The *Ironies of Automation*: A Growth Theory Perspective

Enric Martorell*

Banco de España

February 23, 2026

[Link to latest version]

Abstract

How does the interaction between AI and human capital accumulation affect economic growth? Humans learn skills by doing tasks, but AI automates them. At the same time, AI itself needs some human supervision for its own development. I formalize this tension in a tractable endogenous growth model where human capital and AI co-evolve. I show that if AI augments learning only modestly, the economy converges to a constant ratio of human to AI capital and endogenous growth stops. At the critical threshold, constant returns sustain endogenous balanced growth. Beyond it, the economy either explodes or collapses depending on the initial level of human capital. Task complementarity in production makes the price mechanism that should protect scarce labor instead accelerate displacement. With heterogeneous workers, a thin tail of high-ability workers sustains the economy while the rest become obsolete. The binding constraint on AI-driven growth is the economy's capacity to sustain the human capital that AI development requires. Designing AI to enhance learning rather than replace it is the only policy that ignites endogenous growth.

JEL: O40, O33, J24, E24

Keywords: endogenous growth, automation, human capital, artificial intelligence

*The views expressed in this manuscript are solely those of the author and do not necessarily reflect the views of Banco de España or the Eurosystem.

1. Introduction

Humans build skills by doing tasks. When robots or artificial intelligence (AI) take over those tasks, they destroy opportunities for learning-by-doing, and with them the very human capital that is required for AI development and supervision. A junior programmer who has never debugged production code might find it difficult to develop the judgment needed to oversee the AI that replaced her. Bainbridge (1983) first identified this tension in the human factors engineering literature. Four decades later, these *Ironies of Automation* are gaining relevance in a global economy undergoing AI adoption across a growing number of sectors.

This paper formalizes these dynamics in an endogenous growth model in which human capital and AI co-evolve. The economy produces output by combining a continuum of tasks that are complements. Each task is assigned to whichever factor, human or AI, can perform it at lower cost, making the automation threshold endogenous to relative productivity (Acemoglu and Restrepo, 2018). Workers accumulate human capital through learning-by-doing, but the opportunity to learn depends on the quantity of tasks that remain available to them. AI shapes this process through two opposing forces. First, it crowds out human learning by automating tasks, but it can augment the human learning capacity on the non-automated ones. Second, AI capital also evolves dynamically. It improves through the existing stock, requires human supervision, and decays via model drift when it is not maintained. A growth decomposition identifies three channels through which automation shapes output: crowding out practice, reweighting factor shares, and shifting the depreciation mix. These forces create a co-evolutionary system in which automation erodes the learning opportunities that sustain both human and AI accumulation. That is where the ironies are embedded: automation can bound itself by eroding the very human capital that sustains it.

The model produces a bifurcation in long-run growth governed by whether AI helps workers learn fast enough to offset the practice that automation destroys. In the *mediocrity trap*, AI augments human learning only modestly. The ratio of human to AI capital is constant in the long run as endogenous growth stops. At the exact boundary, constant returns in reproducible factors sustain *endogenous balanced growth*: both human capital and AI grow at a constant positive rate, as in the AK model (Frankel, 1962; Romer, 1986; Rebelo, 1991). When AI augments learning powerfully enough that each generation of human capital and technology compounds the next, the fate of the economy depends on initial conditions. If the economy begins with enough human capital to sustain the learning feedback, it enters a regime of *superagency*: explosive co-growth in which human capital and AI

reinforce each other.[1] If not, the same reinforcing loop works in the opposite direction and the economy enters an *obsolescence spiral* where both human and AI capital collapse.

Another irony arises from the production structure. When tasks are highly complementary, the scarcity market premium on human skill that should slow displacement—the *weak links* (Jones, 2011; Jones and Tonetti, 2026)—instead accelerates it. Complementarity makes the few remaining human-performed tasks expensive bottlenecks. The natural response to a bottleneck is for firms to automate the costly human tasks. The irony is self-reinforcing. As human capital erodes, the bottleneck worsens, the price signal strengthens, and the incentive to automate grows.

I extend the model to include workers who differ in learning ability. Even within the mediocrity trap, the distribution of human capital polarizes: most workers cluster near full automation and become obsolete while a thin Pareto tail of high-ability workers becomes super-productive and sustains aggregate output. A representative agent with the mean technology ratio would underestimate the automation experienced by the typical worker. Moreover, task complementarity makes the rising wage per unit of scarce human capital fail to translate into a larger labor share. Endogenous task reallocation shifts production toward AI faster than scarcity raises the marginal product.

The model yields clear policy implications. Designing AI systems to help workers learn, rather than merely replace humans, dominates traditional education investment as a tool to reignite endogenous growth. The central lesson is that the binding constraint on AI-driven growth is not the economy's capacity to automate, but its capacity to sustain the human capital that automation demands.

Related Literature. This paper builds on the task-based automation framework (Zeira, 1998; Acemoglu and Restrepo, 2018, 2019, 2022), which treats the allocation of tasks between labor and capital as endogenous. I add a dynamic human capital feedback through learning-by-doing. Displacement does not merely reallocate tasks, it erodes the practice that sustains human capital, and this erosion feeds back into the automation decision. Existing models of new task creation and worker augmentation (Autor, 2015, 2024; Agrawal et al., 2026; Freund and Mann, 2026) use static human capital endowments or take skill profiles as fixed. Ide and Talamàs (2025) show that autonomous AI degrades worker-solver matches in a static knowledge-hierarchy model, providing an organizational micro-foundation for the human capital erosion I endogenize dynamically.

[1] I borrow the term *superagency* from Hoffman (2025) to describe a regime where AI amplifies human capability faster than it replaces it.

Several papers analyze whether AI can sustain or accelerate growth (Aghion et al., 2019; Nordhaus, 2021; Jones, 2024; Trammell and Korinek, 2024; Brynjolfsson and McAfee, 2014; Korinek and Juelfs, 2025), focusing on idea production or knowledge accumulation as the binding constraint. I identify a distinct bottleneck: the human skills required to develop and supervise AI are themselves endogenous to the automation process. Even if AI eventually automates idea production, the transition requires human oversight that automation may be degrading in real time.

The growth and learning-by-doing literature (Arrow, 1962; Frankel, 1962; Ben-Porath, 1967; Romer, 1986; Lucas, 1988) established that human capital accumulation and knowledge externalities can drive sustained growth. I apply this mechanism to a setting where the opportunity to learn is endogenous to the automation decision. The distributional analysis relates to work on automation and inequality in growth frameworks (Prettner and Strulik, 2020; Hemous and Olsen, 2022); the novel channel is that automation erodes learning heterogeneously, generating polarization through differential exposure to crowding out. The weak-links production structure (Jones, 2011; Jones and Tonetti, 2026) motivates the complementarity between tasks: in my model, this amplifies the price feedback governing task allocation, making crowding-out more aggressive when human capital becomes scarce.

Experimental evidence documents both augmentation and atrophy from AI adoption (Noy and Zhang, 2023; Brynjolfsson et al., 2023; Dell'Acqua et al., 2023), yet aggregate effects remain muted (Yotzov et al., 2026). Agrawal et al. (2026) read this as supporting pure augmentation. I draw a different lesson: the net effect depends on whether AI augments learning faster than automation erodes practice.

Outline. The remainder of the paper is organized as follows. Section 2 develops an endogenous growth model where human capital and AI co-evolve. Section 3 extends the model to include workers who differ in learning ability. Section 4 discusses policy implications. Section 5 concludes.

2. The Model

I develop an endogenous growth model in which human capital and AI co-evolve through learning-by-doing and task automation. The model has three building blocks: a task-based production technology with endogenous automation, a human capital accumulation process, and an AI accumulation process.

2.1 Production Technology and Task Allocation

The economy produces final output from a continuum of tasks $j \in [0,1]$. Each task j has output $y(j)$ and can be performed by either AI capital or human labor. Task-level outputs aggregate via a constant-elasticity-of-substitution (CES) technology:

$$Y_t = Z_t \left[\int_0^1 y(j)^{\frac{\sigma-1}{\sigma}} \, dj \right]^{\frac{\sigma}{\sigma-1}}, \tag{1}$$

where Z_t denotes Hicks-neutral total factor productivity growing at exogenous rate $g_Z \geq 0$, and $\sigma \in (0,1)$ is the elasticity of substitution between tasks. The restriction $\sigma < 1$ imposes that tasks are gross complements. A threshold $\beta_t \in [0,1]$ partitions the task space: tasks $j \in [0, \beta_t]$ are automated at productivity A_t, while tasks $j \in (\beta_t, 1]$ are performed by human labor at productivity H_t.

With uniform within-group productivity, task-level output is $y(j) = A_t$ for $j \in [0, \beta_t]$ and $y(j) = H_t$ for $j \in (\beta_t, 1]$. Substituting into (1) and integrating:

$$Y_t = Z_t \left[\beta_t \, A_t^{\frac{\sigma-1}{\sigma}} + (1-\beta_t) \, H_t^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

Because $\sigma < 1$, the factor in shorter effective supply becomes the bottleneck that constrains output—a *weak links* structure in the sense of Jones (2011); Jones and Tonetti (2026).

The aggregate wage equals the marginal product of human capital:

$$w_t = \left. \frac{\partial Y_t}{\partial H_t} \right|_{\beta_t} = Z_t^{\frac{\sigma-1}{\sigma}} (1-\beta_t) \left(\frac{Y_t}{H_t} \right)^{1/\sigma}, \tag{3}$$

and the aggregate return to AI capital is

$$r_t = \left. \frac{\partial Y_t}{\partial A_t} \right|_{\beta_t} = Z_t^{\frac{\sigma-1}{\sigma}} \beta_t \left(\frac{Y_t}{A_t} \right)^{1/\sigma}. \tag{4}$$

The derivative holds the automation share β_t fixed at its optimum, as justified by the envelope theorem: the marginal task is indifferent between factors, so $\partial Y_t / \partial \beta_t = 0$. The $Z_t^{(\sigma-1)/\sigma}$ term cancels from the factor price ratio r_t/w_t and from all income shares, which depend only on H_t/A_t.

Define the *technology ratio* $x_t \equiv H_t/A_t$. This sufficient statistic summarizes the relative abundance of human capital and AI, and governs both the automation share and the growth rates of both stocks.

2.2 Endogenous Automation

The automation share β_t is not a parameter but an equilibrium outcome. Tasks are assigned to AI or human labor based on comparative advantage (Eaton and Kortum, 2002). For each task j, the productivity of human labor is $H_t z_{H,j}$ and that of AI is $A_t z_{A,j}$, where $z_{H,j}$ and $z_{A,j}$ are independent draws from a Fréchet distribution with shape parameter $\psi > 0$. Each task is assigned to the factor with lower unit cost. Solving for β_t (see Appendix A) yields:

$$\beta_t = \frac{1}{1 + x_t^{\zeta}}, \qquad 1 - \beta_t = \frac{x_t^{\zeta}}{1 + x_t^{\zeta}}, \tag{5}$$

where the *composite elasticity*

$$\zeta \equiv \frac{\psi(\sigma + 1)}{\sigma(1 + \psi)} \tag{6}$$

governs the responsiveness of automation to shifts in the technology ratio x_t.

