

Abundant Agents, Scarce Organizations

Implementation Capacity and the Economics of AI Adoption

Jorge Menéndez-Pidal

Project Frontier, Madrid

ABSTRACT. AI agents can perform tasks, but unlike workers they can be copied at the price of compute. This paper asks what that does to the economics of production. In a task-based model, a firm can let agents perform a task only after implementing it: integrating the agent, redesigning the workflow and preparing the data. Implementation uses a distinct kind of labour and is a fixed cost per task, independent of scale. Five results follow. Automation depends on market size relative to organisational frictions, so AI-native entrants can out-automate larger incumbents. Cheaper agents raise the demand for implementation labour: the technology substitutes for the workers who perform tasks and complements those who deploy it. When implementation capacity is inelastic, a fall in the price of AI raises the implementation wage rather than the number of automated tasks, and its effect on unit costs is bounded by the current automation share, the Hulten term that existing macroeconomic estimates take as given. The same bottleneck shields production workers: displacement is triggered by the expansion of implementation capacity, not by the falling price of AI. If AI lowers its own implementation cost, the bottleneck persists only while that rate stays below a threshold the paper derives. Read through the model, the modest ten-year productivity estimate of Acemoglu (2025) presumes implementation capacity growing at roughly 8–22% a year. Agents are abundant; what is scarce is the organisation able to deploy them.

Preliminary draft (v0.2). This is a theory note. Sections 3–6 develop the model, with all propositions proved in Appendix A and verified symbolically and numerically in the replication code; Section 7 illustrates them with stylised parameters that are not estimates; Section 8 sets out testable predictions. Comments welcome; please do not cite without permission.

Keywords: AI agents, automation, implementation, organisational capital, factor demand, technology adoption, productivity. **JEL:** D24, E23, J23, L23, O33.

1 Introduction

The price of machine intelligence is falling faster than that of almost any input in economic history. The cost of a given level of model capability has fallen by orders of magnitude in a few years, and AI agents, systems that execute multi-step tasks with limited supervision, are now sold by the unit of compute. Yet the measured economic footprint of the technology remains small. Only a few per cent of US firms reported using AI to produce goods or services in 2023–24 (Bonney et al., 2024). Labour-market effects in the first years of generative AI are close to zero (Humlum and Vestergaard, 2025). Workers adopted the technology faster than the organisations that employ them (Bick et al., 2024). At the same time, firms are hiring a new class of workers, forward-deployed engineers, AI engineers and implementation consultants, whose job is not to perform tasks but to make it possible for agents to perform them.

This paper argues that these facts have a common explanation and builds a model to make it precise. The starting question is whether AI agents should be treated as a new factor of production. They perform tasks, as labour does, but they are owned and bought, as capital is. They are also reproducible: once a model exists, an additional instance costs only the compute it consumes. The answer the model gives is that the agent is not the scarce factor. Because agents can be bought in whatever quantity is needed at price p_A , they earn no scarcity rent. The factor that becomes scarce, and that captures rents as AI becomes cheaper, is the capacity to implement them.

The model is a task-based model of production in the tradition of Zeira (1998) and Acemoglu and Restrepo (2018), with one addition. Before agents can perform a task inside a firm, the task has to be implemented: the agent integrated with the firm’s systems, the workflow redesigned, the data prepared and the output evaluated. Implementation uses a distinct input, implementation labour, and it is a fixed cost per task. Once a workflow is built, agents replicate at constant cost, whatever the scale of the firm. Agents are thus non-rival in design and rival in deployment, and implementation is a fixed cost in organisation. Firms have increasing returns that come from the organisation, not from the code.

Five results follow.

Automation (Proposition 1). A firm automates more tasks when agents are cheaper, when its market is larger, and when implementation is cheaper. Market size and implementation costs enter only through their ratio, the firm’s market size per unit of organisational friction. A start-up without legacy systems can therefore automate more than an incumbent with a far larger market. The advantage of AI-native firms is organisational rather than technological: both buy the same model at the same price.

Complementarity (Proposition 2). Cheaper agents raise the firm’s demand for implementation labour. The same technology substitutes for the workers who perform automated tasks and complements the workers who deploy it. This is a Jevons effect operating on the complement rather than on the input itself. It explains why firms hire more people to implement AI as AI gets cheaper.

The bottleneck (Proposition 3). When the supply of implementation labour is inelastic, a fall in the price of agents does not change the set of automated tasks. It raises the implementation wage instead, which absorbs the value of the tasks that would otherwise be automated. The

effect of cheaper AI on unit costs is then exactly the current automation share. This is the Hulten term on which macroeconomic estimates of the impact of AI are built (Hulten, 1978; Acemoglu, 2025). Here, the share is endogenous and its growth is governed by implementation capacity. The gap between what the technology can do and what the economy gets from it is a property of organisations, not of models.

Labour (Proposition 4). The bottleneck also shields production workers. With implementation capacity fixed, cheaper agents only reduce costs on tasks already automated, so the firm expands and hires more production labour. Displacement requires that the set of automated tasks grow, and that in turn requires implementation capacity. Whether a fall in the price of AI raises or lowers employment depends on the elasticity of supply of implementation labour, and it switches sign at a threshold. The prediction concerns timing: small or positive employment effects at first, followed by displacement as implementation capacity accumulates.

Self-implementation (Proposition 5). If better AI lowers its own implementation requirement, the extensive margin opens even with fixed capacity, at the rate at which the requirement falls. The scarcity rent of implementation labour survives only while that rate stays below a threshold that depends on the automation share and on how fast cost savings decline across tasks. Above it, the bottleneck dissolves. Section 7.1 uses the same logic in reverse: taken through the model, the ten-year productivity estimate of Acemoglu (2025) presumes implementation capacity growing by roughly 8–22% a year, depending on the share of tasks automated today.

The paper is a theory paper. All propositions are proved in Appendix A and checked symbolically and numerically in the replication code. Section 7 illustrates them with stylised parameters that are not estimates. Section 8 sets out testable predictions. Section 9 derives the self-implementation result and discusses further extensions.

