Is Software Eating the World? Measuring the Progress and Diffusion of AI

Filippo Bontadini*, Carol Corrado†, Jonathan Haskel‡
Cecilia Jona-Lasinio§

April 29, 2026
(preliminary draft, not for quotation)

Abstract

How should artificial intelligence be measured in the national accounts, and what observable evidence can assess its diffusion across the economy? This paper develops a framework mapping AI into existing national accounting categories and identifies two first-order diagnostics of AI diffusion: the AI capital factor income share, and the rate of change in the AI capital rental price. Both are subject to systematic mis-measurement from three sources: the asset boundary problem created by cloud delivery, in which capital services and/or capacity-building expenditures are classified as intermediate consumption; quality-change gaps in AI services price indexes that cause observed price declines to understate true reductions; and concentration of AI investment in a small number of frontier firms. We address these through modeling, analysis of the cloud boundary's treatment for factor shares, and a proof-of-concept price index for AI capital services that, alone, suggests extensive diffusion in the years to come. Our empirical estimates of cross-price input elasticities and simulation evidence of the multi-input case present a more nuanced picture consistent with insufficient complementary investments in organizational capital and the "weak links" environment described in Jones and Tonetti (2026).

1 Introduction

How should artificial intelligence (AI) be measured in the national accounts, and what observable evidence can be used to assess its effects on productivity and the distribution of factor incomes? Despite considerable interest in recent AI advances, no standard framework exists for connecting these developments to the core statistics economists use to track economic activity.

*LUISS University. Email: fbontadini@luiss.it.
†Georgetown University. Email: carol.corrado@georgetown.edu.
‡Imperial College Email: j.haskel@imperial.ac.uk.
§LUISS Business School Email: cjonalasinio@luiss.it.

The difficulty is not that the underlying economic mechanisms are unknown. The ways AI is expected to affect the economy—automation of tasks, augmentation of labor, and the reorganization of production—fit naturally within existing models. The challenge is measurement. AI does not appear as a separate input factor in the national accounts; it is embedded within existing asset categories such as software, computing equipment, and data-related assets.

Aspects of AI's impact are already visible in the data, even if not separately identified. Rising capital intensity, increasing income shares associated with software-related assets, and elevated rates of return in AI-producing sectors are all consistent with rapid technological change—and all observable in existing accounts.[1]

This paper develops a framework for extracting information about AI from those existing data. We conceptualize AI as a composite form of capital—software, data, and compute—that maps directly onto national accounts asset categories. Two measurement problems follow from this conceptualization. First, cloud delivery increasingly bundles compute and software into service subscriptions, making the standard boundary between current expenditure and capital formation genuinely ambiguous. Second, AI represents a tighter integration of hardware and software than did the IT era, whose familiar adage that "what Intel giveth, Microsoft taketh away" reflected a production structure in which computers and software could be priced as separate identifiable products. By contrast, a deployed AI model reflects the entire AI production stack, and its price should capture that. We therefore conceptualize the price of AI as "what the frontier model delivereth": a quality-adjusted cost per unit of productive AI output, regardless of which layer of the stack generated the gain, and provide a proof-of-concept application of this approach.

Approach: Two Diagnostics

To organize the analysis, we adopt a two-sector framework. An upstream sector produces AI-related capital; a downstream sector uses that capital in production. This separation allows us to distinguish technological progress in AI production from its diffusion throughout the economy. The framework also captures a self-amplifying dynamic: as AI raises upstream productivity, it lowers the relative price of AI capital, which through the user cost of capital accelerates downstream substitution. The magnitude of that response depends on the elasticity of substitution between AI capital and labor in the downstream sector, moderated by the accumulation of complementary organizational capital.

This framework points to two diagnostics that should be particularly informative about AI diffusion. The first is the behavior of factor income shares associated with AI-related capital, which provides indirect evidence on the elasticity of substitution between AI capital and labor. The second is the behavior of the AI capital rental price. Observed declines in AI capital services prices are a signal of upstream productivity advantage and a leading indicator of downstream diffusion.

