the production structure of our economy allows the typical firm to allocate "surplus" labor to a separate plant and thereby neutralize diminishing returns to capital. As a result, firms face incentives to accumulate capital and to develop more capital-intensive technologies that reinforce each other. As the advanced technology becomes $AK$, either in finite time or asymptotically, labor is still an input in the overall production structure but the economy *creates* a technology (equivalently, a production sector) that uses reproducible inputs only and thus satisfies the condition for endogenous growth discussed in Rebelo (1991).

A similar logic drives the economy's behavior at the origin. In its primitive state $\alpha = K = 0$, the economy nonetheless has output because capital is not essential. The economy therefore can produce capital according to the accumulation technology (5) and figure out how to use it according to the research technology (6). Interestingly, the economy starts by building up knowledge first, and only when it has developed technologies with sufficient capital elasticity (once $\alpha$ reaches $\overline{\alpha}$) does it start building capital. The reason is that with low values of both $K$ and $\alpha$ the advanced sector does not generate enough output to overcome depreciation, making construction of the capital unprofitable.

We can characterize the time paths of several variables of interest from the foregoing results. To keep the discussion tractable, we confine attention to the path that starts at the origin and moves along the arbitrage locus in Fig. 1. The capital elasticity $\alpha$ begins growing at once, rising along an upward sloping concave path. It either reaches its upper bound of 1 in finite time at $t_1$ or it grows forever, going to 1 asymptotically. Capital $K$ remains at zero until time $\overline{t}$, after which it grows monotonically without bound. The time paths for the rates of return $w$, $r_K$, and $r_\alpha$ are shown in Fig. 2. The non-reproducible factor $L$ earns the return $w$, which is fixed by the linearity of the primitive technology and the constant mark-up at the value $A(\varepsilon - 1)/\varepsilon$.[9] Capital's return $r_K$ is net of depreciation, so it starts negative and becomes positive at time $\overline{t}$, after which it rises

[8] The only result that is different between the two cases is that the economy temporarily shuts down the primitive sector when the labor constraint binds. That behavior characterizes many firms in the real world, but it is not an accurate description of whole economies, even the most advanced. For example, all economies we know of continue to use unskilled labor paired with little or no physical capital for janitorial services, stocking shelves, some forms of harvesting, and so on. Thus the unconstrained case in fact may be the more realistic one for the economy as a whole.

[9] If $L$ is unskilled labor, then $w$ is the wage paid to it. If all or part of $K$ is embodied in and owned by labor (as human capital would be), then the total wage paid to labor would be $w$ plus the return paid to $K$. Thus it is not wage rates in general that are constant here, only unskilled labor's wage. The same considerations apply if $L$ is any other kind of non-reproducible factor that embodies all or part of $K$.