This note belongs to a series. RN 1 shows that variable inference costs erode the economies of scale of software firms. This paper shows that those economies reappear one level down, in the organisations that deploy the software: scale economies migrate from code to organisation. RN 2 studies how the gains from AI are distributed across workers of different ability. This paper studies how they are distributed across factors.

2 Related literature

Tasks and automation. The model is built on the task framework (Zeira, 1998; Autor et al., 2003; Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018, 2019, 2022). In that framework, the set of automated tasks is pinned down by relative factor costs, and automation has a displacement effect and a productivity effect. The departure here is that the extensive margin of automation is rationed by a scarce input with a fixed cost per task. This changes three conclusions. The price of a complement rises when the price of the automating technology falls. The aggregate effect of cheaper AI is bounded by the current automation share. And the sign of the employment effect depends on the supply of the complement, not on the price of the technology. Hémous and Olsen (2022) also model a cost of automation, but as directed innovation at the aggregate level rather than as organisational work inside the firm.

AI and the macroeconomy. Aghion et al. (2019) emphasise that growth may be constrained by inputs that are essential but hard to improve, a Baumol logic of which implementation capacity

is an organisational instance. Acemoglu (2025) uses Hulten's theorem (Hulten, 1978) to bound the macroeconomic effect of AI by the share of tasks affected times the cost saving, taking that share as given. Proposition 3 endogenises it. Scenario analyses of transformative AI (Korinek and Suh, 2024; Trammell and Korinek, 2023) focus on substitutability between compute and labour. Restrepo (2025) distinguishes between “bottleneck” work, essential for growth, and supplementary work. The word is shared but the concept is different: our bottleneck is the capacity to deploy agents, not a class of tasks.

General purpose technologies and organisational complements. The closest antecedents are empirical and historical. Electricity raised factory productivity only after plants were redesigned around it (David, 1990). General purpose technologies require co-invention in the sectors that adopt them (Bresnahan and Trajtenberg, 1995). Information technology paid off in firms that reorganised and hired skilled workers (Brynjolfsson and Hitt, 2000; Bresnahan et al., 2002), and in firms with better management practices (Bloom et al., 2012). Intangible complements that are not measured produce a productivity J-curve (Brynjolfsson et al., 2019, 2021). Milgrom and Roberts (1990) formalise complementarities through supermodularity, the method used for Proposition 1 (Topkis, 1978). This literature shows that complements matter. The contribution here is a price-theoretic mechanism that links the falling price of AI to the rising price of its complement and to a bound on its aggregate effect.

Organisation capital and absorptive capacity. Organisation capital (Prescott and Visscher, 1980; Atkeson and Kehoe, 2005) and absorptive capacity (Cohen and Levinthal, 1990) are stock concepts. Here implementation is a flow of labour with a market price. The stock version, in which implementation capacity accumulates through experience, is the dynamic extension in Section 9. Corrado et al. (2009) discuss the measurement of these intangibles.

AI in organisations. Agrawal et al. (2018, 2022) argue informally that the returns to AI require “system” rather than “point” solutions, that is, redesign of the surrounding workflow. This paper formalises that argument. Knowledge-hierarchy models with AI (Garicano, 2000; Ide and Talamàs, 2025) study how AI reshapes the organisation of expertise. The present model is complementary: it treats the organisation as the scarce input in deploying AI.

Diffusion and evidence. Adoption lags are a long-standing theme (Griliches, 1957; Comin and Hobijn, 2010). Recent evidence on AI adoption fits the mechanism. Adoption is concentrated in large and young firms (McElheran et al., 2024). Firms that adopt report organisational changes and few employment reductions (Bonney et al., 2024). Employment effects are small (Humlum and Vestergaard, 2025). Task-level productivity gains are large (Brynjolfsson et al., 2025; Eloundou et al., 2024).

3 The model

3.1 Environment

Demand. A firm faces isoelastic demand $y = D p^{-\varepsilon}$ with $\varepsilon > 1$, as in monopolistic competition (Dixit and Stiglitz, 1977). D measures the size of its market.

Tasks. Output is produced from a unit continuum of tasks,

$$\ln y = \int_0^1 \ln y(i) di. \quad (1)$$

Task i can be performed by production labour, $y(i) = \gamma_L(i) l(i)$, at wage w . If the task has been implemented, it can also be performed by agents, $y(i) = \gamma_A(i) a(i)$, at price p_A per unit of compute. Within a task, labour and agents are perfect substitutes. Comparative advantage $\psi(i) \equiv \gamma_A(i)/\gamma_L(i)$ is continuously differentiable and strictly decreasing, so tasks are ordered from most to least automatable.

Agents. Model weights are non-rival: a model, once trained, serves any number of users. Each deployed instance is rival, because it consumes compute. The firm buys agent services at price p_A with perfectly elastic supply, so agents are reproducible. One can think of $p_A = (1 + \mu)c_A$, a markup over the marginal cost of inference charged by the model provider to recover the fixed cost of training. In the baseline p_A is exogenous. Write $\theta \equiv -\ln p_A$, so that an increase in θ means cheaper AI.

Implementation. Agents cannot perform a task inside the firm until the task has been implemented: the agent integrated with the firm's systems and data, the workflow redesigned around it, and its output evaluated. Implementing a task requires η units of implementation labour H , paid wage v . The requirement is per task, incurred once and independent of how much of the task the firm performs. Once a workflow exists, agents replicate at constant unit cost. Implementation is therefore a fixed cost in organisation, while production with agents has constant marginal cost in compute.

Markets. In Sections 4 and 6, the production wage w is given: the firm, or the sector, is small in the market for production labour. Section 5 closes the market for implementation labour.

3.2 Costs

Let

$$s(i) \equiv \ln \frac{\psi(i) w}{p_A} \quad (2)$$

denote the log cost saving from performing task i with agents rather than labour. It is strictly decreasing in i and increasing one-for-one in θ . Let $\bar{I}(\theta)$ be the task at which $s = 0$ (or 1 if $s > 0$ everywhere). Implementing tasks beyond $\bar{I}$ yields nothing, because agents are more expensive there, and since s is decreasing an optimal firm implements an interval $[0, I]$ with $I \leq \bar{I}$. Write $S(I) \equiv \int_0^I s(i) di$ for $I \leq \bar{I}$.