Equation (5) has three properties central to the dynamics.

First, β_t is a decreasing, sigmoidal function of x_t. When human capital is abundant relative to AI, tasks remain manual; when AI pulls ahead, automation expands.

Second, ζ encodes the interaction between task heterogeneity and price feedback. The Fréchet parameter ψ controls the dispersion of task-level productivities. A low ψ means wide variation across tasks, so that moderate cost shifts flip only marginal tasks. A high ψ compresses this heterogeneity, and even small shifts in relative costs trigger large reallocations.

Third, the elasticity of substitution σ shapes the price mechanism. When $\sigma < 1$, factor prices respond sharply to relative scarcity, amplifying the cost gap that governs task assignment. Accordingly, ζ decreases in σ. Stronger complementarity means a given change in x_t generates a larger equilibrium shift in β_t.[2] In the Leontief limit ($\sigma \to 0$), $\zeta \to \infty$ and automation becomes a knife-edge; conversely, when σ is large, $\zeta \to \psi/(1 + \psi) < 1$ and automation adjusts sluggishly.

The composite elasticity ζ captures how the price mechanism amplifies comparative advantage. Even modest task-level heterogeneity can generate sharp automation responses if tasks are sufficiently complementary, because the price signal does the heavy lifting. This interaction determines how rapidly automation crowds out learning opportunities as AI improves.

[2] The dependence of ζ on σ is a general equilibrium effect. In partial equilibrium, holding factor prices w_t and r_t fixed, the task assignment depends on the cost ratio $(w_t H_t)/(r_t A_t)$, so the responsiveness of β to the technology ratio x_t is governed by ψ alone. All σ dependence in ζ enters through the endogenous response of factor prices to task reallocation.

2.3 Dynamic System

2.3.1 Human capital dynamics

Human capital evolves through a learning-by-doing process (Ben-Porath, 1967). The key assumption is that learning requires practice: workers build and maintain skill by actively performing tasks. AI plays a dual role: it helps workers learn faster, but it also displaces the tasks through which learning occurs. The aggregate law of motion is

$$dH_t = \Bigg[\underbrace{\phi\, H_t^{\rho} \quad \overbrace{A_t^{\theta}}^{\text{augmentation}} \quad \overbrace{(1-\beta_t)}^{\text{practice}}}_{\text{learning-by-doing}} - \underbrace{\delta_h\, H_t}_{\text{atrophy}} \Bigg] dt, \tag{7}$$

where $\phi > 0$ is learning efficiency, $\rho \in (0,1)$ is the elasticity of learning with respect to the current stock of human capital, $\theta \geq 0$ is the elasticity of learning with respect to AI capability, $\delta_h > 0$ is the rate of human capital atrophy, and the automation share β_t is given by (5).

Three channels shape human capital accumulation. The first is *returns to experience*. The term H_t^{ρ} with $\rho \in (0,1)$ implies that a larger stock of human capital supports faster learning, but at a diminishing rate. This concavity is standard in the learning-by-doing literature and ensures that human capital cannot grow explosively on its own.

The second channel is *AI augmentation*. The term A_t^{θ} captures the positive effect of AI on the learning process (coding assistants, personalized tutoring, diagnostic tools). When $\theta > 0$, improvements in AI raise the productivity of human practice. When $\theta = 0$, AI is a pure substitute that displaces labor.

The third channel is *crowding-out*. The term $(1-\beta_t)$ is the share of tasks still performed by human labor. As automation expands, the opportunities for learning-by-doing shrink. A radiologist who no longer reads scans cannot improve at reading scans, regardless of how much AI tutoring is available.[3]

The ironies of automation arise from the interaction of these three channels. Automation reduces $(1-\beta_t)$, which in turn reduces the practice that sustains H_t. However, AI simultaneously augments learning through A_t^{θ} so that the net effect depends on which force dominates: the augmentation that helps workers learn faster, or the crowding-out that deprives them of the opportunity to learn at all.

[3]The model abstracts from the possibility of learning a task without performing it.

Dividing (7) by H_t and substituting $A_t = H_t/x_t$, the human capital growth rate is

$$g_H = \phi\,(1-\beta(x_t))\,x_t^{-\theta}\,H_t^{\rho+\theta-1} - \delta_h, \tag{8}$$

where $x_t = H_t/A_t$ is the technology ratio. The derivation uses $H_t^{\rho-1}A_t^{\theta} = H_t^{\rho+\theta-1}\,x_t^{-\theta}$. The residual factor $H_t^{\rho+\theta-1}$ reflects combined returns $\rho+\theta$ in the two reproducible factors. When $\rho+\theta<1$, the exponent is negative and g_H declines as H_t grows (diminishing returns). When $\rho+\theta>1$, the exponent is positive and g_H rises with H_t (increasing returns). At the boundary $\rho+\theta=1$, the level term vanishes and g_H depends only on the ratio x_t. The qualitative implications of this distinction are characterized in Section 2.5.

2.3.2 AI dynamics

For the dynamics of AI capital, I adopt a specification from the endogenous growth theory literature (Romer, 1990; Jones, 1995). There, researchers produce new ideas by combining effort with the existing stock of knowledge (*standing-on-shoulders*). Here, human supervision combined with the existing AI stock produces new AI capital.

The flow I_t of new AI capital follows a Cobb-Douglas technology:

$$I_t = \lambda\,H_t^{\omega}\,A_t^{1-\omega}, \tag{9}$$

where H_t is human supervision, A_t is the existing stock of AI capital, $\lambda>0$ is AI development total factor productivity, and $\omega\in(0,1)$ is the *supervision share*—the elasticity of AI improvement with respect to human capital. The Cobb-Douglas form implies that both inputs are essential ($I_t=0$ if $H_t=0$ or $A_t=0$). AI cannot improve without human oversight, and oversight alone without an existing stock to build on is unproductive. Moreover, $\omega\in(0,1)$ implies constant returns in the two inputs.

The first input in the AI flow production function is *human supervision*: the stock of human capital H_t determines the capacity to maintain, debug, retrain, and extend AI systems. The same workers who perform tasks in goods production also provide the oversight that keeps AI functional.[4]

The second input is *standing on shoulders*: the term $A_t^{1-\omega}$ captures the idea that existing AI capital enables further AI development.[5]

[4] Human capital operates through the quality of oversight, not a time-allocation tradeoff. The model also treats the skill used in production and the skill required for supervision as the same stock H_t. This follows the original insight of Bainbridge (1983).

[5] This is the standard channel from endogenous growth theory (Jones, 1995). The larger the existing technological stock, the easier it is to produce new capital. In the AI context, each generation of models provides the

Finally, AI systems suffer from some form of depreciation: *model drift*. They degrade as the environment evolves unless retrained on new data.

The key addition to the canonical specification is not the ideas production function itself, but its interaction with the rest of the model. In standard R&D-based growth models, research labor H_A is exogenous or drawn from a fixed pool. Here, the human supervision input H_t is itself endogenous to the automation process. Automation erodes the practice through which workers build and maintain the very skills that AI supervision requires. This creates a novel feedback loop absent from standard growth models.

AI capital therefore evolves as gross investment net of depreciation:

$$dA_t = \left[I_t - \gamma\, A_t\right] dt, \tag{10}$$

where $\gamma > 0$ is the drift rate. Substituting (9):

$$dA_t = \left[\underbrace{\lambda \overbrace{H_t^{\omega}}^{\text{supervision}} \overbrace{A_t^{1-\omega}}^{\text{standing on shoulders}}}_{\text{AI development}} - \underbrace{\gamma\, A_t}_{\text{drift}}\right] dt. \tag{11}$$

Dividing by A_t, the AI growth rate is

$$g_A = \lambda\, x_t^{\omega} - \gamma, \tag{12}$$

where $x_t = H_t/A_t$ is the technology ratio. The derivation uses $H_t^{\omega} A_t^{-\omega} = x_t^{\omega}$: all level terms in A_t cancel (constant returns), so the growth rate depends only on the technology ratio.

Human supervision and model drift are the two forces governing AI accumulation. The term x_t^{ω} captures the role of human oversight: when $A_t \gg H_t$ (x_t small), the stock of human capital is small relative to the AI it must supervise, and the capacity to maintain and extend AI systems degrades through depreciation.

The interaction between (7) and (11) encodes the paper's core mechanism. Automation raises β_t, reducing practice $(1-\beta_t)$ and lowering H_t. Weaker human capital then undermines AI supervision and slows AI development. The two stocks are locked in a race mediated by the supervision channel. This is the core irony of automation in the model.[6]

foundation (pre-trained weights, documented bottlenecks, ...) on which the next generation is built.

[6]Equation (7) and (11) are technological primitives rather than outcomes of household optimization. This places the model in the tradition of learning-by-doing (Arrow, 1962) and R&D-based growth (Jones, 1995).

2.3.3 Equilibrium

The economy consists of a representative firm that produces the final good using AI capital and effective human labor, and a unit mass of identical workers who supply labor inelastically.

Definition 1 (Equilibrium)**.** *An equilibrium is a collection of paths* $\{Y_t, H_t, A_t, \beta_t, w_t, r_t\}_{t\geq 0}$ *such that:*

(i) *the task allocation* β_t *maximizes output given factor supplies, satisfying* (5)*;*

(ii) *factor prices equal marginal products:* w_t *and* r_t *satisfy* (3)–(4)*;*

(iii) *human capital accumulates according to the learning-by-doing technology* (7)*;*

(iv) *AI capital accumulates according to the AI development technology* (11)*.*

2.4 Growth Accounting

The laws of motion (7)–(11) describe how each stock evolves. I now derive a decomposition of output growth.

In equilibrium, because the automation share β_t satisfies the first-order condition, the envelope theorem implies that changes in output arise through changes in TFP and the two factors:

$$dY_t = \frac{Y_t}{Z_t}\, dZ_t + w_t\, dH_t + r_t\, dA_t. \tag{13}$$

Dividing by $Y_t\, dt$ yields

$$g_Y = g_Z + s_H\, g_H + s_A\, g_A, \tag{14}$$

where $s_H \equiv w_t H_t / Y_t$ and $s_A \equiv r_t A_t / Y_t$ are the factor income shares, with $s_H + s_A = 1$.

From the factor prices (3)–(4) and the CES production function (2), the labor share is

$$s_H = \frac{(1-\beta_t)\, x_t^{(\sigma-1)/\sigma}}{\beta_t + (1-\beta_t)\, x_t^{(\sigma-1)/\sigma}}, \tag{15}$$

where $x_t = H_t/A_t$ is the technology ratio, and $s_A = 1 - s_H$. When $\sigma < 1$, the labor share s_H is increasing in x_t: human capital scarcity raises its factor share, a consequence of the complementarity structure.