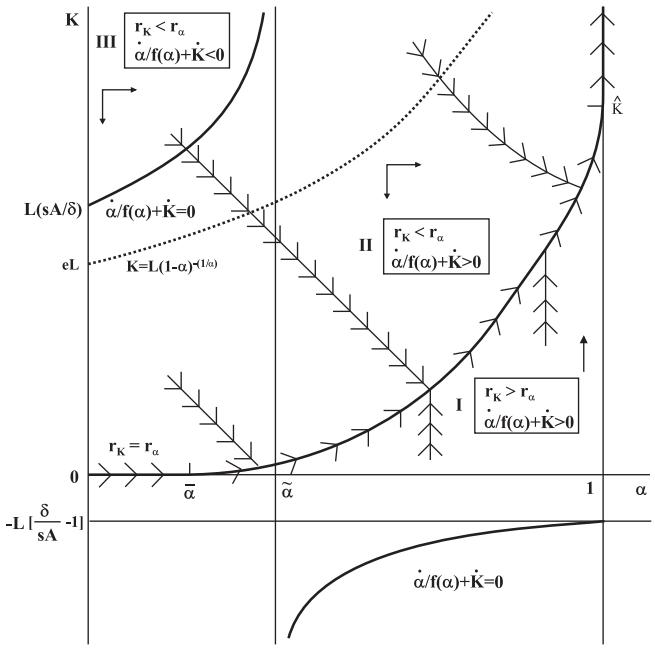


**Fig. 1.** Phase diagram: high saving.

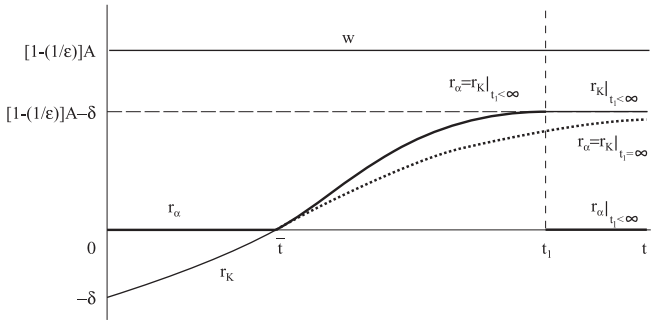


**Fig. 2.** Time path: rates of return.

monotonically until time $t_1$, when $\alpha$ reaches its upper bound of 1. If $t_1$ is infinite, $r_K$ rises forever. The solid path in Fig. 2 applies when $t_1$ is finite, and the dotted path applies when $t_1$ is infinite.

The rate of return $r_\alpha$ is zero until $\bar{t}$, then rises monotonically until $t_1$, when it drops to zero. Between $\bar{t}$ and $t_1$, the two rates of return $r_K$ and $r_\alpha$ are equal. Fig. 3 shows the time path for the growth rate of output. Initially, the growth rate remains constant at zero until the economy starts building capital at time $\bar{t}$. The growth rate then grows monotonically until time $t_1$, when $\alpha$ reaches its limit of 1. The growth rate's limiting value is the growth rate for the *AK* model, which is $sA-\delta$. If $\alpha$ reaches 1 in finite time, output's growth rate overshoots its limit, reaching a peak at the moment $t_1$ that $\alpha$ arrives at 1. At that moment, the growth rate discontinuously jumps down, and then falls monotonically to its limit. The discontinuity arises because at $t_1$ investment in R&D abruptly drops to zero. When the economy transfers investment from R&D to physical capital, the kick to growth that had been delivered by growth in $\alpha$ ends discontinuously, leading to the drop in output's growth.

Fig. 4 shows the phase diagram when $sA/\delta < 1$. The figure is drawn for the case where the arbitrage locus converges with smooth pasting to the vertical line $\alpha = 1$ above the point where the stationarity locus intersects $\alpha = 1$. The results are exactly the same if the arbitrage locus is asymptotic to $\alpha = 1$. In either case, the arbitrage locus intersects the stationarity locus at some $\alpha \in (0, 1)$. The intersection of the two loci creates the new region IV in which total accumulation $\dot{K} + \dot{\alpha}/f(\alpha) < 0$ and $r_\alpha < r_K$, implying that R&D is zero and net capital investment is negative. The equilibrium paths point south. The dynamics in regions I–III are as before. The portion of the stationarity locus below the arbitrage locus, shown as the heavy dashed arc between point **x** and the vertical line $\alpha = 1$, is a set of "Solow" steady states in which the economy has no growth because there is no exogenous growth in population or technology. Even the point where the stationarity locus hits $\alpha = 1$ and production in the advanced plant becomes *AK* is a no-growth steady state. From any starting point, the economy converges to levels of $K$ and $\alpha$ at which the amount of saving is exactly enough to maintain $K$.