Lemma 1 (Costs and factor shares). *Given the implemented set $[0, I]$, the firm's unit cost is $c(I) = c_0 e^{-S(I)}$, where $c_0 = \exp \int_0^1 \ln(w/\gamma_L)$ is the unit cost with no automation. Agents account for a share I of variable cost and production labour for $1 - I$. Total variable cost, revenue R and operating profit π satisfy*

$$TC(I) = \frac{\varepsilon - 1}{\varepsilon} R = (\varepsilon - 1)\pi = (\varepsilon - 1)\kappa D c_0^{1-\varepsilon} e^{(\varepsilon-1)S(I)}, \quad \kappa \equiv \frac{(\varepsilon - 1)^{\varepsilon-1}}{\varepsilon^\varepsilon}. \quad (3)$$

The firm chooses I to maximise

$$\Pi(I) = \frac{TC(I)}{\varepsilon - 1} - v \eta I. \quad (4)$$

The first term is operating profit, which rises with automation through lower costs and through the larger scale that lower prices bring. The second is the cost of implementation.

Assumption 1. *At the optimum, $(\varepsilon - 1)s(I)^2 + s'(I) < 0$.*

Assumption 1 is the second-order condition. It requires that comparative advantage decline faster across tasks than the scale feedback from automation to profit rises. Propositions 1 and 2 do not use it. Propositions 3 and 4 use it to obtain closed forms. Interior optima satisfy

$$TC(I^*) \cdot s(I^*) = v \eta. \quad (5)$$

Because each task accounts for an equal share of variable cost, $TC \cdot s$ is the cost bill of the marginal task times the log saving from automating it. The firm implements the marginal task when that saving covers the cost of implementation.

4 Automation and the demand for implementation

Proposition 1 (Automation). *The largest optimal automation set I^* is weakly increasing in θ (decreasing in p_A) and in D , and weakly decreasing in v and η . Market size and implementation costs affect I^* only through the ratio $D/(v\eta)$.*

The proof uses monotone comparative statics. A lower price of agents raises the marginal value of automation for two reasons: each automated task is cheaper, and a lower unit cost expands the firm's scale, which raises the value of every cost saving. No concavity is needed.

The ratio $D/(v\eta)$ is the firm's market size per unit of organisational cost. It collapses two very different sources of advantage into one number.

Corollary 1 (AI-native advantage). *Consider an entrant with (D_e, η_e) and an incumbent with (D_i, η_i) , facing the same p_A , w and v . The entrant automates more tasks if and only if $D_e/\eta_e > D_i/\eta_i$.*

An incumbent with ten times the market of an entrant automates fewer tasks if its implementation requirement is more than ten times larger, because of legacy systems, fragmented data or processes that must be renegotiated. The entrant has no technological advantage: both buy the same model at the same price. Its advantage is that it has less organisation to change.

Proposition 2 (Jevons for complements). *The firm's demand for implementation labour, $H^d = \eta I^*$, is weakly increasing in θ : cheaper agents raise the demand for implementation labour. Under Assumption 1, at an interior optimum with v fixed,*

$$\left. \frac{dI^*}{d\theta} \right|_v = \frac{(\varepsilon - 1)I^* + 1/s}{Q} > 0, \quad Q \equiv -\left[(\varepsilon - 1)s + \frac{s'}{s} \right] > 0, \quad (6)$$

with s and s' evaluated at I^* .

The same technology is a substitute for production labour on the tasks it automates and a complement to the labour that deploys it. The two relations follow from the same fact. Agents enter production with a constant marginal cost, while implementation is a fixed cost that unlocks their use. When agents get cheaper, each implemented task is worth more and the firm implements more of them. This is Jevons (1865) applied to a complement: cheaper agents raise the demand for the input that makes them usable.

The numerator of (6) has two terms. $(\varepsilon - 1)I^*$ is the scale effect: lower costs on already-automated tasks expand output and raise the value of every task. $1/s$ is the direct effect: cheaper agents raise the saving on the marginal task, proportionally more when that saving is small.

5 The organisational bottleneck

Implementation labour is supplied by people who know both the technology and how organisations work: forward-deployed engineers, AI engineers, solution architects. Its supply is limited in the short run and expands as people are trained and gain experience. Consider a unit mass of identical firms and a supply of implementation labour with constant elasticity $\zeta \geq 0$:

$$\eta I = \bar{H} v^\zeta. \quad (7)$$

Because demand for H falls with v (Proposition 1), the market has a unique equilibrium (I, v) .

Proposition 3 (The bottleneck). *Under Assumption 1, at an interior equilibrium:*

(a) *The automation set responds to cheaper agents according to*

$$\frac{dI}{d\theta} = \frac{(\varepsilon - 1)I + 1/s}{Q + 1/(\zeta I)}, \quad (8)$$

which is increasing in ζ , equal to zero when $\zeta = 0$, and converges to (6) as $\zeta \rightarrow \infty$.

(b) *The implementation wage rises when agents get cheaper:*

$$\frac{d \ln v}{d\theta} = \frac{(\varepsilon - 1)I + 1/s}{1 + \zeta IQ} > 0, \quad (9)$$

and the response is largest when supply is least elastic.

(c) *The elasticity of the unit cost with respect to the price of agents is*

$$\frac{d \ln c}{d \ln p_A} = I + s \frac{dI}{d\theta} \geq I, \quad (10)$$

with equality if and only if $\zeta = 0$.

Part (c) is the central result. By Hulten's theorem, the first-order effect of a fall in an input price on unit cost equals the cost share of that input. Here the share of agents in variable cost is I , the automation share. When implementation capacity is fixed, this first-order effect is the whole effect. The extensive margin, automating tasks that are not yet automated, is closed. A fall in the price of AI then lowers costs only on tasks already automated, and today that share is small.

This is the economic content of the gap between technological and economic capability. A model that becomes ten times cheaper and twice as capable does not raise productivity in proportion if the organisations using it cannot implement new workflows. In Acemoglu (2025), the macroeconomic effect of AI is bounded by the share of tasks affected, which is taken as given. Proposition 3 says what governs that share: the supply of implementation capacity.