Substituting (8) and (12) into (14):

$$g_Y = g_Z + \underbrace{s_H\,\phi\,(1-\beta_t)\,x_t^{-\theta}\,H_t^{\rho+\theta-1}}_{\text{learning}} + \underbrace{s_A\,\lambda\,x_t^{\omega}}_{\text{AI development}} - \underbrace{\left[s_H\,\delta_h + s_A\,\gamma\right]}_{\bar{\delta}}, \tag{16}$$

where $\bar{\delta} \equiv s_H\delta_h + s_A\gamma$ is the share-weighted depreciation rate. Output growth decomposes into gross accumulation from human learning and AI development minus effective depreciation. The automation share β_t enters through three distinct channels:

(i) *Practice.* Higher β_t reduces $(1-\beta_t)$ in the learning term, shrinking the practice share and slowing down human capital accumulation.

(ii) *Share reweighting.* Higher β_t shifts the output weights from s_H toward s_A, amplifying the AI contribution and dampening the human capital contribution.

(iii) *Depreciation mix.* Higher β_t shifts $\bar{\delta}$ from the atrophy rate δ_h toward the drift rate γ, with net effect depending on which factor depreciates faster.

Equation (16) makes the role of β_t analytically concrete. In the learning term, automation destroys practice through $(1-\beta_t)$. The AI development term λx_t^{ω} depends on the technology ratio but not on β_t directly. Automation affects AI development only indirectly, by eroding the human capital that enters through x_t. The shares s_H and s_A themselves depend on β_t through (15), creating an additional indirect channel. The net effect of automation on output growth is therefore ambiguous.

2.5 Steady State Analysis

TFP Z_t enters multiplicatively but is absent from the accumulation equations. The long-run analysis therefore reduces to the two-dimensional system in (H_t, A_t).

Definition 2 (Balanced growth path)**.** *A balanced growth path (BGP) is a path along which x_t, β_t, s_H, s_A, and the growth rates g_H and g_A are all constant. When $g_H = g_A = 0$, the BGP reduces to a steady state in levels and output grows at rate $g_Y = g_Z$. When $g_H = g_A = g^* > 0$, both stocks grow at the endogenous rate g^* and output grows at $g_Y = g_Z + g^*$.*

When $\rho + \theta \neq 1$, the BGP corresponds to fixed points (H^*, A^*) where $dH = dA = 0$. At the boundary $\rho + \theta = 1$, no such fixed point exists; the BGP is instead a path along which

the technology ratio x_t is constant and both stocks grow at a common positive rate. I now characterize both cases.

Proposition 1. *Consider the dynamic system*

$$dH = \left[\phi\, H^{\rho}\, A^{\theta}\,(1-\beta(H/A)) - \delta_h\, H\right] dt, \tag{17}$$

$$dA = \left[\lambda\, H^{\omega}\, A^{1-\omega} - \gamma\, A\right] dt, \tag{18}$$

where $\beta(x) = 1/(1+x^{\zeta})$ *and all parameters are positive:* $\phi, \delta_h, \lambda, \gamma > 0$, $\omega \in (0,1)$, $\zeta > 0$. *Moreover,* $A_0 > 0$ *and* $H_0 > 0$.

Then:

(i) *The nullcline ratio* $x^* = H^*/A^*$ *solving* $dA = 0$ *is*

$$x^* = \left(\frac{\gamma}{\lambda}\right)^{1/\omega}. \tag{19}$$

(ii) ***Mediocrity trap*** *(*$\rho + \theta < 1$*). There exists a unique interior steady state* $(H^*, A^*) \in \mathbb{R}^2_{++}$ *with technology ratio* x^**. If* $\omega\gamma > \delta_h(\rho + \zeta\beta^* - 1)$*, it is locally asymptotically stable.*

(iii) ***Superagency vs. Obsolescence*** *(*$\rho + \theta > 1$*). There exists a unique interior steady state* (H^*, A^*) *with technology ratio* x^**. It is a saddle point: economies starting above* H^* *enter explosive co-growth; those starting below* H^* *collapse to the boundary.*

(iv) ***Endogenous balanced growth*** *(*$\rho + \theta = 1$*). No interior steady state in levels exists. The technology ratio* x_t *converges to* x^{**}*, the unique solution to*

$$\phi\,(1-\beta(x))\,x^{-\theta} - \delta_h = \lambda\, x^{\omega} - \gamma. \tag{20}$$

Along the BGP, both stocks grow at the constant endogenous rate $g^* = \lambda(x^{**})^{\omega} - \gamma$.

Proof. See Appendix B.

The stability of the mediocrity trap requires that AI capital depreciates fast enough relative to human capital feedback: $\omega\gamma > \delta_h(\rho + \zeta\beta^* - 1)$. This holds with margin at the baseline parametrization.[7]

[7] AI model generations turn over in 1–3 years (new architectures, distribution shift, technical debt). Human skills erode over decades.

The proposition presents a long-run bifurcation based on $\rho + \theta$, which measures the combined returns to learning in the two reproducible factors. I now discuss the economic intuition behind each case.

2.5.1 Discussion

The parameter ρ captures returns to own experience (i.e. how much current skill amplifies new learning). The parameter θ captures AI's augmentation of the learning process (i.e. how much AI capability amplifies human practice). Their sum $\rho + \theta$ measures the total returns to learning in the bundle of reproducible factors.[8]

When $\rho + \theta < 1$, diminishing returns to learning dominate. Even with AI augmentation, the learning process exhibits decreasing returns in the factors that the economy can accumulate. The system admits a stable interior steady state at which all endogenous variables are constant: the technology ratio x^*, the automation share β^*, the factor shares s_H and s_A, and the levels H^* and A^*. Human capital and AI co-exist at finite levels, automation remains incomplete, and output grows at rate $g_Y = g_Z$, driven entirely by exogenous TFP. I call this regime the *mediocrity trap* because the engine of endogenous growth stops and the technology ratio converges to a constant x^* where production factors stagnate.

When $\rho + \theta > 1$, AI augmentation is strong enough to generate increasing returns. The learning process becomes self-reinforcing: more human capital raises AI productivity, which augments learning, which builds more human capital. The interior steady state becomes dynamically unstable, acting as a knife-edge, and no stable BGP exists. On one hand, economies starting above H^* enter a regime of explosive co-growth in which human capital and AI compound each other without bound. I call this regime *superagency* because AI amplifies human capabilities beyond what either factor could achieve in isolation, making workers not redundant but extraordinarily productive. Along the superagency path, $g_H, g_A > 0$ and accelerating, so output grows faster than g_Z and the technology ratio x_t diverges—automation recedes as human capital outpaces AI. On the other hand, economies starting below H^* collapse into an *obsolescence spiral*. Automation erodes practice, lowering skill, which makes further tasks worth automating. The loop continues until both human capital and AI capability converge to zero. Along this path, $g_H, g_A < 0$, $\beta_t \to 1$, and output growth falls below g_Z. The set of initial conditions from which the economy converges to this boundary is the *obsolescence basin*.

[8] The role of $\rho + \theta$ as the aggregate return to reproducible factors parallels Frankel (1962), who showed that an externality from capital accumulation can raise aggregate returns above the private diminishing return, sustaining endogenous growth.

When $\rho + \theta = 1$, AI augmentation exactly offsets diminishing returns to own learning. The residual factor $H_t^{\rho+\theta-1} = H_t^0 = 1$ drops out of the growth rate (8), so both g_H and g_A depend only on the technology ratio $x_t = H_t/A_t$. No interior steady state in levels exists—the equation that pins down H^* becomes degenerate—but the ratio x_t converges to a well-defined value x^{**} at which both stocks grow at the same constant rate g^*. This is the same mechanism as in the AK model: constant returns in reproducible factors sustain permanent endogenous growth without explosive dynamics. Each unit of human capital generates just enough AI-assisted learning to maintain the same growth rate, regardless of the current level. The balanced growth rate $g_Y = g_Z + g^*$ depends on structural parameters. I call this regime *endogenous balanced growth*.

The knife-edge bifurcation formalizes the irony of AI. The parameter θ governs whether AI augments the learning process or merely displaces it. When $\theta = 0$, AI is a pure substitute: automation raises static efficiency but can only harm human capital accumulation. As θ increases, AI becomes a complement to learning at the same time that it introduces fragility as it approaches the knife-edge. The augmentation effect that could send the economy to superagency is the same mechanism that, if human capital falls below a threshold, accelerates collapse.

The composite elasticity ζ governs the size of the obsolescence basin. Stronger complementarity between tasks (lower σ) raises ζ, which amplifies the automation response to shifts in x_t and widens the basin of attraction for the boundary. Under complementarity, the few remaining non-automated tasks become expensive bottlenecks. The price signal that should protect labor, by rising wages as human skill becomes scarce, instead accelerates reallocation toward machines, because firms cannot tolerate bottlenecks when tasks are complements. Economies with complementary task structures therefore face a larger obsolescence basin.

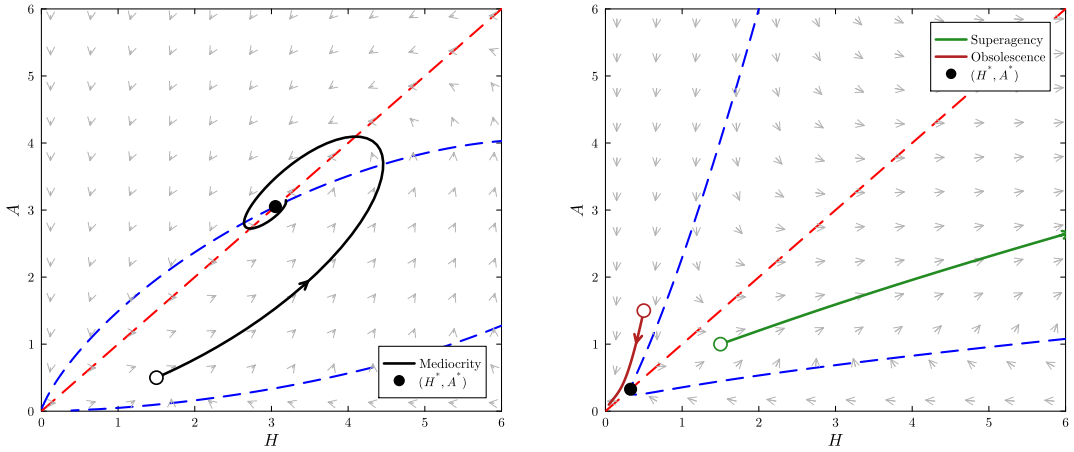
Finally, the ratio x^* is determined entirely by the AI development technology: AI development productivity λ, drift γ, and supervision share ω. The automation share does not enter, so x^* is invariant to the task allocation. This closed-form solution reflects the constant-returns structure of the AI development technology: setting $g_A = \lambda x^\omega - \gamma = 0$ pins down the ratio at which human supervision exactly offsets AI drift.

2.6 Numerical Simulation

I simulate the model to illustrate its dynamics. All parameters are shared except the AI augmentation elasticity θ, which I vary to move the economy across the bifurcation threshold

$\rho + \theta = 1$. These exercises are illustrative and demonstrate the qualitative mechanisms of the model rather than match specific empirical targets.