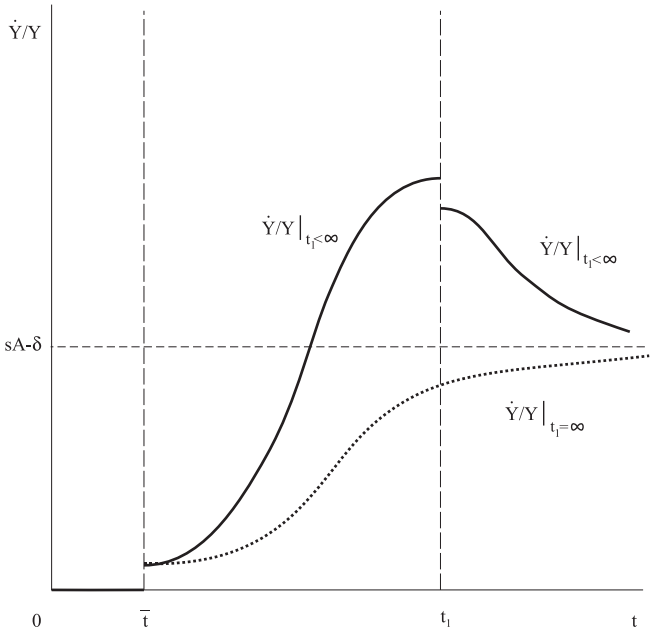


**Fig. 3.** Time path: growth rate of output.

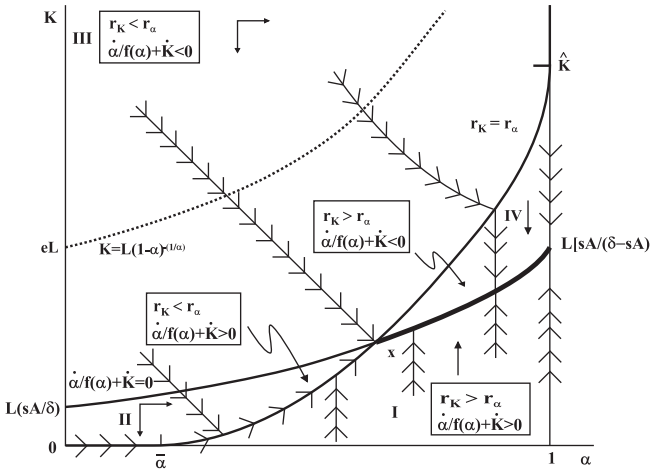


**Fig. 4.** Phase diagram: low saving.

If the arbitrage locus converges with smooth pasting to the vertical line $\alpha = 1$ at or below the point where the stationarity locus intersects that line, the heavy arc in Fig. 4 collapses to the single point where the arbitrage locus intersects $\alpha = 1$, and that point is the unique steady state of the economy. This outcome is quite interesting and highlights the difference between our model and the standard *AK* story. As the economy reaches $\alpha = 1$ and creates the sector that supports endogenous growth, a well-defined equilibrium exists even when the saving rate is too low. Specifically, the economy obeys $\dot{K} + \delta K = sA(L + K)$ and converges to $K = LsA/(\delta - sA) > 0$. The property that allows this outcome is that, although the marginal product of capital is constant, the *average* product of capital is a decreasing function of *K*. In the standard *AK* story, in contrast, the marginal and average product of capital coincide and $sA/\delta < 1$ yields perpetual *shrinking* of the capital stock.

A comparison of the high and low saving cases proves the following result.

**Proposition 4** (*Perpetual growth condition*). *The economy exhibits perpetual growth if and only if* $s \geq \delta/A$.

In the foregoing discussion, the economy may approach $\alpha = 1$ asymptotically or reach it in finite time. The determining condition is whether the limit of $m'(\alpha)f(\alpha)$ is zero or positive. If $\lim_{\alpha \to 1} m'(\alpha)f(\alpha) = 0$, the arbitrage locus approaches the vertical line $\alpha = 1$ asymptotically. Otherwise the arbitrage locus converges with smooth pasting to the vertical line $\alpha = 1$ and the economy reaches $\alpha = 1$ in finite time. In the Appendix we explore the microfoundations of the process of knowledge

accumulation that decide which case holds. The discussion highlights the interesting interplay between the forces that justify $f'(\alpha)<0$ and the forces that yield $m'(\alpha)>0$ and $m''(\alpha)>0$. From that discussion we conclude that convergence to $\alpha=1$ in finite time does not require extreme or implausible assumptions and therefore is a sensible case in the framework of our theory. Literally, *AK* production means that no human intervention is needed.[10] Apparently it is feasible not only in fiction but also in economic theory and, to the extent that the theory fits the facts, in reality.