Part (b) says where the value of the closed margin goes. With $\zeta = 0$, the value of every task that would be automated at the lower price of agents accrues to implementation labour as a rent. The workers who deploy AI gain most from cheaper AI, and the rise in their wage measures the bottleneck. Cheaper AI produces the shortage of forward-deployed engineers, rather than relieving it.

The bound in part (c) applies to the short run. As ζ rises, through training, learning by doing or the entry of implementation firms, the extensive margin opens and the response of costs to the price of AI grows towards its frictionless value. The model thus produces a J-curve shaped by organisational capacity, not by unmeasured intangibles alone (Brynjolfsson et al., 2021).

6 Production labour

Production labour is employed on the tasks that are not automated. By Lemma 1, the production wage bill is a share $1 - I$ of variable cost, so employment is $L = (1 - I)TC/w$.

Proposition 4 (The bottleneck shields labour). *Under Assumption 1, at an interior equilibrium of the implementation market,*

$$\frac{d \ln L}{d \theta} = (\varepsilon - 1)I + \left[(\varepsilon - 1)s - \frac{1}{1 - I} \right] \frac{dI}{d \theta}, \quad (11)$$

with $dI/d\theta$ given by (8).

- (a) If $\zeta = 0$, then $d \ln L/d\theta = (\varepsilon - 1)I > 0$: cheaper agents raise employment in production.
- (b) If $(\varepsilon - 1)s(1 - I) < 1$, the second term of (11) is negative and its magnitude increases in ζ . Comparing economies with the same initial equilibrium, there is at most one threshold $\hat{\zeta}$ such that cheaper agents raise production employment for $\zeta < \hat{\zeta}$ and lower it for $\zeta > \hat{\zeta}$.

The first term of (11) is a scale effect. Cheaper agents lower costs on automated tasks, the firm cuts its price and expands, and it needs more labour on the tasks that remain manual. The second term combines displacement, the loss of the newly automated tasks ($-1/(1 - I)$), with the scale effect of automating them $((\varepsilon - 1)s)$. Displacement requires that the automation set grow, and by Proposition 3 that requires implementation capacity.

The implication is about timing. Falling AI prices do not on their own displace workers. While implementation capacity is the binding constraint, they raise employment. Displacement comes when that capacity expands. A finding of small employment effects in the first years of a technology (Humlum and Vestergaard, 2025) is therefore not evidence that later effects will also be small. It is what the model predicts while organisations are the bottleneck.

The labour share of revenue is $(1 - I)(\varepsilon - 1)/\varepsilon$ and changes only through the extensive margin. It too is held in place by the bottleneck.

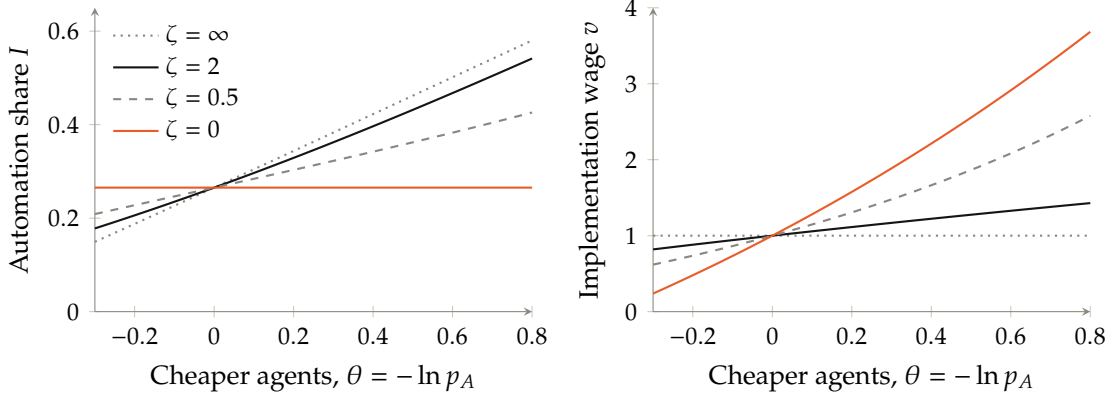


FIGURE 1. Automation share (left) and implementation wage (right) as the price of agents falls, for four elasticities of implementation supply. All economies share the equilibrium at $\theta = 0$.

7 Numerical illustration

This section illustrates the propositions with a linear benchmark for comparative advantage, $\ln \psi(i) = \alpha - \beta i$, so that $s(i) = m_0 + \theta - \beta i$. Parameters are $\varepsilon = 2$, $\beta = 3$, $m_0 = 1.2$ and $\eta = 1$. All values are illustrative and are not estimates. The supply of implementation labour is normalised so that $v = 1$ at $\theta = 0$ for every elasticity ζ . All economies therefore start from the same equilibrium, with automation share $I = 0.27$, and differ only in how implementation capacity responds. The horizontal axis is $\theta = -\ln p_A$. Its right end, $\theta = 0.8$, corresponds to a 55% fall in the price of agents.

Automation and the implementation wage. Figure 1 shows the automation share and the implementation wage as agents get cheaper. With elastic implementation supply, the 55% fall in p_A raises the automation share from 0.27 to 0.58. With $\zeta = 2$, it rises to 0.54. With fixed capacity ($\zeta = 0$), the share does not move, and the implementation wage rises to 3.68 times its initial level, against 1.43 times with $\zeta = 2$. The same technological shock shows up as automation in one economy and as implementation rents in another.

The Hulten bound. Figure 2 plots the elasticity of unit cost with respect to p_A at $\theta = 0$ against ζ . With fixed capacity it equals the automation share, 0.27, the Hulten term. It rises to 0.39 at $\zeta = 2$ and approaches 0.42 as supply becomes perfectly elastic. Over the full range of Figure 1, unit cost falls by 19% with fixed capacity and by 36% with elastic supply. Organisational capacity thus determines almost half of the productivity effect of the same fall in the price of AI.