Interpreting both diagnostics requires care on three fronts. First, investment and returns

[1] For example, Bontadini, Corrado, Haskel and Jona-Lasinio (2026 forthcoming) find that software capital production and use accounted for roughly one-half of US labor productivity growth and its post-2017 acceleration, with software-producing industries — only about 6 percent of the nonfarm economy — contributing more than 40 percent of the acceleration in total factor productivity growth.

are currently concentrated among producers of frontier AI models, so aggregate patterns may reflect AI producers' behavior rather than broad-based substitution across the rest of the economy. Because a small number of firms account for the large majority of global investment in foundational AI models and proprietary training data, the factor shares and capital costs measured at the industry or aggregate level are dominated by the behavior of those frontier firms. Estimated elasticities of substitution will primarily reflect factor substitution at those firms, and may not be representative of the broader economy. Decomposing aggregate factor share trends into contributions from frontier firms and the rest of the economy is therefore an important empirical priority, one that requires linking the framework to firm-level data on market structure, profitability, and factor shares—a distinction that is especially important in cross-country comparisons. Second, a cloud asset boundary problem, described in detail in Section 4.3, means that rapid growth in cloud services bolsters the measured operating surplus of software-producing industries, while a portion of those expenditures is economically capital services in using industries and should be reclassified as such when examining factor shares for evidence of diffusion beyond the software sector. Third, even large price declines and factor reallocation may yield modest output gains if AI continues to expand in tasks that are non-binding constraints on production; the aggregate implications depend on task composition and the availability of complementary organizational capital, as in the "weak links" environment of Jones and Tonetti (2026); Jones (2026).

Financial implications

Financial markets provide a forward-looking counterpart to the two diagnostics above. Expected future gains in upstream AI productivity are capitalized into asset valuations today, while announced commitments to data center construction and foundational model R&D represent observed leading indicators of the investment flows that will eventually appear in the capital account. From the framework's perspective, both carry information: valuations embed beliefs about the pace of AI capital cost declines and the degree to which those declines will translate into economy-wide output gains, while announced spending commitments—if and when capitalized correctly—will shape the measured capital stock and rental rates that are central to the diagnostics developed here. A full treatment of these financial and investment signals is left for future work; we flag the connection in the conclusion.

2 A Two-Sector Framework for AI

We begin by distinguishing between the production of AI and its use. This distinction is central to both measurement and interpretation. Consider an economy with two sectors. The upstream sector produces AI-related capital, primarily in the form of software and related intangible assets. The downstream sector uses this capital as an input into production. Formally, output in the upstream sector is given by

$$X_S = A_S F_S(K_S^{AI}, K_S^O, L_S), \tag{1}$$

while output in the downstream sector is

$$Y_C = A_C F_C(K_C^{AI}, K_C^O, L_C, Z_C). \tag{2}$$

In both equations, X_S and Y_C denote real output; A_S and A_C are sector-specific total factor productivity (TFP) levels; $F_S(\cdot)$ and $F_C(\cdot)$ are constant-returns production functions; K^{AI} denotes the stock of AI-related capital (software, trained models, and associated data assets); K^O denotes all other capital; and L denotes labor input. Subscripts S and C index the upstream (AI-producing) and downstream (AI-using) sectors respectively. The downstream production function additionally includes Z_C to denote complementary organizational capital—process redesign, co-invention, and firm-specific human capital—that must co-evolve with AI capital before its full productivity benefits are realized.

Capital–Labor Substitution(s)

The two-sector framework also clarifies that there are two distinct notions of capital–labor substitution in the context of AI.

The first is substitution across the economy. In this case, firms use AI-related capital to replace or augment labor in a wide range of activities. This is the channel typically emphasized in discussions of automation.

The second is substitution within the AI-producing sector itself. Here, AI-related capital is used in the production of new AI-related capital. For example, machine learning systems may assist in writing code or developing new models. In this case, substitution occurs within the upstream sector, as AI capital increasingly replaces labor in its own production.

These two channels are conceptually distinct but empirically intertwined. Observed changes in aggregate capital shares reflect both, but it is diffusion across downstream sectors that drives broad-based improvements in aggregate productivity.

Coordination, Organzational Capital, and Tasks

Milgrom and Roberts (1990) identify a fundamental lacuna in the standard production function: it treats the firm as a costless optimizer, selecting input combinations from a menu of technically feasible plans without friction. But in practice, production requires what they call *complementary activities*—choices about technology, organization, and incentives that must be made jointly and consistently to be productive at all. A firm adopting flexible manufacturing cannot simply swap in new machinery; it must simultaneously reorganize job assignments, renegotiate supplier relationships, redesign information flows, and restructure incentive systems. The complementarities are pervasive, and getting them wrong is costly. The production function, $Y = AF(K, L)$, is silent on all of this: it assumes the firm is already *inside* the plan, with coordination solved.