## 5. Some important implications

The most important implication of our theory is that factor-eliminating technical change can generate perpetual growth without imposing the assumptions that have been subject to the "linearity critique" discussed by Jones (2005) and Growiec (2007). Jones notes that all existing growth models have a linearity in the system of differential equations governing growth, and Growiec provides a formal proof that such a linearity is in fact necessary. Consider a vector of variables $Z$ whose growth rate, denoted $\hat{Z}$, is governed by the equation $\hat{Z}=F(Z)$. A balanced growth path exists if $d\hat{Z}/dt=DF(Z)\cdot\dot{Z}=0$, that is, if either $\dot{Z}=0$ or $DF$ is singular. The first possibility means that $Z$ does not grow, so the necessary condition for perpetual growth is singularity of $DF$. As Growiec notes, such a condition is imposed either explicitly or implicitly in *all* growth models developed to date. Jones and Growiec both argue that there is no compelling reason for any version of the necessary singularity assumption. It is imposed simply because steady-state growth requires it. Our model in contrast imposes no *a priori* singularity on the dynamical system but delivers perpetual growth anyway because factor-eliminating technical change allows an economy to transform its production function from one that cannot sustain perpetual growth into one that can.[11]

In our model the origin is not a steady state. Output at the origin is positive since the primitive production of intermediate goods requires only labor. Positive output is accompanied by positive investment in R&D and physical capital, which generates economic growth. This behavior seems realistic. Humans started with nothing but their labor and through their own efforts moved away from that state. The image we have in mind is the insightful ape in the first act of the movie *2001: A Space Odyssey*, who has no capital but discovers the use of tools. In our theory, however, the ape figures out with his own wits how to use tools rather than getting his inspiration from the Monolith. No exogenous, supernatural spark is required to light the fire of discovery.

Pricing above marginal cost is not necessary for positive R&D. The crucial element is excludability of the fruits of firm-specific R&D. If we let the elasticity of substitution $\epsilon$ go to infinity, the firm's pricing power disappears but the arbitrage and stationarity loci remain well-behaved and our qualitative results are unchanged.

## 6. Evidence

Our theory yields several testable implications. Extended tests of the theory are beyond the scope of this paper, so we present only a brief discussion of three testable implications and their relation to readily available data. Our cursory examination suggests good conformity between the theory and the facts.

Our theory predicts that the share of national income paid to non-reproducible factors

$$\frac{wL}{Y}=\frac{1}{1+m(\alpha)\frac{K}{L}}\left(1-\frac{1}{\epsilon}\right), \tag{29}$$

falls over time. Data on income shares are limited but what data are available suggest that the model's prediction agrees with the facts. Begin with two types of non-reproducible factors: unskilled labor and land.

Bound and Johnson (1995) and Krueger (1999) present evidence that unskilled labor's income share of the US economy has been falling. Krueger reports that the share was down to 6 percent by the mid-1990s. At the same time, the income share of skilled labor has been rising (Blanchard, 1997). Recall that our variable $K$ is broadly defined and includes human capital. An increase in $\alpha$ therefore tends to increase skilled labor's income share.

The return on land is difficult to measure because much of it comes as capital gains, not included in national income accounts. Nevertheless, estimates of land's income share are available for England for the period 1600–2000. Land's share was about 25% in 1600 (Clark, 2001), but by 2000 it had dropped to about 0.1% (Bar and Leukhina, 2010).[12]

Our theory predicts that output always grows and that the capital elasticity $\alpha$ grows most of the time (everywhere except in the interior of region I in Fig. 1 and the interior of regions I and IV in Fig. 4). Thus the model predicts that output and capital elasticity should be positively correlated on average. Our theory includes elements of imperfect competition, so the capital elasticity is not the same as the factor share. However, if we are willing to use factor shares as approximate measures of factor elasticities, as is routinely done in calculating the Solow residual, we can test the model. We do not have sufficiently

[10] Science fiction buffs will recognize this as the Krell technology from the movie *Forbidden Planet*.

[11] The linearity critique refers to the economy's dynamical system only. It does not refer to other aspects of the economy's structure, such as whether there are constant returns to scale. Our theory assumes constant returns to scale on the basis of the replication argument, which is not subject to the critique.

[12] The 2000 estimate is based on data from the UK National Statistics, kindly provided to us by Oksana Leukhina.

long time series data on factor shares to do a time series test, but we can do a cross-section test using cross-country data that recently have become available.