Production labour. The right panel of Figure 2 plots the elasticity of production employment with respect to the fall in p_A , $d \ln L / d \theta$. It is 0.27 with fixed capacity, where cheaper agents expand the firm without displacing anyone. It falls with ζ , crosses zero at $\hat{\zeta} = 1.32$ and reaches -0.11 with elastic supply. The same price shock raises employment when organisations are the bottleneck and reduces it when they are not.

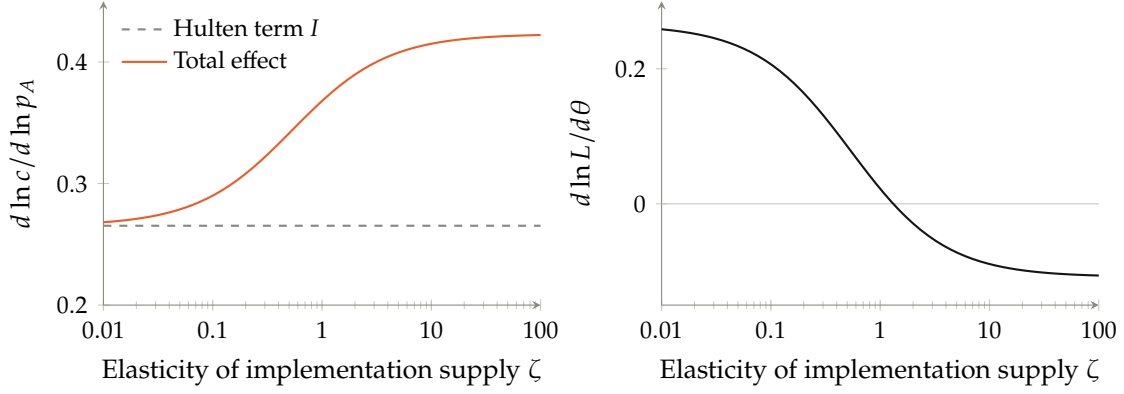


FIGURE 2. Left: elasticity of unit cost with respect to the price of agents. Right: response of production employment to cheaper agents. Both at $\theta = 0$, as functions of the elasticity of implementation supply.

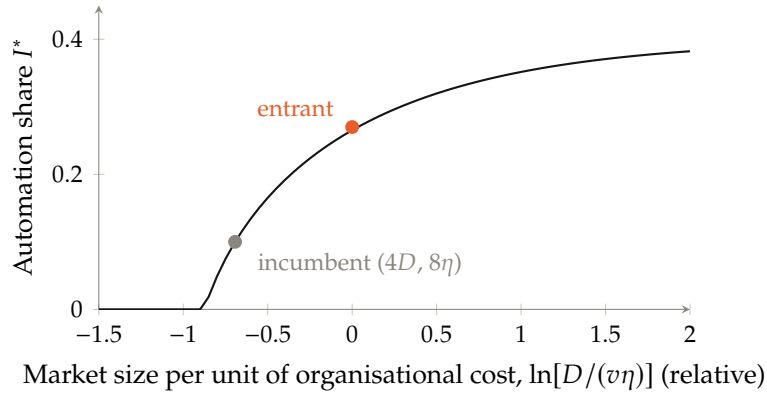


FIGURE 3. Automation share against market size per unit of organisational cost. Firms on the same curve differ only in $D/(v\eta)$.

Who automates. Figure 3 plots the automation share against $\ln[D/(v\eta)]$, relative to the baseline. An incumbent with four times the market of an entrant, but eight times its implementation requirement, automates a share 0.10 of its tasks, against 0.27 for the entrant (Corollary 1).

7.1 What the Hulten term requires

Acemoglu (2025) computes the ten-year productivity effect of AI as the share of tasks affected times the average cost saving. Twenty per cent of US tasks are exposed to AI; following Svanberg et al. (2024), 23% of those can be profitably automated within ten years, so 4.6% of tasks are affected; with average cost savings of 14.4% the effect on total factor productivity is about $0.046 \times 0.144 \approx 0.66\%$ over ten years. In that calculation the share of affected tasks is a parameter. In this model it is an equilibrium outcome governed by implementation capacity, and with $\zeta = 0$ and accumulation as in Section 9, $I_t = I_0 e^{\phi t}$. Reaching a given share in ten years therefore requires implementation capacity to grow at a specific rate.

Read through the model, Acemoglu's modest number is not only a statement about technology. If 1% of tasks are automated today, it presumes that the economy's capacity to implement AI grows by about 15% a year for a decade, a doubling roughly every five years sustained over ten. If capacity grows more slowly, the effect is smaller still; if AI learns to implement itself (Proposition 5), it can be larger without any growth in the implementation

TABLE 1. Growth of implementation capacity required by the Hulten term

Share of tasks automated today, I_0	0.5%	1%	2%
Annual growth needed to reach 4.6% in ten years	22%	15%	8%
Annual growth needed to reach 20% (all exposed tasks)	37%	30%	23%

Notes: $\phi = \ln(I_{10}/I_0)/10$, with fixed implementation requirement per task ($e_\eta = 0$) and inelastic short-run supply ($\zeta = 0$), so that the automation share grows at the rate of implementation capacity. Targets from Acemoglu (2025). Illustration, not an estimate: today's automated share is not measured, and the calculation ignores the intensive-margin gains on tasks already automated, which Proposition 3(c) shows are bounded by I_0 .

workforce. Measured growth in implementation occupations, and its response to model releases, is therefore the statistic that disciplines macroeconomic forecasts of AI.

8 Testable predictions

The model yields five predictions. Each can be tested with data that exist or can be assembled.