This is precisely the gap that organizational capital fills in the intangibles framework (Corrado et al., 2005, 2009). Organizational capital is investment in the *design and implementation* of production arrangements—business process re-engineering, management consulting, the codified knowledge embedded in operating procedures, and the relational capital that sustains supply-chain and partner networks. It is, in effect, the capitalized value of

solving the coordination problem Milgrom and Roberts identify. The expenditure is current, but the return is durable: a well-designed organizational architecture continues to generate value in subsequent periods, exactly as physical capital does. That durability is what makes it investment rather than intermediate consumption, even though standard national accounts treat much of it as the latter.

The task framework (Autor et al., 2003; Acemoglu and Restrepo, 2018a, 2019) offers a complementary micro-foundation by decomposing production into discrete *tasks* rather than undifferentiated labor and capital inputs. In this view, production requires a bundle of tasks—cognitive, manual, and communicative—that must be assigned to workers or machines and sequenced coherently. Coordination is itself a task, or more precisely a *meta-task*: the activity of specifying, monitoring, and adjusting the assignment of other tasks. It is non-routine, judgment-intensive, and context-dependent—the canonical features of tasks that resist automation and that command wage premia in the empirical literature.

Organizational capital, in task-framework terms, is the accumulated stock of solutions to the meta-task problem. It includes the codified procedures that reduce the cognitive load of routine coordination (standard operating procedures, ERP system configurations, decision rules), and the relational and reputational assets that lower the cost of non-routine coordination (trusted supplier relationships, management teams with shared mental models, organizational cultures that align incentives without constant monitoring). Investment in organizational capital is therefore investment in reducing the unit cost of the coordination meta-task—or, equivalently, in raising the productivity of the workers who perform it.

This framing has a useful implication for measurement. If coordination is a task, it is in principle separable, assignable, and—increasingly—automatable. Firms already outsource portions of it to management consultants, enterprise software vendors, and platform intermediaries. As AI systems become capable of performing more of the cognitive sub-tasks involved in coordination (monitoring supply chain states, synthesizing performance information, flagging incentive misalignments), the boundary between organizational capital as a *stock* and AI-assisted coordination as a *service flow* becomes analytically important. In national accounts terms, the question is whether firm expenditure on AI-mediated coordination tools constitutes investment in a new organizational-capital asset, or purchase of an intermediate service—a question that maps directly onto the cloud-delivery problem discussed in Section 4.

Recursive Production and the Self-Amplifying Mechanism

The second channel introduces a recursive element into the production process. The mechanism operates like a flywheel: each turn lowers the cost of the next. Software is used to produce software, and AI systems are increasingly used to develop new AI systems—so productivity gains accrue not only to the efficiency of production but to the pace of innovation itself.

More precisely, the flywheel operates upstream. AI capital deployed in sector S raises A_S, which lowers the acquisition price P_S of new AI capital, which induces further upstream accumulation. From the perspective of downstream firms, this self-amplifying mechanism continuously shifts the frontier of productive AI capacity available to them: a falling r^{AI} represents an expanding opportunity to substitute AI capital for labor, independent of anything

downstream firms themselves do.

Whether and how quickly downstream firms take up that opportunity is a separate matter, governed by the accumulation of complementary organizational capital Z_C. Firms must co-invest in Z_C before AI capital can be deployed productively, so the full response of downstream factor shares to a declining r^{AI} is realized only gradually. A recursive element is present here too — more Z_C enables more productive use of AI capital, which raises the return to further Z_C investment — but it is muted and lagged relative to the upstream flywheel. The result is a time series in which the supply-side mechanism accelerates the expansion of the available frontier while the demand-side transition creates a persistent gap between that frontier and observed factor shares.

These two forms of recursion pull in opposite directions in the time series: the supply-side mechanism accelerates AI capital diffusion (in the upstream sector), while the demand-side transition creates a period during which observed factor shares lag behind the underlying substitution impulse—the output and factor-share implications of which are governed by the Z_C and task-composition mechanisms discussed above.

3 Two Diagnostics

To connect the framework to observable data, we focus on the relationship between factor inputs, factor income shares, and relative factor input prices.

3.1 Capital–Labor Substitution and Factor Shares

We first examine the behavior of factor income shares associated with AI-related capital. A natural starting point is a CES production function for sector $j \in \{S, C\}$:

$$Y_j = A_j \left[\alpha_j (K_j^{AI})^{\rho_j} + (1-\alpha_j) L_j^{\rho_j}\right]^{1/\rho_j}. \tag{3}$$

Here $\alpha_j \in (0,1)$ is a distribution parameter governing the intensity of AI capital in production, $\rho_j = (\sigma_j - 1)/\sigma_j$ is the substitution parameter, and L_j denotes labor input. The elasticity of substitution between AI capital and labor is

$$\sigma_j = \frac{1}{1-\rho_j}. \tag{4}$$

It is useful to interpret this parameter. If $\sigma < 1$, capital and labor are complements. If $\sigma = 1$, factor shares are constant. If $\sigma > 1$, capital and labor are substitutes, and increases in capital raise the capital share.