Caselli and Feyrer (2007) use World Bank data to construct income shares for reproducible capital (physical capital) and natural capital (land and natural resources). We regress their income share series on output per worker. The income share for total capital is significantly negatively related to output per worker, as previous investigators have found. The striking new finding is that the income share for reproducible capital is significantly *positively* related to output per worker, as our theory predicts. The two results together imply that the income share of natural capital is significantly negatively related to output per worker, also as our theory predicts. Table 1 reports the regression results, and Figs. 5 and 6 plot the shares as well

**Table 1**
Factor shares vs. output per worker $\alpha_j = b_0 + b_1(Y_j/L_j) + \epsilon_j$.

| Dependent variable | $b_0$ | $b_1$ | $\overline{R}^2$ |
|---|---|---|---|
| *Caselli and Feyrer* | | | |
| Total capital | 0.4000 | −2.01E−06 | 0.1189 |
| | (0.0273) | (7.27E−07) | |
| Reproducible capital | 0.1437 | 1.54E−06 | 0.1055 |
| | (0.020040) | (5.55E−07) | |
| *Zuleta* | | | |
| Total human factors | 0.7815 | −0.0193 | 0.8580 |
| | (0.0041) | (0.0011) | |
| Human capital | 0.1659 | 0.0620 | 0.4415 |
| | (0.0276) | (0.0076) | |

*Note*: Heteroskedasticity-consistent standard errors are in parentheses.

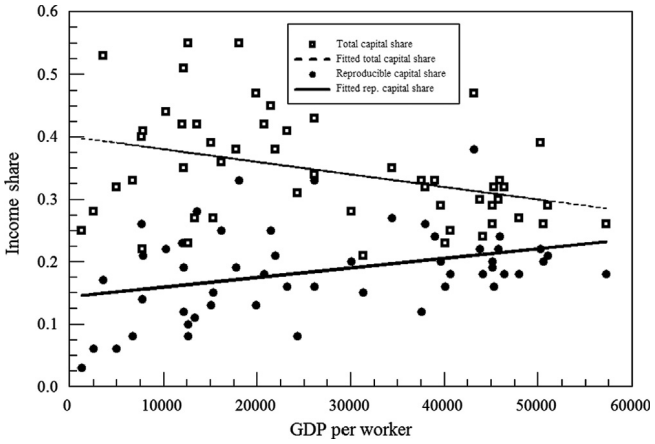


**Fig. 5.** Income shares: total and reproducible physical capital vs. GDP per worker.

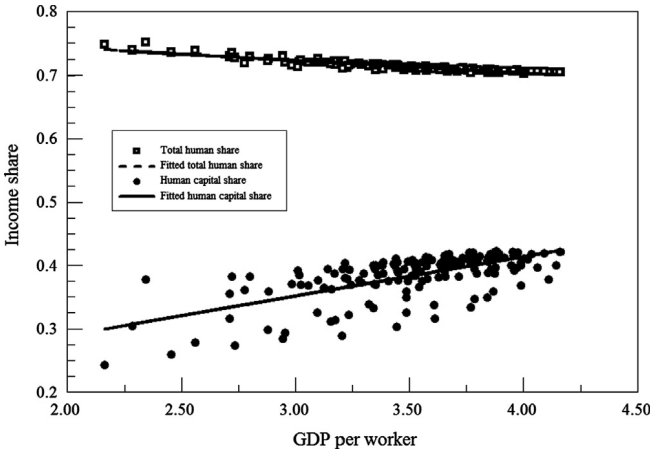


**Fig. 6.** Income shares: total labor and human capital along vs. GDP per worker.

as their regression lines against output per worker. Caselli and Feyrer do not provide data on the income share of human capital, but Zuleta has kindly provided us with such data for the OECD countries. Those data show a slight and statistically insignificant negative relation between the income share of total human factors and output per worker, but a strong and statistically significant positive relation between the income share of just the human capital component of human income and income per worker, as our theory predicts.[13]

Measuring the division of income between labor and everything else is not difficult but dividing labor income between the parts paid to raw labor and human capital is. It is also not easy to divide the return to capital between the parts paid to the reproducible and non-reproducible components. One therefore should maintain some skepticism about the foregoing discussion. Nonetheless, the behavior of the various factor shares measured in various ways uniformly agree with the predictions of our theory, which suggests that the agreement is not merely a spurious result of bad data.