I set returns to experience at $\rho = 0.5$, implying diminishing returns to own learning, consistent with the human capital literature. When $\theta = 0.3$, the combined returns are $\rho+\theta = 0.8 < 1$, placing the economy in the mediocrity regime. When $\theta = 0.5$, the combined returns are $\rho + \theta = 1$, placing the economy at the balanced growth boundary. When $\theta = 0.7$, the combined returns are $\rho + \theta = 1.2 > 1$, and the interior steady state becomes unstable. The remaining parameters are set to round values that produce clear dynamics: learning efficiency $\phi = 0.25$, atrophy rate $\delta_h = 0.10$, AI development productivity $\lambda = 0.5$, drift rate $\gamma = 0.15$, supervision share $\omega = 0.3$, Fréchet shape $\psi = 1$, and substitution elasticity $\sigma = 0.5$. Exogenous TFP growth is set to $g_Z = 0$, so that all simulation results are interpreted net of exogenous productivity growth.



(a) Stable ($\rho + \theta = 0.8 < 1$)

(b) Unstable ($\rho + \theta = 1.2 > 1$)

Figure 1: Phase diagrams. (a) Stable regime: dashed lines are nullclines; the black trajectory converges to the interior steady state (H^*, A^*). (b) Unstable regime: the interior steady state is a saddle point. The green trajectory diverges; the red trajectory collapses to the origin.

Figure 1a presents the phase diagram under the stable regime with $\rho + \theta = 0.8 < 1$. The $dH = 0$ and $dA = 0$ nullclines intersect at the interior steady state (H^*, A^*). The $dA = 0$ locus is a line through the origin with slope $1/x^*$, reflecting the fixed long-run ratio at which supervision sustains AI capital against drift. The trajectory converges to the mediocrity trap: the economy builds human capital and AI in tandem, but growth decelerates as diminishing returns bind, and the system settles at finite levels of automation and output.

Figure 1b presents the phase diagram under the unstable regime with $\rho + \theta = 1.2 > 1$. The interior steady state is now a saddle point. A trajectory starting above H^* enters superagency: both human capital and AI accelerate without bound as augmentation compounds faster than atrophy erodes. A trajectory starting below H^* falls into the obsolescence spiral. There is no mediocrity trap to land in and the economy either escapes to unbounded growth or collapses entirely.

2.6.1 Transitional dynamics

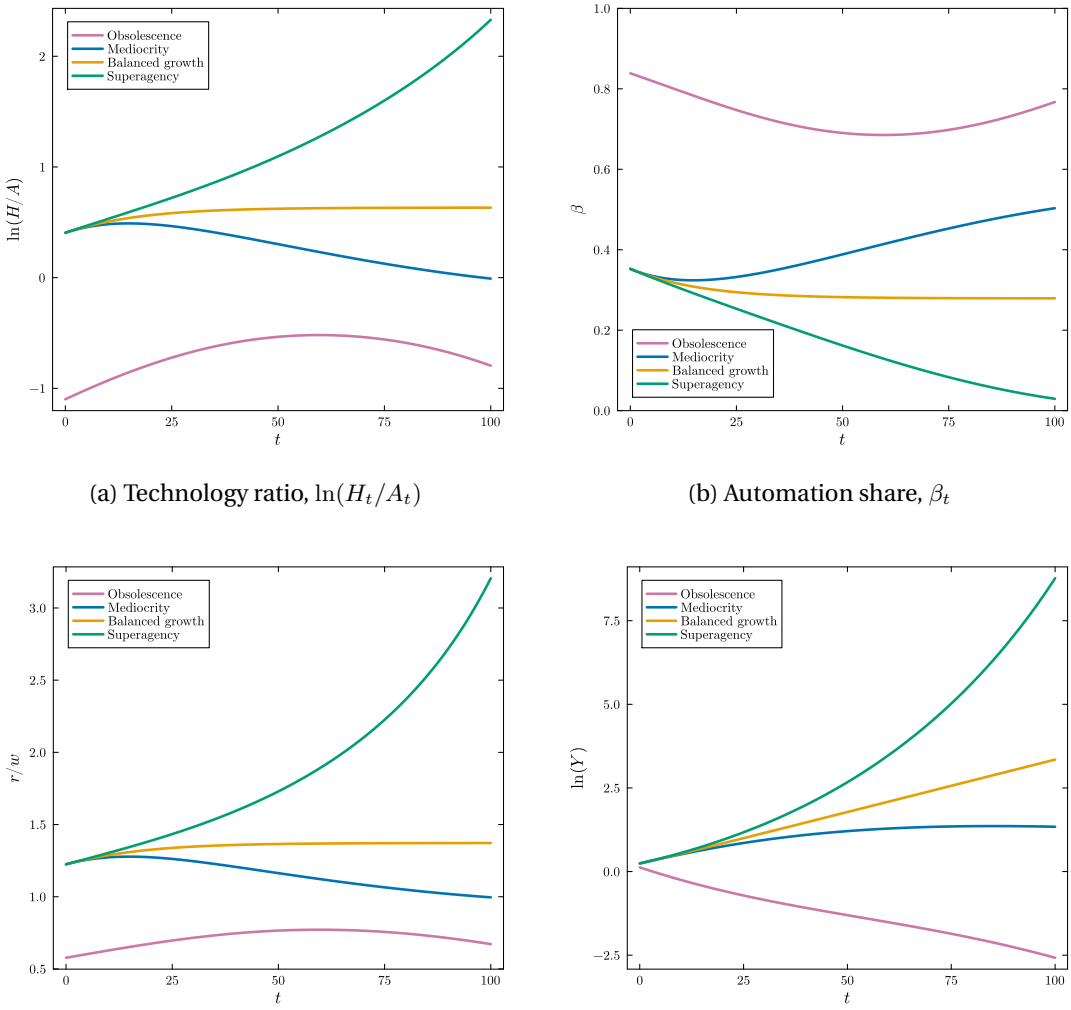
Figure 2 simulates four economies. The mediocrity economy ($\theta = 0.3$, $\rho+\theta = 0.8 < 1$) starts from $(H_0, A_0) = (1.5, 0.5)$ and converges to a stable interior steady state: automation stabilizes at an intermediate level and output reaches a finite level. The endogenous balanced growth economy ($\theta = 0.5$, $\rho + \theta = 1$) starts from the same initial conditions: the technology ratio converges to x^{**}, the automation share stabilizes, and output grows at a constant endogenous rate g^*, visible as a linear trajectory in $\ln(Y_t)$. The superagency and obsolescence economies share the same structural parameters ($\theta = 0.7$, $\rho + \theta = 1.2 > 1$) but differ in initial conditions. The superagency economy starts above the basin boundary ($H_0 > H^*$): AI augmentation compounds human learning so strongly that both state variables accelerate, automation recedes as human capital outpaces AI, and output grows explosively. The obsolescence economy starts below the basin boundary ($H_0 < H^*$): automation rises toward one, factor prices collapse, and output declines monotonically. These two economies are structurally identical—same technology, same parameters—yet their fates diverge depending on the initial stock of human capital.

In economic terms, the difference between mediocrity, balanced growth, and the unstable regime hinges on whether AI augments the learning process strongly enough to generate increasing returns ($\rho + \theta \gtrless 1$). Within the unstable regime, the difference between superagency and obsolescence hinges on initial conditions: an economy that enters the AI era with sufficient human capital escapes to unbounded growth, while one that enters with too little collapses.

2.7 Discussion

Three features distinguish this model from existing frameworks for AI and growth.

First, the binding constraint on growth is human capital formation, not idea production. Models in the tradition of Aghion et al. (2019) and Jones (2024) ask whether AI can automate the research process. The bottleneck there is the production of new knowledge.



(a) Technology ratio, $\ln(H_t/A_t)$

(b) Automation share, β_t

(c) Relative factor price, r_t/w_t

(d) Output, $\ln(Y_t)$

Figure 2: Transitional dynamics for four parameterizations. Mediocrity: $\rho + \theta = 0.8$. Balanced growth: $\rho + \theta = 1.0$. Superagency and obsolescence: $\rho + \theta = 1.2$, differing only in initial H_0.

Here, the bottleneck is different: maintaining and supervising AI requires human capital that the automation process itself degrades. The irony that AI development depends on the skills automation erodes is absent from models where research labor is exogenous or drawn from a fixed pool. Second, the feedback between automation and skill formation is missing from standard task-based models. In Acemoglu and Restrepo (2018) and Acemoglu and Restrepo (2019), automation displaces workers from tasks, but skill endowments do not respond to displacement. Here, displacement destroys the practice through which skills form. This feedback generates the bifurcation at $\rho + \theta = 1$: a condition absent from models with static human capital. Third, the model inverts the standard intuition about weak links. In Jones (2011), complementarity between inputs protects the scarce factor by raising its marginal product. Here, the same price mechanism makes human-performed tasks expensive bottlenecks—incentivizing their automation and accelerating displacement. These channels are not independent but reinforcing loops within a single dynamic system.

The model's parsimony isolates the feedback between automation and skill formation and makes the regime classification analytically tractable. The exact bifurcation at $\rho+\theta = 1$ reflects the multiplicative separability of the learning function and constant returns in AI development. This qualitative regime classification is robust to perturbations of the functional form. Replacing the linear practice term $(1 - \beta)$ with a general decreasing function $f(\beta)$ does not alter the returns structure. β depends on the ratio $x_t = H_t/A_t$, not on levels, the term $f(\beta(x_t))$ is level-free and $H_t^{\rho+\theta-1}$ remains the sole level-dependent factor. The novelty of the regime classification lies not in the existence of a threshold between stagnation and growth but in the economic channels that govern the threshold: AI augmentation of learning (θ) and the feedback between automation and practice erosion.

However, the model's parsimony comes at a cost. Human capital accumulates mechanically through learning-by-doing rather than through forward-looking investment. Also, the augmentation parameter θ is exogenous, though in practice it reflects design choices that respond to market incentives. Endogenizing the augmentation vs automation design of AI is a highly relevant direction for future research.[9] Finally, the task space is fixed, ruling out the creation of new tasks that historically restored labor demand after automation (Acemoglu and Restrepo, 2019).[10]

[9] A directed technical change framework (Acemoglu, 2002) provides a starting point.

[10] Task creation would expand the practice share by adding tasks that only humans can initially perform, counteracting crowding out. The bifurcation, however, is governed by the returns to learning in the reproducible factors, not by the size of the task space. With task creation, the practice share would be higher for any given technology ratio, shrinking the obsolescence basin, but the qualitative instability when $\rho + \theta > 1$ would remain.

3. Heterogeneous Workers

The representative agent model delivers the regime classification but suppresses cross-sectional variation. I now introduce heterogeneity in learning ability. The economy consists of N workers who differ in learning efficiency ϕ_i. Each worker-entrepreneur operates a competitive production unit that combines her own embodied human capital $h_{i,t}$ with the non-rival AI stock A_t. Human capital is specific to each worker and cannot be traded across production units.