Factor income shares provide an observable diagnostic of the diffusion of AI capital. Let $s_{K,j}^{AI}$ denote the share of sector j value added accruing to AI capital, and $s_{L,j}$ the corresponding labor share. Under cost minimization:

$$\frac{s_{K,j}^{AI}}{s_{L,j}} = \frac{\alpha_j}{1-\alpha_j} \left(\frac{K_j^{AI}}{L_j}\right)^{\rho_j}. \tag{5}$$

At the same time, firms respond to relative factor prices. Let w_j denote the wage and r_j^{AI} the rental rate (user cost) of AI capital in sector j. Cost minimization implies:

$$\frac{K_j^{AI}}{L_j} \propto \left(\frac{w_j}{r_j^{AI}}\right)^{\sigma_j}. \tag{6}$$

Combining these expressions yields:

$$\ln\left(\frac{s_K^{AI}}{s_L}\right)_j = \text{const}_j + (\sigma_j - 1)\ln\left(\frac{w}{r^{AI}}\right)_j. \tag{7}$$

The elasticity of substitution σ between AI capital and labor organizes the central trade-off: if $\sigma > 1$, falling r^{AI} raises the capital share and amplifies the self-reinforcing mechanism of Section 2; if $\sigma < 1$, the same price decline compresses capital's share and the distributional and growth dynamics are qualitatively different. The value of σ therefore enters directly into projections of factor income distribution and assessments of AI's labor-displacing potential.

The two-input framework is, however, insufficient for rigorous policy analysis, for two reasons. First, estimates of σ capture only the *net* substitution effect after complementary organizational capital Z_C, which is unmeasured in practice, has been accumulated; during the transition, the observed factor share ratio understates the long-run response to a given decline in r_C^{AI} (Brynjolfsson et al., 2021). Second, even a well-identified σ does not determine aggregate output gains: those depend on task composition and whether AI is deployed in tasks that represent binding constraints on production. σ governs the responsiveness of factor proportions to relative prices; Z_C and task composition jointly govern how those changes in factor proportions translate into output. We therefore use the two-input case as an expository and diagnostic device, and develop the multi-input Morishima framework with a proxy for Z_C to govern the empirical work in Section 7.

3.2 Capital Costs and the Role of Prices

The second diagnostic concerns the cost of AI capital, which appears directly in the CES factor demand system. The relevant concept is the user cost (Jorgenson, 1963), which converts the acquisition price of a capital asset into the rental price a firm must pay (or implicitly charge itself) for one unit of capital services for one period. In the two-sector framework, sector-C firms evaluate this cost in units of their own output price P_C. The user cost of AI capital for sector C is therefore:

$$r_C^{AI} = P_S\left(\varrho + \delta - \pi\right), \tag{8}$$

where P_S is the acquisition price of a new constant-quality unit of AI capital, ϱ is the required real rate of return, δ is the economic depreciation rate of AI capital, and

$$\pi \equiv \frac{\dot{P}_C}{P_C} - \frac{\dot{P}_S}{P_S} \tag{9}$$

is the rate of decline of the acquisition price of AI capital *relative to* sector-C output prices.[2]

The relative price π is the central empirical object of this paper. Introducing a capability index Q to adjust an observed nominal price per unit P_S^{obs} for quality ($P_S \equiv P_S^{obs}/Q$), π decomposes into three additive components:

$$\pi = \underbrace{\dot{P}_C/P_C}_{\text{(i) inflation in } C} + \underbrace{\dot{Q}/Q}_{\text{(ii) capability improvement}} + \underbrace{(-\dot{P}_S^{obs}/P_S^{obs})}_{\text{(iii) nominal price decline}} . \tag{10}$$

Component (i) reflects the fact that π is the change in a relative price: even holding AI capital prices fixed, a rising P_C lowers the relative cost of AI capital and induces substitution. Components (ii) and (iii) are the capability improvement and nominal unit price decline terms that are the primary focus of the empirical work in Section 5. In the two-sector framework of Section 2, the result of Oulton (2012) connects π directly to relative productivity: in steady state, $\pi \approx \mu_S - \mu_C$, where μ_S and μ_C are the TFP growth rates of the AI-producing and AI-using sectors. The observed rate of relative AI capital cost decline is therefore an empirical signal of the upstream productivity advantage of sector S, not an independent cause of downstream substitution, and it is π—not the absolute level of P_S—that is capitalized into asset valuations.