Variable factor shares have important implications for explanations of cross-country differences in output per worker. The standard approach is to assume that factor shares are the same across countries, estimate TFP from the data, and then decompose cross-country income differences into the parts arising from differences in TFP and from other observables, such as physical and human capital. The typical finding is that about half the cross-country differences arise from differences in TFP and half from differences in the other observables (e.g., Klenow and Rodriguez-Clare, 1997). Sturgill (2012), however, finds that factor shares differ enormously across countries. Physical capital's share ranges from 0.04 to 0.41, human capital's share from 0.08 to 0.53, natural capital's share from 0.06 to 0.45, and unskilled labor's share from 0.13 to 0.49. The large cross-country differences in factor shares suggest that existing estimates of TFP variation suffer from serious measurement error and raises the question of what fraction of cross-country income differences are attributable to cross-country variation in factor shares, something not considered in the literature heretofore. It could well be that factor-elimination is the most important source of cross-country income differences and perhaps the dominant form of technical progress. Physical and human capital are reproducible factors, and natural capital (land, natural resources) and unskilled labor are non-reproducible factors. Our theory predicts that the shares of physical and human capital should be positively related to income per person and the shares of natural capital and unskilled labor should be negatively related to income per person. That is exactly what Sturgill finds. Sturgill's results are consistent with our theory.

A related but somewhat less direct test of our theory concerns the relative prices of factors of production. According to our theory, the relative prices of non-reproducible factors fall for a time if the economy is in the constrained equilibrium where the primitive technology shuts down. Each factor's price reflects the value of that factor's marginal product, which depends on the factor's exponent in the Cobb–Douglas function, so a decrease in the exponent for a non-reproducible factor implies a fall in that factor's relative price.

This last point shows that our theory provides a theoretical foundation for the Prebisch–Singer hypothesis that commodity prices fall as time passes. Recently, Harvey et al. (2010) have provided strong confirmation that the hypothesis holds for many commodities. They examine the relative prices of 25 commodities over 400 years, finding a statistically significant negative trend for 11 of them and an insignificant trend for the remaining 14. Their results are consistent with our theory.

## 7. Conclusion

We have proposed a theory of endogenous technical progress that alters output's factor elasticities. The theory can deliver perpetual economic growth without any sort of factor augmenting technical change. In particular, the *AK* model is the asymptotic limit of the economy if the saving rate is sufficiently high. The theory avoids the singularity/linearity critique by making the property of technology that is necessary for perpetual growth an endogenous outcome of the growth process itself. An economy that initially does not satisfy the necessary restriction changes its technology so that eventually it does satisfy the restriction. The theory also offers an explanation of economic transition that does not rely on exogenous technical change or on the existence of latent modern technologies that the economy activates when the right conditions arise. Human progress is entirely the result of human activity and human choices.

If we think of unskilled labor as a non-reproducible factor, our theory has important implications for income distribution dynamics. Assuming that the economy starts at or near the origin, unskilled labor's share drops along the equilibrium path. It never disappears, but it goes asymptotically to zero. This outcome is similar to what already has happened to land's share in industrialized economies. On the other hand, *skilled* labor's share rises because human capital, which is subsumed under our broadly defined “capital,” earns a wage that increases faster than the growing national income.

Our theory's predictions are consistent with both the time series and cross-section evidence documenting that factor shares change over time, with the shares of reproducible factors rising and those of non-reproducible factors falling.

Time variation in factor elasticities has an important implication for the measurement of total factor productivity. The standard practice is to compute such measurements under the assumption that factor shares are the same across countries. Studies using that approach find that cross-country differences in TFP are of paramount importance in explaining cross-country

[13] Gollin (2002) reports a statistically insignificant relation between capital shares and output per person, and Bernanke and Gürkaynak (2001) report a statistically significantly negative relation. Their results have no bearing on our theory, which says nothing about the income shares of *total* capital or *total* labor but rather about the income shares of *reproducible* and *non-reproducible* inputs.

differences in economic performance. If, however, factor elasticities differ substantially across countries, it is not appropriate to summarize all differences in technology by differences in TFP. Variation in factor elasticities has implications for econometric tests in general. It is standard when using Cobb–Douglas production functions to assume that the exponents are constant. That assumption is false if there is factor-eliminating technical change, and imposing it will lead to invalid estimates or interpretations.