Each worker i operates a production unit in which a threshold $\beta_{i,t} \in [0,1]$ partitions tasks between AI and human labor. The CES aggregator yields individual output:

$$y_{i,t} = Z_t \left[\beta_{i,t} \, A_t^{\frac{\sigma-1}{\sigma}} + (1-\beta_{i,t}) \, h_{i,t}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}. \tag{21}$$

The AI stock A_t is non-rival across workers, that is, each production unit draws on the same AI capability.[11] The aggregate production function (2) is the special case in which $h_{i,t} = H_t$ for all i.

The individual wage equals the marginal product of worker i's human capital:

$$w_{i,t} = \left. \frac{\partial y_{i,t}}{\partial h_{i,t}} \right|_{\beta_{i,t}} = Z_t^{\frac{\sigma-1}{\sigma}} (1-\beta_{i,t}) \left(\frac{y_{i,t}}{h_{i,t}} \right)^{1/\sigma}. \tag{22}$$

and the marginal product of AI capital at the individual level is

$$r_{i,t} = \left. \frac{\partial y_{i,t}}{\partial A_t} \right|_{\beta_{i,t}} = Z_t^{\frac{\sigma-1}{\sigma}} \, \beta_{i,t} \left(\frac{y_{i,t}}{A_t} \right)^{1/\sigma}. \tag{23}$$

The derivative holds the automation share $\beta_{i,t}$ fixed at its optimum, as justified by the envelope theorem. Because the CES aggregator is homogeneous of degree one in $(h_{i,t}, A_t)$ at optimal $\beta_{i,t}$, factor payments exhaust output within each production unit: $w_{i,t} \, h_{i,t} + r_{i,t} \, A_t = y_{i,t}$.

Aggregate output is

$$Y_t = \frac{1}{N} \sum_{i=1}^{N} y_{i,t}. \tag{24}$$

Because Y_t is a mean of heterogeneous CES outputs, not itself a CES function, the representative-agent marginal products from (3)–(4) do not describe the cross-section of individual marginal products. Under complementary tasks ($\sigma < 1$), workers with low $h_{i,t}$ automate heavily and

[11] Non-rivalry reflects the near-zero marginal cost of deploying trained AI models to additional users.

their individual $r_{i,t}$ is large: AI relaxes their binding bottleneck. These same workers have high $w_{i,t}$ per unit of human capital—the scarce factor commands a premium—even as total human capital earnings $w_{i,t}h_{i,t}$ collapse. When all workers are identical, the individual marginal products coincide with the representative agent predictions from (3)–(4).

The worker-level automation share follows the same structure as the aggregate, with the individual technology ratio $h_{i,t}/A_t$ replacing the aggregate x_t:

$$\beta_{i,t} = \frac{1}{1 + (h_{i,t}/A_t)^{\zeta}}. \tag{25}$$

I call $\beta_{i,t}$ the *automation risk* of worker i. The aggregate automation share β_t from (5) is the special case when $h_{i,t} = H_t$ for all i.

Each worker i accumulates human capital according to

$$dh_{i,t} = \left[\underbrace{\phi_i \, h_{i,t}^{\rho} \overbrace{A_t^{\theta}}^{\text{augmentation}} \overbrace{(1-\beta_{i,t})}^{\text{practice}}}_{\text{learning-by-doing}} - \underbrace{\delta_h \, h_{i,t}}_{\text{atrophy}} \right] dt, \tag{26}$$

where $\phi_i > 0$ is worker i's learning efficiency. When all workers are identical, the individual equation reduces to the aggregate law of motion (7).

The shared AI stock evolves as a knowledge externality from aggregate supervision:

$$dA_t = \left[\lambda \, \bar{H}_t^{\omega} \, A_t^{1-\omega} - \gamma \, A_t \right] dt, \qquad \bar{H}_t \equiv \tfrac{1}{N} \textstyle\sum_{i=1}^{N} h_{i,t}. \tag{27}$$

Each worker's supervision contributes to AI improvement through aggregate human capital $\bar{H}_t$. This is the only general equilibrium link across production units.

Workers differ in a single dimension: base learning efficiency ϕ_i, drawn from a Pareto distribution,

$$f(\phi) = \frac{\alpha \, \phi_{\min}^{\alpha}}{\phi^{\alpha+1}}, \qquad \phi \geq \phi_{\min}, \tag{28}$$

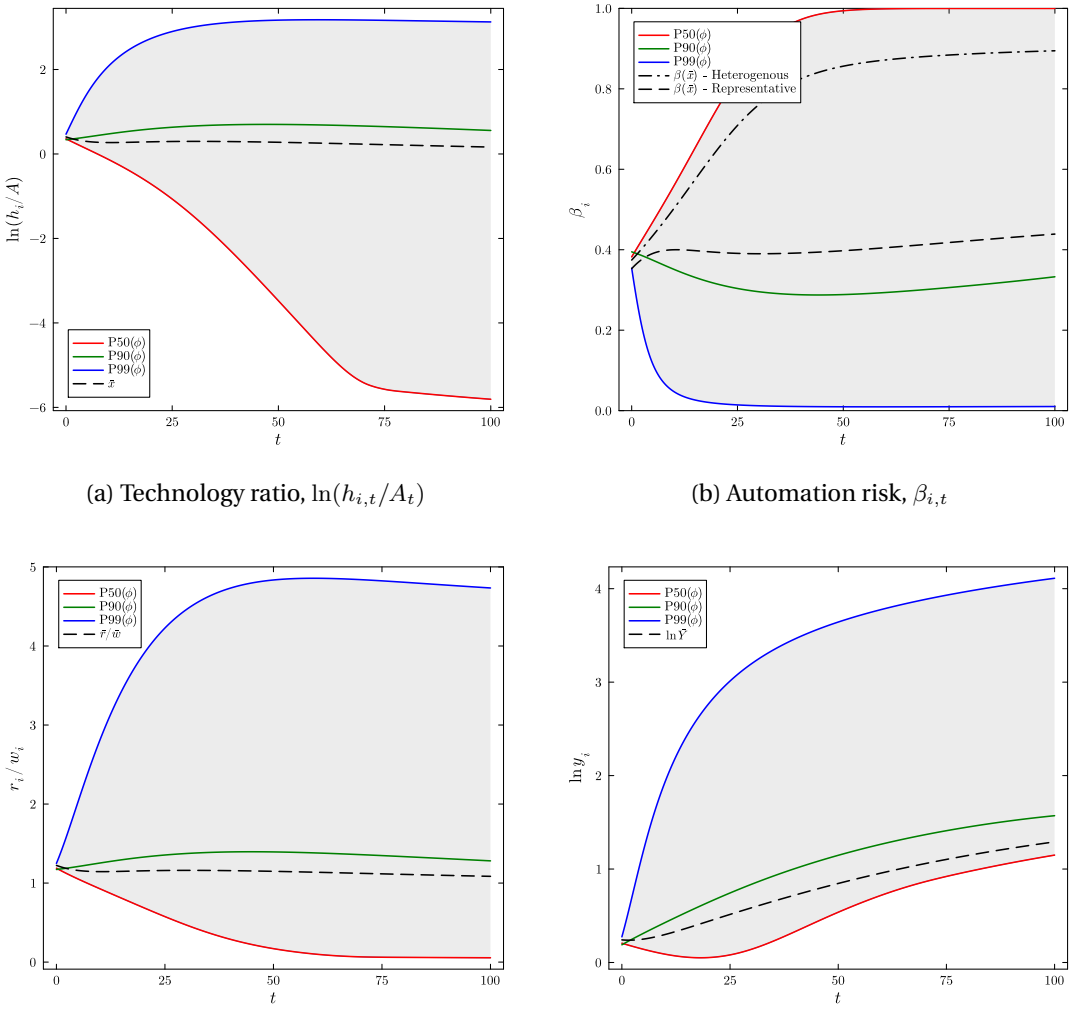
where $\alpha > 1$ is the shape parameter and $\phi_{\min} > 0$ is the lower bound on ability. The Pareto distribution generates the heavy right tail observed in earnings distributions, and it nests the degenerate case $\alpha \to \infty$ in which all workers have ability $\phi_{\min}$ and the model reduces to the representative agent system of Sections 2.1–2.5.

3.1 Numerical Simulation

The representative agent analysis characterizes aggregate dynamics but suppresses the cross-sectional variation. I now simulate the heterogeneous agent model under the mediocrity regime ($\rho + \theta = 0.8$) to explore how the distribution of individual technology ratios $h_{i,t}/A_t$ and automation shares $\beta_{i,t}$ evolves over time and to characterize the long-run ergodic distribution. The simulation draws $N = 10{,}000$ workers with learning abilities ϕ_i from a Pareto distribution ($\alpha = 2.5$). All other parameters match the ones in Section **??**.

Figure 3 tracks the trajectories of three workers located at the P50, P90, and P99 of the learning ability distribution ϕ_i. All four panels show the same individuals, so the link between human capital, automation, factor prices, and output is visible worker by worker. The median-ability worker's automation share rises toward one as her h_i/A ratio falls: AI capital grows while her human capital stagnates, leaving her with no practice opportunities and accelerating atrophy. The P90 ability worker accumulates human capital fast enough to resist displacement. Her β_i slightly falls as h_i/A rises, and then stabilizes in the interior of the automation risk. The P99 worker's automation share drops to near zero, reflecting a technology ratio large enough that the firm's optimal task allocation overwhelmingly favors her to work over AI. The gap between the median and high-ability workers widens during the transition and stabilizes as the economy approaches its ergodic state. The factor price ratio in panel 3c translates these skill dynamics into price signals. Since $r_i/w_i = [\beta_i/(1-\beta_i)](h_i/A)^{1/\sigma}$ automation intensity and relative scarcity pin down the ratio. For the median worker, the collapse in h_i/A dominates. Even though $\beta_i/(1-\beta_i)$ is large, $(h_i/A)^{1/\sigma}$ shrinks faster, driving r_i/w_i toward zero—the per-unit wage rises as human capital becomes scarce, but there is almost no human capital left to earn it. The P99 worker's ratio rises sharply as her large h_i/A compounds through the exponent $1/\sigma$, reflecting the high value of AI in a production unit richly endowed with human capital. Panel 3d shows that log output diverges across workers despite their shared access to the same AI stock.