In competitive equilibrium, r_C^{AI} and P_S move in tandem: a firm that owns AI capital and rents it out will set the rental rate equal to the user cost, so both fall at the same rate as P_S declines (holding ϱ and δ constant). This means that the rate of decline of observable AI capital *services* prices is a valid empirical proxy for the rate of decline of P_S, without requiring a separately observed acquisition price. This point matters because, as the next section establishes, sector-C firms purchasing AI capabilities via cloud subscriptions observe only a services pricec rather than an acquisition price. What Section 5 constructs is therefore a quality-adjusted index of AI capital *services* prices—and under competitive equilibrium, that is precisely what is needed to track the decline in P_S that drives r_C^{AI}.

From diagnostics to measurement

The two diagnostics identified in this section—the AI capital factor share and the user cost of AI capital—are jointly informative about the pace and depth of AI diffusion, but only if they can be correctly measured. Both turn out to be systematically distorted by mismeasurement: the delivery of AI capabilities through cloud services rather than via owned assets means that a portion of capital services used by downstream firms is largely invisible in national accounts, and the price decline that drives r^{AI} is not captured by any existing official deflator. The next two sections address these problems in turn: Section 4 defines the asset boundary and formalizes the shadow capital distortion; Section 5 constructs a proof-of-concept quality-

[2]The nominal user cost for sector C is $P_S(i - \dot{P}_S/P_S + \delta)$, where i is the nominal required rate of return. Applying the Fisher relation $i = \varrho + \dot{P}_C/P_C$ and rearranging yields equation (8) with π as defined in equation (9). For the upstream sector S, which produces AI capital and prices its output at P_S, the corresponding user cost is $P_S(\varrho_S + \delta - \dot{P}_S/P_S)$ — the standard Hall-Jorgenson expression with the acquisition price's own rate of change as the capital gains term. The downstream expression differs because sector-C output is priced at $P_C \neq P_S$; it is the relative price P_S/P_C that governs the substitution incentive in sector C, not the absolute level of P_S.

adjusted price index for AI capital services. Together they establish the empirical foundations on which the factor share analysis of Section 7 rests.

4 Defining AI Capital

Both diagnostics hinge on a clear prior definition of what counts as AI capital. The question is more than technical: the producer–user distinction that is straightforward for owned assets becomes genuinely ambiguous when AI capabilities are delivered through the cloud. The remainder of this section sets out the AI production stack to clarify why the treatment of cloud expenditures in national accounts is not a "one size fits all" issue—specifically, why it affects users rather than producers of AI capital, and how this confounds the factor-share diagnostic. The stack taxonomy also motivates the price measurement approach in Section 5.

4.1 The AI Production Stack

The AI production stack provides a useful taxonomy for thinking through asset boundary questions. AI capabilities reach final users through a sequence of functionally distinct layers, each with different economic characteristics, different relationships to the producer–user distinction, and different implications for the national accounts.

Figure 1 summarizes the stack; we describe each layer in turn. The first four layers describe the delivery chain from infrastructure to end user; the fifth identifies the upstream supplier of the physical capital on which that chain depends.

Layer one: AI-native software and applications. The outermost layer consists of purpose-built AI applications and platforms—generative AI assistants, AI-enabled vertical-industry software, autonomous workflow tools. These are recognizable as final goods or services delivered to business or consumer users. Revenue at this layer is relatively straightforward to classify: business-to-business AI software subscriptions are primarily intermediate consumption or capital formation by user firms; business-to-consumer applications enter final consumption. The rapid growth at this layer is captured, at least partially, in existing software market estimates, though AI-specific components are not separately identified in national accounts.

Layer two: Foundation models and APIs. Immediately below the application layer sits the layer of large-scale foundation models—the frontier language, image, and reasoning models developed by the major AI labs—accessed through application programming interfaces. This layer raises the producer–user boundary problem most acutely. A firm calling an API to generate text, analyze documents, or power a customer-facing product is, from the accounts perspective, purchasing a current service. But the same subscription may simultaneously fund fine-tuning, proprietary embeddings, or workflow development that generates benefits well beyond the current period — activities that are economically investment regardless of how they are billed. Whether the expenditure constitutes capital formation depends on the nature of what is being created, not on the form of delivery. The current convention—expense the subscription, capitalize nothing—handles both cases identically and will misclassify a substantial fraction of what is economically investment.