- 1. Implementation wages rise when AI prices fall** (Propositions 2 and 3b). Large cuts in API prices by model providers are discrete, public and plausibly exogenous to any individual employer. The model predicts that job postings and posted wages for forward-deployed engineers, AI engineers and implementation consultants rise after such cuts, and more so in occupations and regions where implementation supply is less elastic. Posting data with occupation and wage fields allow an event-study design.
- 2. Adoption depends on scale relative to legacy** (Proposition 1, Corollary 1). Holding size fixed, firms with more legacy complexity (older IT systems, more fragmented data, higher shares of on-premise software) automate less. Holding legacy fixed, larger firms automate more. Survey data on AI adoption linked to firm characteristics (McElheran et al., 2024; Bonney et al., 2024) can test both margins and the ratio between them. The distinctive prediction relative to the standard task model is about firm age: at a given size, younger firms, with less organisation to change, should automate more. The Census Business Trends and Outlook Survey reports AI use by firm size and age and allows this comparison directly.
- 3. Short-run productivity responds in proportion to the AI cost share** (Proposition 3c). The short-run response of firm or industry productivity to a fall in AI prices should be close to the current share of AI in costs, and larger where implementation labour is more elastically supplied.
- 4. Employment effects change sign over time** (Proposition 4). Early in adoption, employment effects in exposed occupations are small or positive. They turn negative as implementation capacity accumulates. The model predicts heterogeneity by the local supply of implementation labour, not only by task exposure.
- 5. Adopters report organisational change before labour change** (Propositions 3 and 4). Firms using AI should report new workflows, training and integration work well before they report changes in headcount. The Census Business Trends and Outlook Survey finds exactly this pattern (Bonney et al., 2024). The model turns the pattern from a description into a

prediction about which comes first.

9 Extensions

AI that implements itself. The most important extension lets the implementation requirement fall with model capability. If models can integrate themselves into workflows, write their own connectors and evaluate their own output, the bottleneck loosens from within. Let $\eta = \eta(\theta)$ and define the rate at which cheaper, better AI lowers its own implementation cost, $e_\eta \equiv -d \ln \eta / d\theta \geq 0$.

Proposition 5 (Self-implementing AI). *Under Assumption 1, at an interior equilibrium:*

(a) *The automation set responds according to*

$$\frac{dI}{d\theta} = \frac{(\varepsilon - 1)I + 1/s + e_\eta(1 + 1/\zeta)}{Q + 1/(\zeta I)},$$

which, with fixed implementation capacity ($\zeta = 0$), equals $e_\eta I$. The extensive margin opens at the rate at which AI lowers its own implementation requirement.

(b) *With $\zeta = 0$, the unit-cost elasticity is $d \ln c / d \ln p_A = I(1 + s e_\eta)$, so the Hulten bound of Proposition 3(c) is exceeded in proportion to e_η .*

(c) *With $\zeta = 0$, the implementation wage responds according to*

$$\frac{d \ln v}{d\theta} = (\varepsilon - 1)I + \frac{1}{s} + e_\eta(1 - QI).$$

If $QI \leq 1$ it rises for any e_η . If $QI > 1$, it rises, and the bottleneck persists in prices, if and only if

$$e_\eta < \bar{e} \equiv \frac{(\varepsilon - 1)I + 1/s}{QI - 1}. \quad (12)$$

The proposition turns the race between capability and organisation into a single comparison. Part (a) is the reason self-implementation matters: with fixed capacity, cheaper AI alone automates nothing, but AI that lowers η automates new tasks at rate e_η . Part (c) says when the scarcity rent survives. Below the threshold $\bar{e}$, demand for implementation from cheaper agents grows faster than self-implementation releases capacity, and implementation wages keep rising: the economy is still in the transition this paper describes. Above it, implementation wages fall as AI improves, and the bottleneck dissolves. The threshold is observable in principle, because it depends on the automation share and on the curvature of cost savings across tasks. The direction of implementation wages after model releases that improve agentic capability, holding API prices fixed, is the cleanest test. The derivatives are checked against finite differences in the replication code.

Accumulation. Implementation capacity is built through experience. Letting $\bar{H}$ grow with past implementation, $\dot{\bar{H}} = \phi \eta I$, gives an endogenous J-curve (Brynjolfsson et al., 2021). With fixed requirement η and $\zeta = 0$, $I = \bar{H}/\eta$, so the automation share grows exponentially, $I_t = \min\{\bar{I}, I_0 e^{\phi t}\}$, whatever happens to the price of AI. Productivity responds slowly at first

and accelerates as capacity accumulates, and displacement (Proposition 4) lags the price of AI by the time it takes capacity to grow. Section 7.1 uses this path to ask what growth of implementation capacity existing macroeconomic estimates require. The timing prediction can be confronted with the precisely estimated near-zero earnings effects two years after adoption in Danish data (Humlum and Vestergaard, 2025). This is the dynamic counterpart of organisation capital (Prescott and Visscher, 1980; Atkeson and Kehoe, 2005).

Rents in the AI stack. With $p_A = (1 + \mu)c_A$ and competition among model providers, a fall in the cost of inference c_A is shared among the provider (through μ), implementation labour (through v) and the adopting firm. When implementation is the bottleneck, a larger part of the surplus from cheaper inference goes to v . This links the paper to RN 1, where competition passes common cost reductions to customers.

Heterogeneous firms. With a distribution of (D, η) , implementation labour is allocated to firms with high D/η , raising concentration in automation. Whether cheaper AI increases or reduces concentration depends on whether it raises the demand for implementation more in large incumbents or in entrants.

General equilibrium. Closing the market for production labour lets w fall when automation displaces workers. That raises s and accelerates automation, and it moves real income toward the owners of implementation capacity. Allowing an elasticity of substitution between tasks different from one changes the share formulas but not the logic of the bottleneck.

Jagged integration. Tasks may differ in how hard they are to implement, $\eta = \eta(i)$. Firms then automate tasks in order of $s(i)/\eta(i)$ rather than $s(i)$, and the tasks that are easiest to automate technically may not be the first to be automated in practice. This parallels the jagged frontier of RN 2.

10 Conclusion

AI agents look like a new factor of production. They perform tasks as labour does and are owned as capital is. Because they are reproducible at the price of compute, however, they do not become scarce, and a factor that is not scarce earns no rent and imposes no constraint. This paper argues that the scarce factor in an economy of abundant agents is the capacity to deploy them. Implementation is a fixed cost per task, carried out by a distinct kind of labour. That single feature implies that cheaper AI raises the demand for, and the price of, the people who implement it. The effect of cheaper AI on productivity is bounded by the current automation share for as long as implementation capacity is fixed. And displacement of production workers is triggered by the expansion of that capacity, not by the fall in the price of AI.