Panel 3b also reveals a systematic wedge between the cross-sectional mean $\mathrm{E}[\beta_i]$ and the representative agent prediction $\beta(\bar{x})$, where $\bar{x} = \bar{H}/A$ is the mean technology ratio. Because $\beta(\cdot)$ is a bounded, S-shaped function of the individual ratio h_i/A, the cross-sectional mean $\mathrm{E}[\beta_i]$ differs from $\beta(\mathrm{E}[h_i/A])$. As the skill distribution polarizes, most workers accumulate a low h_i/A and face β_i near one, while a thin right tail of high-ability workers has β_i near zero. Because β is bounded and flat near its extremes, the tail workers have a large effect on the mean technology ratio $\bar{x}$—pulling it into a region where β is small—but only a modest effect on $\mathrm{E}[\beta_i]$, which is dominated by the mass of nearly-fully-automated workers. The result is



(a) Technology ratio, $\ln(h_{i,t}/A_t)$

(b) Automation risk, $\beta_{i,t}$

(c) Factor price ratio, $r_{i,t}/w_{i,t}$

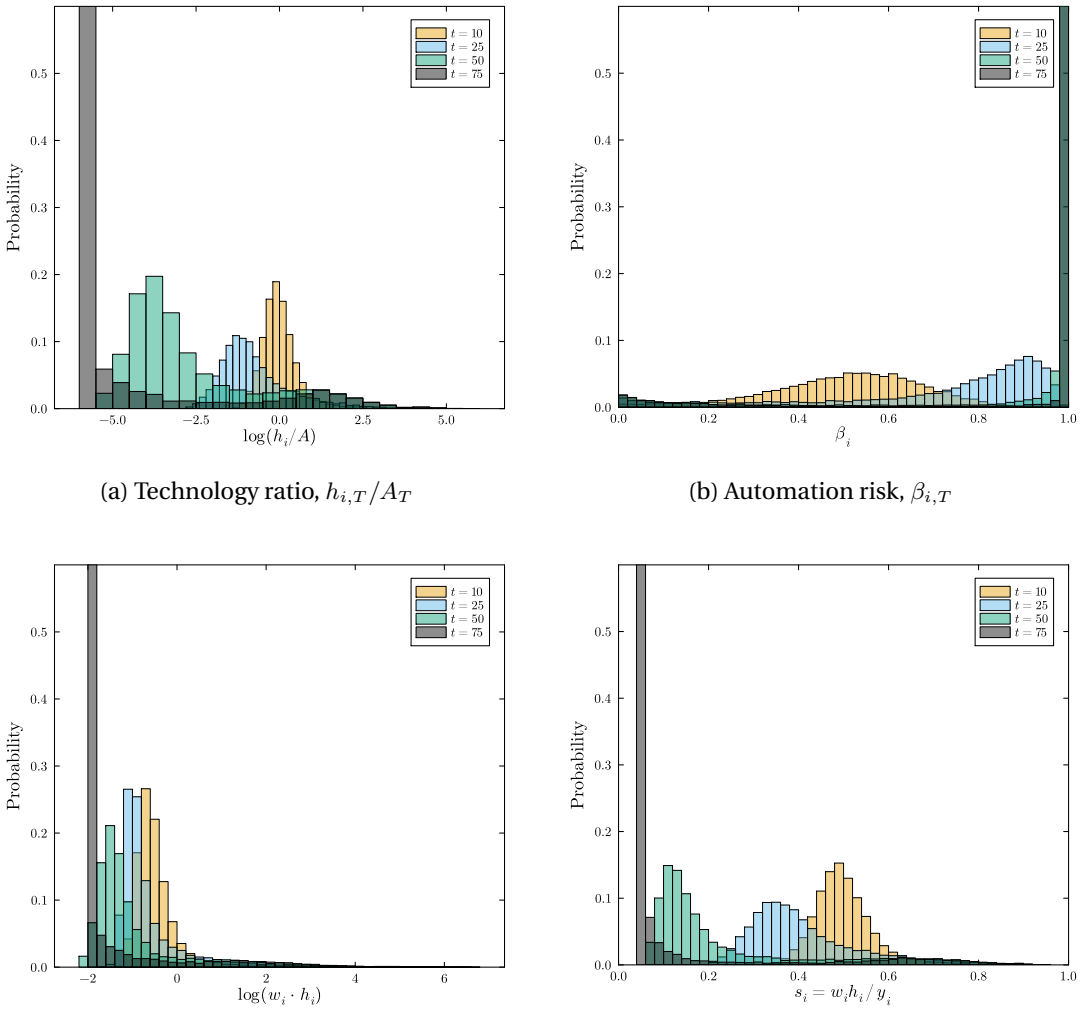
(d) Log output, $\ln(y_{i,t})$

Figure 3: Distributional dynamics under the mediocrity regime ($\rho + \theta = 0.8$). Each panel tracks workers at the P50, P90, and P99 of $\phi_i \sim \text{Pareto}(\alpha = 2.5)$, $N = 10{,}000$. Solid gray lines: cross-sectional mean. Dashed gray in panel (b): representative agent prediction $\beta(\bar{x})$.

$\mathrm{E}[\beta_i] > \beta(\bar{x})$: the average automation share in the heterogeneous economy exceeds the automation share that a representative agent with the mean technology ratio would face. This is relevant because a policymaker relying on aggregate statistics would therefore underestimate the degree of automation experienced by the typical worker. The representative agent model misses the fact that the economy's automation share is driven not by the average worker but by the mass of workers whose skills have already eroded.

Figure 4 traces the cross-sectional distributions at four points in time. At $t = 10$ the distributions are concentrated near their initial values; by $t = 75$ they have polarized into bimodal shapes. The technology ratio in panel 4a develops a large mode at low h_i/A, populated by workers whose human capital has atrophied under near-complete automation, alongside a thin right tail at high h_i/A populated by workers who accumulate skill fast enough to resist displacement. The distribution of automation shares (panel 4b) mirrors this pattern: a mass of workers clusters near $\beta_i = 1$, while a left tail extends toward near-zero automation shares for high-ability workers whose h_i/A is large enough that the optimal β_i from (25) falls well below the median. The bimodal structure emerges as the feedback loop between automation and skill atrophy takes hold.

The earnings distribution in panel 4c is heavily right-skewed. The log-earnings histogram reveals the economic consequences of the polarized skill distribution: highly automated workers cluster at low earnings, while the few workers who resist displacement earn disproportionately more. The individual labor share distribution (panel 4d) rises with skill. The intuition is that endogenous task reallocation dominates the complementarity premium on the scarce factor. A nearly-fully-automated worker's output is mostly AI-generated—human capital contributes little to production—so the labor share is small despite the high marginal product per unit of h. A high-ability worker retains most tasks, human capital dominates output, and the labor share approaches one. This is the distributional face of the complementarity mechanism from Section 2.5: the price signal that should protect labor—rising w_i as h_i becomes scarce—fails to translate into a larger labor share because the firm's optimal response is to automate precisely those tasks the worker can no longer perform. The wage per unit of h rises, but the share of output captured by labor falls, because endogenous task reallocation shifts production toward AI faster than scarcity raises the marginal product.



(a) Technology ratio, $h_{i,T}/A_T$

(b) Automation risk, $\beta_{i,T}$

(c) Log earnings, $\log(w_i h_i)$

(d) Individual labor share, $s_i = w_i h_i / y_i$

Figure 4: Distributional evolution under the mediocrity regime ($\rho + \theta = 0.8$). Each panel overlays the cross-sectional distribution at $t = 10$, 25, 50, and 75 for the same simulation as Figure 3.

3.2 Discussion

Taken together, the numerical analysis reveals several implications. Aggregate output is dominated by the right tail of the human capital distribution. A small number of high-ability workers produce disproportionately because, under complementary tasks ($\sigma < 1$), their balanced endowment of human and AI capital yields far higher output than the unbalanced endowment of highly-automated workers. The typical (median) worker has low skills and near-maximal automation risk, yet the economy appears *stable* because output concentrates at the top. This is the irony of automation operating through the cross-section rather than the aggregate. The representative agent model, which collapses heterogeneity into means, cannot detect the divergence of the human capital distribution.

The distributional dynamics here differ from existing models of automation and inequality. Prettner and Strulik (2020) and Hemous and Olsen (2022) generate inequality through differential exposure to automation across task types or skill groups. In those frameworks, worker skill endowments are fixed or evolve exogenously. Here, the skill distribution is itself endogenous to the automation process: workers who lose tasks also lose the practice through which they maintain skills, and this erosion feeds back into the automation decision. The result is a self-reinforcing polarization mechanism absent from models with static human capital. The nonlinearity wedge between $\mathrm{E}[\beta_i]$ and $\beta(\bar{x})$ is a direct consequence of this endogenous polarization and has a practical implication: aggregate statistics systematically understate the degree of displacement experienced by the typical worker.

Once again, the clarity of these results comes with some model limitations. The economy is a competitive island model in which each worker-entrepreneur operates an independent production unit with embodied, non-tradable human capital. Cross-worker interaction occurs only through the shared AI stock (27): there is no labor market, no reallocation across units, and no search or matching frictions. Workers differ only in learning ability ϕ_i, not in task content or occupational exposure. The key distributional results—polarization of h_i/A, the nonlinearity wedge in automation shares, and output concentration in the right tail—are driven by the individual learning dynamics (26) and the common AI stock, not by cross-worker price interactions. Relaxing the island structure to allow labor market competition would enrich the distributional predictions but would not alter the core mechanism: endogenous skill erosion driven by the feedback between automation and learning.

4. Policy Analysis

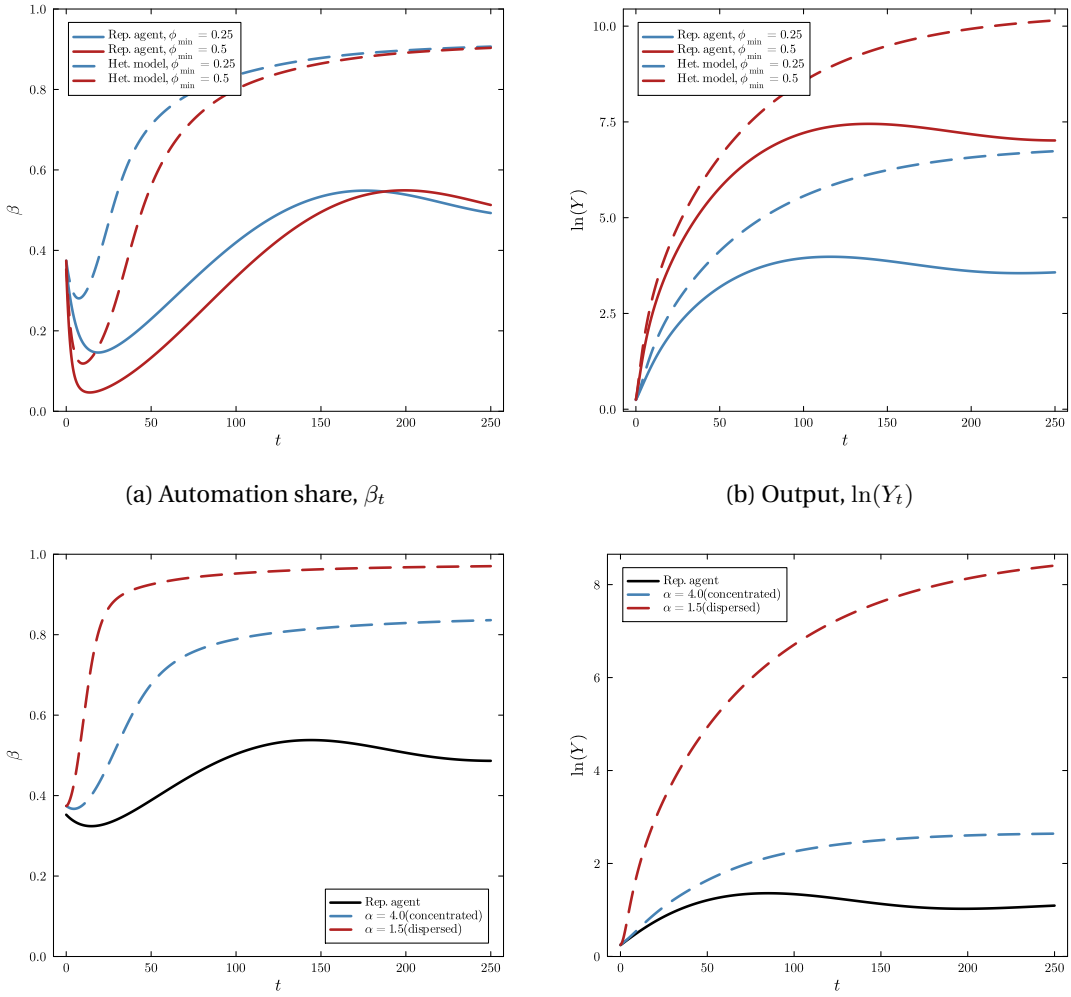
The model's regime structure disciplines which instruments can address the automation–human capital tension and which cannot. I examine two: education investment and AI augmentation strength. They operate on fundamentally different margins. Education raises the level at which the economy stagnates. Augmentation design is the only instrument that can change the regime.