Layer three: Platforms, tooling, and MLOps. The third layer encompasses the infrastructure that enables firms to build, train, evaluate, and deploy AI systems: MLOps platforms, fine-tuning infrastructure, vector databases, evaluation tooling, and the associated orchestration software. Expenditure here is predominantly by firms that are themselves producing AI capabilities—the layer-two model developers and the layer-one application builders. It has the character of intermediate goods in production of AI capital, and a significant fraction qualifies as capital formation under the existing SNA framework, though it is rarely identified as such.

Layer four: Hyperscale cloud AI infrastructure. At the base of the stack sits the compute infrastructure itself—the GPU clusters, custom accelerators, and associated networking owned and operated by hyperscale cloud providers. This layer is the primary site of the producer–user distinction problem for AI. The cloud providers own capital-intensive assets; the firms accessing those assets do not. Training runs and large-scale inference workloads accessed as cloud services generate productive capacity for the user firm, but that capacity appears on no user firm's balance sheet and no national accounts investment table. From the ownership perspective, investment is recorded for the cloud providers. From the residence perspective, investment is understated for the many user firms whose AI capabilities are built on rented compute.

Layer five: AI semiconductor producers. Beneath the cloud infrastructure layer sit the firms that manufacture the physical capital on which the entire stack depends: the producers of GPU clusters, custom accelerators, and associated networking equipment — NVIDIA, AMD, and the custom silicon programs of Google (TPU), Amazon (Trainium), and Microsoft (Maia). Unlike layers one through four, which describe stages in the delivery of AI capabilities to users, layer five describes a supplier industry whose output is purchased as capital equipment by the hyperscalers and large model developers in layer four. Layer-five expenditure is straightforwardly classified as gross fixed capital formation in tangible capital equipment by the purchaser. At approximately $180 billion globally in 2025 (see below), layer five is the largest single component of AI-related investment by value, and it is the primary driver of the price of AI capital P_S in the user-cost formula. Rapid and sustained quality improvement in AI semiconductors—measured by performance per dollar rather than nominal price influences the price measurement approach to the entire stack developed in Section 5.

Figure 1: **The AI Industry Stack: Definition, Representative Vendors, and Revenue, 2025 and 2035E**

Layer	Definition · Representative vendors	Revenue (billions) 2025	2035E
L1 Applications	AI-native software sold directly to end-users: consumer chatbots, enterprise copilots, coding assistants, and vertical AI (healthcare, legal, finance). *e.g. GitHub Copilot · ChatGPT / Claude.ai · Harvey · Microsoft 365 Copilot*	**$19**	$650
L2 Foundation Models	API access to large language models and other foundation models, sold to developers and enterprises. Consumer subscriptions are L1. *e.g. OpenAI API · Anthropic API · Google Gemini API · Mistral · Cohere*	**$20**	$470
L3 Platforms & Tooling	MLOps, vector databases, model orchestration, AI observability, and fine-tuning infrastructure — middleware connecting models to applications. *e.g. Pinecone · Weaviate · Weights & Biases · LangChain · Hugging Face*	**$6.4**	$77
L4 Hyperscale Cloud AI	AI-attributable increment in hyperscaler cloud revenue (AWS, Azure, GCP) plus neocloud GPU-as-a-service. AI increment only — total cloud excluded. *e.g. AWS Bedrock / SageMaker · Azure OpenAI Service · GCP Vertex AI · CoreWeave · Lambda Labs*	**$27**	$715
L1–L4 total (software and services)		**$72**	∼$1.9T
L5 Compute	AI chip revenue: NVIDIA, AMD, and custom silicon. Excluded from L1–L4 software totals to avoid double-counting hardware. *e.g. NVIDIA H100 / B200 · AMD MI300X · Google TPU v5 · AWS Trainium · Microsoft Maia*	**$180**	∼$1.1T

Sources: Menlo Ventures (2026); Epoch AI; Microsoft FY25 Q3 IR; Synergy Research; NVIDIA FY2025 10-K. 2035 figures are base-case projections (authors' estimates). All figures are global, ownership basis. Deduplication: consumer subscriptions in L1 not L2; L4 = AI increment only; L5 excluded from L1–L4 software total.

The interaction among these layers compounds the measurement problem. A downstream firm building an AI-powered product may simultaneously rent GPU-hours from a layer-four provider (layer-four service consumption), access a foundation model via API (layer-two service consumption), use MLOps tooling (layer-three intermediate consumption), and deploy a layer-one application. Across this entire stack, the only expenditure likely to be capitalized under current conventions is any bespoke software development the firm commissions or undertakes directly as R&D. The remainder—potentially the larger part of the firm's total AI-related outlay—is expensed. The systematic undercounting of AI-related capital services that results biases both the factor share diagnostics and the user cost calculations in the same direction: toward understating the pace and depth of AI capital deepening in the downstream sector.