The theory suggests several lines of future research, both theoretical and empirical. An obvious extension is to include both factor-augmenting and factor-eliminating technical change because empirical evidence suggests that both occur. We have done an initial exploration along these lines, and the theory seems tractable and interesting. Another is to make the saving rate endogenous. On the empirical side, it would be useful to re-calculate TFP across countries allowing for cross-country variation in factor shares so that we could judge how much of the variation in cross-country economic performance depends on differences in TFP and how much depends on differences in factor elasticities.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org.10.1016/j.jmoneco.2013.01.005.

## References


Bar, M., Leukhina, O., 2010. Demographic transition and industrial revolution: a macroeconomic investigation. Review of Economic Dynamics 13 (April), 424–451.

Bernanke, Ben S., Gürkaynak, Refet S., 2001. Is growth exogenous? Taking Mankiw, Romer, and Weil seriously. In: NBER Macroeconomics Annual, vol. 16, pp. 11–57.

Blanchard, Olivier J., 1997. The medium run. Brookings Papers on Economic Activity 2, 89–158.

Bound, John, Johnson, George, 1995. What are the causes of rising wage inequality in the United States? Economic Policy Review, Federal Reserve Bank of New York, January 1995, pp. 9–17.

Caselli, Francesco, Feyrer, James, 2007. The marginal product of capital. Quarterly Journal of Economics 122 (May), 535–568.

Clark, Gregory, 2001. The Secret History of the Industrial Revolution. Working Paper, University of California at Davis, October 2001.

Galor, Oded, Weil, David N., 2000. Population technology and growth: from Malthusian stagnation to the demographic transition and beyond. American Economic Review 90 (September), 806–828.

Gollin, Douglas, 2002. Getting income shares right. Journal of Political Economy 110 (April), 458–474.

Growiec, Jakub, 2007. Beyond the linearity critique: the knife-edge assumption of steady-state growth. Economic Theory 31 (June), 489–499.

Hansen, Gary D., Prescott, Edward C., 2002. Malthus to Solow. American Economic Review 92 (September), 1205–1217.

Harvey, David I., Kellard, Neil M., Madsen, Jakob B., Wohar, Mark E., 2010. The Prebisch–Singer hypothesis: four centuries of evidence. Review of Economics and Statistics 92 (May), 367–377.

Jones, Charles I., 2005. In: Philippe, Aghion, Durlauf, Steven N. (Eds.), Growth and Ideas Handbook of Economic Growth, Part 2, vol. 1. Elsevier, pp. 1063–1111. Chapter 16.

Kamien, Morton I., Schwartz, Nancy L., 1968. Optimal 'induced' technical change. Econometrica 36 (January), 1–17.

Klenow, Peter J., Rodriguez-Clare, Andres, 1997. The neoclassical revival in growth economics: has it gone too far? In: NBER Macroeconomics Annual, vol. 12, pp. 73–103.

Krueger, Alan B., 1999. Measuring labor's share. American Economic Review 89 (May), 45–51.

Peretto, Pietro F., 1998. Technological change and population growth. Journal of Economic Growth 3 (December), 283–311.

Peretto, Pietro F., 1999. Cost reduction, entry, and the interdependence of market structure and economic growth. Journal of Monetary Economics 43 (February), 173–195.

Rebelo, Sergio, 1991. Long run policy analysis and long run growth. Journal of Political Economy 99 (June), 500–521.

Sato, Ryuzo, Beckmann, Martin J., 1968. Neutral innovations and production functions. Review of Economic Studies 35 (January), 57–66.

Seater, John J., 2005. Share-altering technical progress. In: Finley, L.A. (Ed.), Economic Growth and Productivity, Nova Science Publishers, Hauppauge, NY, pp. 59–84.

Sturgill, Brad, 2012. The relationship between factor shares and economic development. Journal of Macroeconomics 34 (December), 1044–1062.

Zuleta, Hernando, 2008a. An empirical note on factor shares. Journal of International Trade & Economic Development 17 (September), 379–390.

Zuleta, Hernando., 2008b. Factor saving innovations and factors income share. Review of Economic Dynamics 11 (October), 836–851.