4.1 Education Investment

The natural response to automation-driven human capital erosion is investment in education. In the model, this would be proxied by raising the learning efficiency ϕ. This will increase human capital accumulation without altering the regime boundary. To see why, recall the steady-state characterization from Proposition 1. The steady-state ratio $x^* = (\gamma/\lambda)^{1/\omega}$ does not involve ϕ. The automation share β^* is therefore invariant to education investment. Higher ϕ raises the steady-state levels H^* and A^* proportionally but leaves the ratio, and hence the factor distribution, unchanged. ϕ enters the human capital equation (7) as a scale parameter, not a returns parameter. It cannot push $\rho + \theta$ above one. As a consequence, more education produces a richer stagnation within the mediocrity regime.

Panels (a)-(b) in Figure 5 compare the representative agent and heterogeneous models under two education levels ($\phi_{\min} \in \{0.25, 0.5\}$, with ϕ set to the Pareto mean $\alpha\phi_{\min}/(\alpha-1)$). Panel (a) confirms that the representative agent β_t (solid lines) converges to the same β^* for both education levels. The dashed lines tell a different story: the cross-sectional mean $\mathrm{E}[\beta_{i,t}]$ from the heterogeneous model settles near 0.9, far above the representative agent prediction of $\beta^* \approx 0.5$. This is the nonlinearity wedge from Section 3. Raising $\phi_{\min}$ from 0.25 to 0.5 barely moves the dashed lines: the general equilibrium feedback through A absorbs the level shift. Higher $\phi_{\min}$ scales all abilities proportionally, aggregate H rises, A adjusts, and the technology ratios h_i/A are largely unchanged. Panel (b) shows the output consequences: the representative agent predicts that higher ϕ raises the output level, while the heterogeneous model output departs from it. Education raises steady-state output in both the representative agent and the heterogeneous economy, but it cannot alter the automation share in either. The general equilibrium feedback through A absorbs the level shift in both cases.

While education levels leave the automation share unchanged, the *shape* of the ability distribution does not. Panels (c)-(d) in Figure 5 holds the Pareto mean constant at the base-



(a) Automation share, β_t

(b) Output, $\ln(Y_t)$

(c) Automation share, β_t

(d) Output, $\ln(Y_t)$

Figure 5: Representative agent versus heterogeneous model under two education levels. Solid lines: representative agent with ϕ set to the Pareto mean of $\phi_{\min} \in \{0.25, 0.5\}$. Dashed lines: heterogeneous model, $N = 10{,}000$. Mediocrity regime ($\rho + \theta = 0.8$).

line value while varying the shape parameter α. A concentrated distribution ($\alpha = 4$) clusters abilities near a higher floor; a dispersed distribution ($\alpha = 1.5$) generates a heavy right tail but requires a much lower floor to maintain the same mean. The general equilibrium feedback through A neutralizes proportional scaling but cannot absorb changes in the relative distribution.

Panel (c) of Figure 5 shows that the concentrated economy has a lower steady-state $\mathrm{E}[\beta_i]$ than the dispersed economy. The mechanism is the saturation of $\beta(\cdot)$ at low skill: workers near the floor of the ability distribution are already at $\beta_i \approx 1$, so lowering the floor further changes nothing for them. The Pareto distribution places most mass near $\phi_{\min}$, and maintaining a constant mean while increasing dispersion requires lowering $\phi_{\min}$. The result is more workers trapped at $\beta_i \approx 1$. The heavy right tail of the dispersed distribution does produce a few high-ability workers who resist displacement, but they are too few to offset the mass at the bottom. Panel (d) complicates this finding: the dispersed economy's aggregate output lies far *above* the concentrated economy's, despite identical mean ability. The few high-ability workers in the right tail are productive enough to drive aggregate output, even as the mass of workers remains nearly fully automated. The representative agent understates this tail-driven output gain. The result is a tradeoff: concentration lowers displacement but at significant cost to aggregate output.

4.2 Augmentation Design

If education operates within a regime, the question is which instrument can change the regime itself. As explained in Section 2.5, the answer lies in θ, the degree to which AI enhances the learning process rather than merely displacing it. Unlike ϕ, the parameter θ enters the bifurcation condition directly. The comparative dynamics in Figure 2 already illustrate this: raising θ from 0.3 to 0.5 to 0.7 moves the economy from mediocrity through balanced growth to superagency. Changes in θ near the threshold produce qualitatively different long-run outcomes.

What does raising θ mean in practice? The parameter governs whether AI systems are designed to enhance human learning or to replace human judgment. A coding assistant that explains its suggestions and forces the programmer to engage with the logic raises θ: the human accumulates skill faster because AI augments the learning process. Autor (2024) argues for designing systems that augment workers rather than replace them. In the language of the model, this is a call to raise θ.

Augmentation design is the only instrument that changes the regime. But it is also the

hardest to implement. Raising θ requires shaping the design choices of AI developers—mandating or incentivizing systems that explain their reasoning, preserve user engagement, and maintain the cognitive demands through which learning occurs.

4.3 Discussion

Education operates on levels, augmentation on regimes. The two instruments are complementary but asymmetric: education is feasible but cannot escape the trap, while augmentation design can change the regime but is difficult to mandate. Within the education margin, the distribution of ability matters more than its average level: the general equilibrium feedback through A absorbs uniform increases in $\phi_{\min}$, but the shape of the distribution creates a tradeoff. Concentration—raising the floor relative to the mean—lowers aggregate displacement by shifting more workers above the automation threshold, but at significant cost to aggregate output. The few high-ability workers in the tail of a dispersed distribution drive output even as most workers are displaced. The binding constraint remains the qualitative design of AI systems, not the quantity of human capital investment.

The heterogeneous agent results from Section 3 sharpen the policy challenge. Aggregate statistics mask the distributional reality: the mean automation share understates the displacement experienced by the typical worker, and aggregate output overstates the economy's resilience by weighting toward the few high-ability workers who resist displacement.

The current model has limitations in terms of policy evaluation. First, the output metric abstracts from welfare; a full welfare analysis would require a consumption-savings decision and an explicit social welfare function. Second, the model treats θ as a policy tool, but in practice the augmentation-versus-automation design choice reflects the incentives of AI developers and the preferences of adopting firms. Hemous and Olsen (2022) and Acemoglu and Restrepo (2018) endogenize the direction of technical change between automation and augmentation. Embedding their mechanisms into the present framework would move the analysis from identifying the right policy tool to characterizing how it is determined.

5. Conclusion

This paper formalizes the dynamic feedback between automation and human capital accumulation in an endogenous growth framework where human capital and AI co-evolve. The model reveals a bifurcation governed by whether AI augments human learning enough to generate increasing returns in the reproducible factors.

The model's ironies reinforce each other. Automation erodes the practice through which workers build skills and, with it, the human oversight that AI development demands. Task complementarity amplifies the feedback by making the scarcity premium that should protect labor an incentive to automate.

Below the bifurcation threshold, the economy settles into a *mediocrity trap*: finite human capital, finite AI, incomplete automation, and no endogenous growth. At the knife-edge, constant returns in reproducible factors sustain *endogenous balanced growth*. Above it, the economy either faces explosive co-growth under *superagency* or goes into an *obsolescence spiral* in which automation and atrophy reinforce each other until both stocks collapse. The same AI systems that enable *superagency* also make collapse catastrophic depending on where the economy starts.

Introducing heterogeneous learning ability produces a heavily skewed ergodic distribution. Even when the economy grows, most workers cluster near obsolescence while a Pareto tail of high-ability workers sustains the economy.

I draw some policy implications from the model. Education investment raises the level of output but cannot change the growth regime. Augmentation design—building AI systems that enhance human learning rather than replace human judgment—is the only instrument that helps cross the bifurcation threshold. The binding constraint on long-run growth is not the economy's capacity to automate but its capacity to sustain the human capital that automation demands.

Embedding the ironies of automation formalized here in a framework where forward-looking agents choose how much to invest in their own skills given the automation trajectory, and where AI developers choose how much to invest in human-augmenting AI is a natural and promising avenue for future research.

References

Acemoglu, D. (2002). Directed Technical Change. *Review of Economic Studies*, 69(4):781–809.

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.

Acemoglu, D. and Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica*, 90(5):1973–2016.

Aghion, P., Jones, B. F., and Jones, C. I. (2019). Artificial intelligence and economic growth. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, pages 237–282. University of Chicago Press.

Agrawal, A., McHale, J., and Oettl, A. (2026). Enhancing Worker Productivity Without Automating Tasks: A Different Approach to AI and the Task-Based Model. Technical Report w34781, National Bureau of Economic Research, Cambridge, MA.

Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173.

Autor, D. (2024). Applying AI to Rebuild Middle Class Jobs. Technical Report w32140, National Bureau of Economic Research, Cambridge, MA.

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6):775–779.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, 75(4):352–365.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at work. NBER Working Paper No. 31161.

Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company, New York.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper No. 24-013.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Frankel, M. (1962). The production function in allocation and growth: A synthesis. *The American Economic Review*, 52(5):996–1022.

Freund, L. B. and Mann, L. F. (2026). Job Transformation, Specialization, and the Labor Market Effects of AI.

Hemous, D. and Olsen, M. (2022). The rise of the machines: Automation, horizontal innovation, and income inequality. *American Economic Journal: Macroeconomics*, 14(1):179–223.

Hoffman, R. (2025). *Superagency: How AI Will Amplify Human Capability*. Currency, New York.

Ide, E. and Talamàs, E. (2025). Artificial Intelligence in the Knowledge Economy. *Journal of Political Economy*, 133(12):3762–3800.

Jones, C. I. (1995). R&D-based models of economic growth. *Journal of Political Economy*, 103(4):759–784.

Jones, C. I. (2011). Intermediate goods and weak links in the theory of economic development. *American Economic Journal: Macroeconomics*, 3(2):1–28.

Jones, C. I. (2024). The A.I. dilemma: Growth versus existential risk. Working Paper, Stanford University.