AI Revenue by Layer. Figure 1 reports estimates of global revenue at each layer in 2025 and a base-case projection for 2035, free of duplication across layers. The 2025 estimates are anchored in company reports and industry data; the 2035 projections are based on industry sources, with deduplication by the authors. The orders of magnitude are subject to considerable uncertainty and should be read as illustrative of the scale of diffusion implied by current growth trajectories rather than as point forecasts. Nonetheless, they

underscore the quantitative significance of the measurement issues identified in this section: as AI-related expenditure grows toward these levels, the misclassification of downstream cloud expenditure as intermediate consumption, and the absence of quality-adjusted price deflators for AI capital, will generate increasingly large distortions to existing measures.

The final-demand and intermediate-demand composition of revenues differs markedly across layers, and this structure matters for how measurement errors propagate. Layer-one revenues are split between final demand—consumer and government subscriptions for AI-enabled products and services—and intermediate demand from businesses deploying AI in production. Layer-two revenues (foundation model API access) are almost entirely intermediate: developers and enterprises are the buyers, not final consumers. The same is true of layer-three tooling and MLOps, which sells exclusively to firms building AI systems. Layer-four compute is similarly an intermediate good, purchased by cloud operators who supply it in turn to layer-two and layer-three users, though from the national accounts perspective the investment is recorded for the hyperscaler rather than the ultimate user firm. The implication is that measurement error is concentrated in the intermediate layers, where final output is hardest to observe and where the capital-formation/intermediate-consumption boundary is most contentious. Any understatement of AI-related investment at layers two and three does not simply reduce the measured capital stock; it also overstates intermediate consumption, understates value added, and suppresses measured TFP in the downstream sectors that are actually deploying AI.

The layered structure maps directly onto national accounts asset categories, as follows.

4.2 AI Asset Types

A firm investing in AI capabilities simultaneously incurs expenditures of three qualitatively different types, each posing a distinct challenge for our diagnostic measures. The types are as follows:

Type 1 — own-account software, data, and R&D. This category includes expenditures on software development, model training, and R&D that generate identifiable, firm-specific assets with multi-period benefits. These expenditures satisfy the economic definition of capital formation and are, in principle, within the asset boundary of the System of National Accounts (SNA) although measurement of data remains incomplete in practice. The progressive capitalization of software, R&D, and—most recently—data reflects the gradual incorporation of intangibles into official statistics.

Type 2a — AI capital services consumed in production. API access, inference compute, and cloud AI services purchased by sector-C firms as direct inputs into current production. These expenditures reflect productive capacity embodied in a shadow capital stock held by the provider—they are, in economic substance, rental payments for AI capital services that the purchasing firm consumes but does not own. The shadow capital formalization in Section 4.3 applies directly to this sub-type.

Type 2b — AI capital services consumed in own-account investment. Cloud services, APIs, and compute purchased to build something durable: a fine-tuned model, a bespoke AI system, a proprietary inference pipeline. The expenditure takes the form of a service fee but generates a Type 1 asset. The correct treatment is to impute the cloud expenditure as an input to own-account investment and record the resulting asset as

intangible capital formation—bringing Type 2b within the existing SNA framework for own-account software and R&D, and requiring only that statistical agencies identify and reclassify the relevant portion of cloud service fees. Unlike Type 2a, shadow stock imputation does not apply here: the durable output is recorded as the firm's own investment, not as a rental payment for external capital services.

Type 3 — complementary organizational investment. This category includes process redesign, co-invention, firm-specific human capital, and other organizational adjustments required to deploy AI productively. These expenditures are largely embedded in labor costs and internal operations and remain outside the asset boundary in official accounts, despite their central role in enabling returns to AI capital. To the extent possible, however, these assets are included in the intangibles-extended accounts EU KLEMS & INTANProd (LUISS 2025).

In practice, firms undertake all three types of expenditure jointly. The classification of these expenditures—particularly whether Type 2 and Type 3 are treated as intermediate consumption or as capital formation—has first-order implications for measured investment, capital stocks, and productivity.

4.3 The Cloud and the Shadow Capital Stock

The central problem

The asset-boundary problem formalized in this section concerns Type 2a expenditures—AI capital services consumed directly in production. Type 2b expenditures (cloud services consumed in own-account investment) are addressed through the capitalization route described above and do not require shadow stock imputation. For Type 2a, when sector-C firms purchase foundation model API access, cloud compute, or data pipeline services as inputs to current production, the accounts record those payments as intermediate consumption. Yet these expenditures reflect productive capacity embodied in a shadow capital stock held by the provider—they are, in economic substance, rental payments for AI capital services. The mismatch between economic reality and accounting convention biases both diagnostics developed in Section 3.1: it suppresses the measured AI-capital income share in the using sector and distorts the observed user cost of AI capital.