The intended contribution is to locate the gap between what AI can do and what the economy gets from it. That gap is not a lag that closes on its own, and it is not mainly a measurement problem. It is the price of a scarce input, and it can be observed in the wages of the people who supply it. Policies and strategies that expand implementation capacity, through training, through tools that lower the cost of integration, or through organisations

designed around agents from the start, determine how much of the technology reaches the economy. Agents are abundant. What is scarce is the organisation able to deploy them.

References

- Daron Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025.
- Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4B, pages 1043–1171. Elsevier, Amsterdam, 2011.
- Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6): 1488–1542, 2018.
- Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019.
- Daron Acemoglu and Pascual Restrepo. Tasks, automation, and the rise in US wage inequality. *Econometrica*, 90(5):1973–2016, 2022.
- Philippe Aghion, Benjamin F. Jones, and Charles I. Jones. Artificial intelligence and economic growth. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*, pages 237–282. University of Chicago Press, Chicago, 2019.
- Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, 2018.
- Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, 2022.
- Andrew Atkeson and Patrick J. Kehoe. Modeling and measuring organization capital. *Journal of Political Economy*, 113(5):1026–1053, 2005.
- David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.
- Alexander Bick, Adam Blandin, and David J. Deming. The rapid adoption of generative AI. Working Paper 32966, National Bureau of Economic Research, 2024.
- Nicholas Bloom, Raffaella Sadun, and John Van Reenen. Americans do IT better: US multinationals and the productivity miracle. *American Economic Review*, 102(1):167–201, 2012.
- Kathryn Bonney, Cory Breaux, Cathy Buffington, Emin Dinlersoz, Lucia S. Foster, Nathan Goldschlag, John C. Haltiwanger, Zachary Kroff, and Keith Savage. Tracking firm use of AI in real time: A snapshot from the business trends and outlook survey. Working Paper 32319, National Bureau of Economic Research, 2024.

- Timothy F. Bresnahan and Manuel Trajtenberg. General purpose technologies: ‘engines of growth’? *Journal of Econometrics*, 65(1):83–108, 1995.
- Timothy F. Bresnahan, Erik Brynjolfsson, and Lorin M. Hitt. Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*, 117(1):339–376, 2002.
- Erik Brynjolfsson and Lorin M. Hitt. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives*, 14(4): 23–48, 2000.
- Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*, pages 23–57. University of Chicago Press, Chicago, 2019.
- Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.
- Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025.
- Wesley M. Cohen and Daniel A. Levinthal. Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1):128–152, 1990.
- Diego Comin and Bart Hobijn. An exploration of technology diffusion. *American Economic Review*, 100(5):2031–2059, 2010.
- Carol Corrado, Charles Hulten, and Daniel Sichel. Intangible capital and U.S. economic growth. *Review of Income and Wealth*, 55(3):661–685, 2009.
- Paul A. David. The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review: Papers and Proceedings*, 80(2):355–361, 1990.
- Avinash K. Dixit and Joseph E. Stiglitz. Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3):297–308, 1977.
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024.
- Luis Garicano. Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904, 2000.
- Zvi Griliches. Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4):501–522, 1957.
- David Hémous and Morten Olsen. The rise of the machines: Automation, horizontal innovation, and income inequality. *American Economic Journal: Macroeconomics*, 14(1):179–223, 2022.

- Charles R. Hulten. Growth accounting with intermediate inputs. *Review of Economic Studies*, 45(3):511–518, 1978.
- Anders Humlum and Emilie Vestergaard. Large language models, small labor market effects. Working Paper 33777, National Bureau of Economic Research, 2025.
- Enrique Ide and Eduard Talamàs. Artificial intelligence in the knowledge economy. *Journal of Political Economy*, 2025. doi:10.1086/737233.
- William Stanley Jevons. *The Coal Question*. Macmillan, London, 1865.
- Anton Korinek and Donghyun Suh. Scenarios for the transition to AGI. Working Paper 32255, National Bureau of Economic Research, 2024.
- Kristina McElheran, J. Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas. AI adoption in America: Who, what, and where. *Journal of Economics & Management Strategy*, 33(2):375–415, 2024.
- Paul Milgrom and John Roberts. The economics of modern manufacturing: Technology, strategy, and organization. *American Economic Review*, 80(3):511–528, 1990.
- Edward C. Prescott and Michael Visscher. Organization capital. *Journal of Political Economy*, 88(3):446–461, 1980.
- Pascual Restrepo. We won’t be missed: Work and growth in the AGI world. Working Paper 34423, National Bureau of Economic Research, 2025.
- Maja Svanberg, Wensu Li, Martin Fleming, Brian Goehring, and Neil Thompson. Beyond AI exposure: Which tasks are cost-effective to automate with computer vision? Working paper, MIT FutureTech, 2024.
- Donald M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2):305–321, 1978.
- Philip Trammell and Anton Korinek. Economic growth under transformative AI. Working Paper 31815, National Bureau of Economic Research, 2023.
- Joseph Zeira. Workers, machines, and economic growth. *Quarterly Journal of Economics*, 113(4):1091–1117, 1998.

A Proofs

Throughout, $s(i) = \ln[\psi(i)w/p_A]$ is strictly decreasing in i and $\partial s/\partial \theta = 1$. Every step below is checked symbolically and, for the equilibrium formulas, against finite differences on randomly drawn parameterisations in the replication code (`code/01_model_checks.py`).