Jones, C. I. and Tonetti, C. (2026). Past automation and future a.i.: How weak links tame the growth explosion. Working Paper, Stanford University.

Korinek, A. and Juelfs, M. (2025). Preparing for the (non-existent?) future of work. NBER Working Paper No. 31966.

Lucas, Jr., R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42.

Nordhaus, W. D. (2021). Are we approaching an economic singularity? Information technology and the future of economic growth. *American Economic Journal: Macroeconomics*, 13(1):299–332.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Prettner, K. and Strulik, H. (2020). Innovation, automation, and inequality: Policy challenges in the race against the machine. *Journal of Monetary Economics*, 116:249–265.

Rebelo, S. (1991). Long-run policy analysis and long-run growth. *Journal of Political Economy*, 99(3):500–521.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, 94(5):1002–1037.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

Trammell, P. and Korinek, A. (2024). Economic growth under transformative AI. Working Paper.

Yotzov, I., Barrero, J. M., Bloom, N., Bunn, P., Davis, S., Foster, K., Jalca, A., Meyer, B., Mizen, P., Navarrete, M., Smietanka, P., Thwaites, G., and Wang, B. Z. (2026). Firm Data on AI. Technical Report w34836, National Bureau of Economic Research, Cambridge, MA.

Zeira, J. (1998). Workers, machines, and economic growth. *Quarterly Journal of Economics*, 113(4):1091–1117.

Appendix

A. Derivation of the Automation Share

This appendix derives the closed-form expression for the individual automation share $\beta_{i,t} = 1/(1+(h_{i,t}/A_t)^{\zeta})$ from a standard Eaton and Kortum (2002) model of comparative advantage applied to task allocation within each worker's production unit.

A.1 Setup

For each task $j \in [0,1]$, the productivity of AI and human labor are $A_t z_{A,j}$ and $h_{i,t} z_{H,j}$, respectively, where $z_{A,j}$ and $z_{H,j}$ are independent draws from Fréchet distributions with common shape parameter $\psi > 0$ and scale parameter normalized to one. The unit costs of performing task j in worker i's production unit are

$$C_{A,j} = \frac{r_{i,t}}{A_t \, z_{A,j}}, \qquad C_{H,j} = \frac{w_{i,t}}{h_{i,t} \, z_{H,j}}, \tag{29}$$

where $r_{i,t}$ and $w_{i,t}$ are the marginal products of AI and human capital from equations (23) and (22). In worker i's production unit, a cost-minimizing assignment allocates each task to the factor with lower unit cost.

A.2 Automation Probability

Task j is automated whenever AI is the least-cost provider:

$$C_{A,j} < C_{H,j} \iff \frac{r_{i,t}}{A_t \, z_{A,j}} < \frac{w_{i,t}}{h_{i,t} \, z_{H,j}} \iff \frac{z_{A,j}}{z_{H,j}} > \frac{r_{i,t}}{w_{i,t}} \cdot \frac{h_{i,t}}{A_t}. \tag{30}$$

By the properties of the Fréchet distribution, the probability that AI is the least-cost provider across the continuum of tasks is

$$\beta_{i,t} = \frac{(A_t/r_{i,t})^{\psi}}{(A_t/r_{i,t})^{\psi} + (h_{i,t}/w_{i,t})^{\psi}}. \tag{31}$$

Dividing numerator and denominator by $(A_t/r_{i,t})^{\psi}$:

$$\beta_{i,t} = \frac{1}{1 + \left(\dfrac{h_{i,t}}{w_{i,t}} \cdot \dfrac{r_{i,t}}{A_t}\right)^{\psi}} = \frac{1}{1 + \left(\dfrac{r_{i,t}}{w_{i,t}}\right)^{\psi} \left(\dfrac{h_{i,t}}{A_t}\right)^{\psi}}. \tag{32}$$

A.3 Endogenous Price Feedback

Define the intermediate variable

$$\kappa_{i,t} \equiv \left(\frac{r_{i,t}}{w_{i,t}}\right)^{\psi}, \tag{33}$$

so that $\beta_{i,t} = 1/\left[1 + \kappa_{i,t}(h_{i,t}/A_t)^{\psi}\right]$. From equations (22) and (23), the ratio of marginal products within worker i's production unit is:

$$\frac{r_{i,t}}{w_{i,t}} = \frac{\beta_{i,t}}{1-\beta_{i,t}} \left(\frac{h_{i,t}}{A_t}\right)^{1/\sigma}. \tag{34}$$

Substituting into (33):

$$\kappa_{i,t} = \left(\frac{\beta_{i,t}}{1-\beta_{i,t}}\right)^{\psi} \left(\frac{h_{i,t}}{A_t}\right)^{\psi/\sigma}. \tag{35}$$

A.4 Solving for $\beta_{i,t}$

Substituting (35) into (32) and rearranging for the odds ratio $(1-\beta_{i,t})/\beta_{i,t}$:

$$\frac{1-\beta_{i,t}}{\beta_{i,t}} = \left(\frac{\beta_{i,t}}{1-\beta_{i,t}}\right)^{\psi} \left(\frac{h_{i,t}}{A_t}\right)^{\psi(1+1/\sigma)}. \tag{36}$$

Multiplying both sides by $\left[(1-\beta_{i,t})/\beta_{i,t}\right]^{\psi}$ collects all $\beta_{i,t}$ terms on the left:

$$\left(\frac{1-\beta_{i,t}}{\beta_{i,t}}\right)^{1+\psi} = \left(\frac{h_{i,t}}{A_t}\right)^{\psi(\sigma+1)/\sigma}, \tag{37}$$

Taking the $(1+\psi)$-th root:

$$\frac{1-\beta_{i,t}}{\beta_{i,t}} = \left(\frac{h_{i,t}}{A_t}\right)^{\psi(\sigma+1)/[\sigma(1+\psi)]} \equiv \left(\frac{h_{i,t}}{A_t}\right)^{\zeta}, \tag{38}$$

with the composite elasticity

$$\zeta \equiv \frac{\psi(\sigma+1)}{\sigma(1+\psi)}. \tag{39}$$

Solving for the automation share:

$$\beta_{i,t} = \frac{1}{1+(h_{i,t}/A_t)^{\zeta}}, \qquad 1-\beta_{i,t} = \frac{(h_{i,t}/A_t)^{\zeta}}{1+(h_{i,t}/A_t)^{\zeta}}. \tag{40}$$

When all workers are identical ($h_{i,t} = H_t$ for all i), this reduces to $\beta_t = 1/(1+x_t^{\zeta})$ with $x_t \equiv H_t/A_t$.

B. Proofs

B.1 Proof of Proposition 1

Proof. I proceed in three steps.

(i) Nullcline ratio.

Setting $dA = 0$ in (18) and writing $x = H/A$ yields $\lambda\, x^{\omega} = \gamma$. Since $\omega \in (0,1)$, the function $g(x) \equiv \lambda\, x^{\omega}$ is strictly increasing on $(0,\infty)$ with $g(0) = 0$ and $g(x) \to \infty$. There is therefore a unique solution:

$$x^* = \left(\frac{\gamma}{\lambda}\right)^{1/\omega}, \tag{41}$$

establishing (19).

(ii-iii) Interior steady state and stability ($\rho + \theta \neq 1$).

Setting $dH = 0$ in (17) and substituting $A^* = H^*/x^*$ gives

$$(H^*)^{\rho+\theta-1} = \frac{\delta_h}{\phi} \cdot \frac{1 + (x^*)^{\zeta}}{(x^*)^{\zeta-\theta}}. \tag{42}$$

Existence and uniqueness. The right-hand side is strictly positive for $x^* > 0$ and the exponent is nonzero, so this equation has a unique positive solution H^*. Together with $A^* = H^*/x^*$, the interior steady state exists and is unique.

Stability: proportional perturbations. Scale both stocks by a common factor: $(H, A) \to (1+\varepsilon)(H^*, A^*)$. The technology ratio $x = H/A$ is unchanged, so β does not move and $g_A = \lambda x^{\omega} - \gamma$ is unaffected. The only effect is on human capital growth through $H^{\rho+\theta-1}$. When $\rho + \theta < 1$, higher H *reduces* g_H: the system is self-correcting and the interior steady state attracts. When $\rho + \theta > 1$, higher H *raises* g_H: small advantages compound and the interior steady state is a saddle. The formal Jacobian analysis confirms this: $\det(J) = -\omega\gamma\delta_h\,(\rho + \theta - 1)$, whose sign is entirely determined by whether $\rho + \theta$ exceeds unity.

Stability: compositional perturbations. Raise H while holding A fixed. Two competing effects arise: (i) the practice share $(1-\beta)$ increases, accelerating human capital accumulation (positive feedback); (ii) higher H also raises AI investment $\lambda H^{\omega} A^{1-\omega}$, which accelerates AI growth and restores the ratio. For the mediocrity trap ($\rho + \theta < 1$), stability requires the AI adjustment channel to dominate: $\omega\gamma > \delta_h(\rho + \zeta\beta^* - 1)$. Formally,

$\mathrm{tr}(J) = \delta_h(\rho + \zeta\beta^* - 1) - \omega\gamma < 0$. When $\rho + \theta > 1$, the saddle classification follows from $\det(J) < 0$ regardless of the trace.

Boundary. At $(H, A) = (0, 0)$, both drift terms vanish. Near $H = 0$, $(1 - \beta(H/A)) \sim (H/A)^\zeta \to 0$: workers have no tasks to practice on. Depreciation dominates accumulation, and H spirals to zero. The result is verified numerically in Section 2.6.

(iv) Endogenous balanced growth ($\rho + \theta = 1$).

When $\rho + \theta = 1$, equation (42) reduces to $1 =$ RHS, which generically has no solution. No interior steady state in levels exists.

At $\rho + \theta = 1$, the factor $H^{\rho+\theta-1} = 1$ drops out, so both growth rates depend only on x:

$$g_H(x) = \phi(1 - \beta(x))\, x^{-\theta} - \delta_h, \tag{43}$$

$$g_A(x) = \lambda\, x^{\omega} - \gamma. \tag{44}$$

The ratio evolves as

$$\frac{\dot{x}}{x} = g_H(x) - g_A(x) \equiv \Phi(x). \tag{45}$$

Existence and uniqueness. Since $\Phi(0^+) = \gamma - \delta_h > 0$ and $\Phi(x) \to -\infty$ as $x \to \infty$, continuity gives at least one zero $x^{**} > 0$. Uniqueness follows because g_A is strictly increasing while g_H is hump-shaped and eventually decreasing; under the regularity condition that Φ crosses zero exactly once, x^{**} is unique.

Stability. At x^{**}, $\Phi'(x^{**}) = g_H'(x^{**}) - g_A'(x^{**}) < 0$, since $g_A' > 0$ and g_H is locally decreasing near x^{**}. The balanced growth ratio is locally asymptotically stable.

Finally, along the balanced growth path, both stocks grow at

$$g^* = g_A(x^{**}) = \lambda(x^{**})^{\omega} - \gamma = g_H(x^{**}). \tag{46}$$