Proof of Lemma 1. With the Cobb–Douglas aggregator (1), the unit cost of y is $\ln c = \int_0^1 \ln c(i) di$, where $c(i)$ is the minimum unit cost of task i . For $i \leq I$ this is $p_A/\gamma_A(i)$, because $s(i) \geq 0$ there. For $i > I$ it is $w/\gamma_L(i)$. Hence $\ln c = \int_0^I \ln(w/\gamma_L) - \int_0^I s = \ln c_0 - S(I)$. Under Cobb–Douglas, expenditure is equal across tasks, so agents receive a share I of variable cost and labour a share $1 - I$; equivalently $\partial \ln c / \partial \ln p_A = I$ and $\partial \ln c / \partial \ln w = 1 - I$. With constant marginal cost c and demand $Dp^{-\varepsilon}$, the profit-maximising price is $p = \varepsilon c / (\varepsilon - 1)$. Revenue is $R = Dp^{1-\varepsilon}$, operating profit is $\pi = (p - c)y = R/\varepsilon$, and variable cost is $TC = cy = (\varepsilon - 1)R/\varepsilon = (\varepsilon - 1)\pi$. Substituting p gives $\pi = \kappa D c^{1-\varepsilon}$, and with $c = c(I)$ this yields (3). $\square$

Proof of Proposition 1. Extend S to $[0, 1]$ as $S(I) = \int_0^I \max\{s(i), 0\} di$. Implementing a task beyond $\bar{I}(\theta)$ costs $v\eta > 0$ and saves nothing, so maximising (4) on $[0, 1]$ is equivalent to the original problem. Π is continuous on the compact set $[0, 1]$, so maximisers exist and the set of maximisers has a largest element. Write $TC(I) = D g(I; \theta)$ with g nondecreasing in I . Then:

- $\partial \Pi / \partial \theta = TC(I) \cdot \min\{I, \bar{I}(\theta)\}$, a product of two nonnegative functions that are nondecreasing in I . So Π has increasing differences in (I, θ) .
- $\partial \Pi / \partial D = g(I; \theta) / (\varepsilon - 1)$, nondecreasing in I : increasing differences in (I, D) .
- $\partial \Pi / \partial (-v\eta) = I$: increasing differences in $(I, -v\eta)$.

By Topkis (1978), the largest maximiser is nondecreasing in θ , D and $-v\eta$. Finally, $\Pi/(v\eta) = [D/(v\eta)] g(I; \theta) / (\varepsilon - 1) - I$, so the set of maximisers depends on (D, v, η) only through $D/(v\eta)$. The corollary follows because entrant and incumbent share θ , w and v , and I^* is nondecreasing in D/η . $\square$

Proof of Proposition 2. $H^d = \eta I^*$ is nondecreasing in θ by Proposition 1. At an interior optimum, $s(I^*) > 0$ because $TC \cdot s = v\eta > 0$, so the first-order condition can be written as $F(I, \theta) \equiv \ln TC(I, \theta) + \ln s(I, \theta) - \ln(v\eta) = 0$. Since $\partial \ln TC / \partial I = (\varepsilon - 1)s$ and $\partial \ln TC / \partial \theta = (\varepsilon - 1)I$,

$$\frac{\partial F}{\partial I} = (\varepsilon - 1)s + \frac{s'}{s} = -Q, \quad \frac{\partial F}{\partial \theta} = (\varepsilon - 1)I + \frac{1}{s}.$$

Differentiating the first-order condition gives $\Pi''(I^*) = TC [(\varepsilon - 1)s^2 + s']$. Assumption 1 therefore makes the optimum strict, and dividing it by $s > 0$ gives $Q > 0$. The implicit function theorem yields (6). $\square$

Proof of Proposition 3. Equilibrium. Demand $\eta I^*(v)$ is nonincreasing in v and supply $\bar{H}v^\zeta$ is strictly increasing for $\zeta > 0$, so there is at most one equilibrium. For $\zeta = 0$, $I = \bar{H}/\eta$, and $v = TC s/\eta$ from the first-order condition, which is positive when $\bar{H}/\eta < \bar{I}$.

(a) For $\zeta > 0$, substituting $\ln v = [\ln(\eta I) - \ln \bar{H}]/\zeta$ into the first-order condition gives

$$G(I, \theta) \equiv \ln TC + \ln s - \ln \eta - \frac{1}{\zeta} [\ln(\eta I) - \ln \bar{H}] = 0,$$

with $\partial G / \partial I = -Q - 1/(\zeta I) < 0$ and $\partial G / \partial \theta = (\varepsilon - 1)I + 1/s$. The implicit function theorem gives (8). At a given equilibrium point, the denominator is decreasing in ζ , tends to Q as $\zeta \rightarrow \infty$ and to $+\infty$ as $\zeta \rightarrow 0$. For $\zeta = 0$, $I = \bar{H}/\eta$ does not depend on θ .

(b) For $\zeta > 0$, $d \ln v / d\theta = (\zeta I)^{-1} dI / d\theta$, which gives (9). For $\zeta = 0$, $\ln v = \ln TC + \ln s - \ln \eta$ with I fixed, so $d \ln v / d\theta = (\varepsilon - 1)I + 1/s$, which is (9) at $\zeta = 0$. Positivity follows from $s > 0$ and $Q > 0$, and the expression is decreasing in ζ .

(c) $\ln c = \ln c_0 - S(I, \theta)$ and c_0 does not depend on θ . Hence $d \ln c / d \ln p_A = -d \ln c / d\theta = \partial S / \partial \theta + s dI / d\theta = I + s dI / d\theta$. Since $s > 0$ and $dI / d\theta \geq 0$, with equality if and only if $\zeta = 0$ by part (a), the inequality and its equality case follow. $\square$

Proof of Proposition 4. $\ln L = \ln(1 - I) + \ln TC - \ln w$, and w is fixed. Differentiating,

$$\frac{d \ln L}{d\theta} = -\frac{1}{1 - I} \frac{dI}{d\theta} + (\varepsilon - 1) \left(I + s \frac{dI}{d\theta} \right),$$

which is (11). (a) With $\zeta = 0$, $dI / d\theta = 0$. (b) Let $B \equiv (\varepsilon - 1)s - 1 / (1 - I)$, which is negative when $(\varepsilon - 1)s(1 - I) < 1$. Fix the initial equilibrium (I, v) and vary ζ , adjusting $\bar{H}$ so that $\bar{H}v^\zeta = \eta I$. Then (I, s, Q) are unchanged and, by part (a) of Proposition 3, $X(\zeta) \equiv dI / d\theta$ is strictly increasing from $X(0) = 0$. So $d \ln L / d\theta = (\varepsilon - 1)I + B X(\zeta)$ is strictly decreasing in ζ and positive at $\zeta = 0$. It has at most one zero $\hat{\zeta}$, and it has one if and only if $(\varepsilon - 1)I + B [(\varepsilon - 1)I + 1/s] / Q < 0$. $\square$