The 2025 revision of the System of National Accounts addresses the cloud asset boundary problem. As set out in Reinsdorf (2023), long-term contracts granting a user firm dedicated access to cloud hardware should be treated as financial leases—and hence as fixed capital formation by the user—when the user bears the operating risk of the asset. It also clarifies that software licenses of one year or more are fixed assets of the licensee regardless of whether the software is hosted in the cloud. These provisions address the cases in which legal and economic ownership are most clearly separable. What they do not address is the dominant mode of AI capital delivery: standard pay-per-use API access, foundation model subscriptions, and inference compute, where no long-term dedicated contract exists and no ownership risk transfers to the user firm. It is precisely this class of expenditure—the fastest-growing and most economically significant component of AI-related cloud spending—that the shadow capital framework formalizes below.

The practical gap is wider still. The October 2025 compilation guidance prepared for BPM7 and SNA 2025 implementation (Sakai et al., 2025) confirms that even the conceptually preferred approach—attributing cloud service transactions to the location of actual service delivery—remains operationally infeasible for most statistical agencies. Resource pooling and dynamic workload allocation mean that neither users nor compilers can reliably identify where computing services are physically provided, and the guidance note recommends recording transactions on the basis of actual payments as a practical second-best. The misclassification of Type 2a expenditure as intermediate consumption is therefore not merely a conceptual gap in the SNA framework but an acknowledged practical constraint that current data sources cannot resolve.

Formalization

To make this precise, decompose total cloud expenditure by firms as

$$M = M^{AI} + M^{comp} + M^{legacy}, \tag{11}$$

where M^{AI} denotes expenditure on AI-related cloud services, M^{comp} other compute services, and M^{legacy} legacy cloud and software services. Within M^{AI}, a critical distinction arises between expenditures by sector-S firms (AI producers, building models and capabilities) and sector-C firms (AI users, deploying AI in production):

$$M = \sum_{k \in \{AI,\, comp,\, legacy\}} \left(M_S^k + M_C^k\right). \tag{12}$$

For sector-S firms, M_S^{AI} is an intermediate input in producing AI capabilities and is correctly treated as such in the accounts. For sector-C firms, M_C^{AI} provides access to capabilities that enter directly into production—and it is here that the misclassification arises.

We capture this through a shadow reclassification:

$$M_C^{AI} = r_C^{AI} \cdot K_{shadow}^{AI}, \tag{13}$$

where K_{shadow}^{AI} denotes the unobserved stock of AI-related capital effectively used by sector-C firms and r_C^{AI} is its user cost. Equation (13) does not require that sector-C firms own the underlying assets, nor does it assert that all cloud expenditure should be capitalized. It states that the services purchased by sector-C firms can, in part, be interpreted as the flow of services from a capital stock that is not recorded in the accounts.

Corrected factor shares by sector

Because M_C^{AI} is treated as intermediate consumption, it is excluded from sector-C value added and therefore from measured sector-C capital income. The corrected AI-capital income for the using sector is

$$\Pi_C^{AI,true} = \Pi_C^{AI,obs} + M_C^{AI}, \tag{14}$$

and the corrected factor-share ratio for sector C is

$$\left(\frac{s_K^{AI}}{s_L}\right)_C^{true} = \frac{\Pi_C^{AI,obs} + M_C^{AI}}{w_C L_C}. \tag{15}$$

This is the direct mapping from the asset-boundary problem to the empirical framework of Section 3.1: the observed factor-share ratio for sector C omits M_C^{AI} from the numerator and therefore understates the true importance of AI capital in downstream production. The same omission implies an overstated rental rate r_C^{AI}—since the shadow capital stock is excluded from measured investment, the denominator of the user-cost expression is too small—so the relative price ratio w/r_C^{AI} entering equation (7) is also mismeasured. Both effects bias the estimated elasticity of substitution $\hat{\sigma}_C$ downward.

The counterpart in sector S runs in the opposite direction. Because M_C^{AI} is recorded as revenue accruing to cloud and model providers, it inflates the measured operating surplus of sector-S firms. The observed factor-share ratio for the producing sector therefore overstates the true return to AI capital there:

$$\left(\frac{s_K^{AI}}{s_L}\right)_S^{obs} = \frac{\Pi_S^{AI,obs}}{w_S L_S} = \frac{\Pi_S^{AI,true} + M_C^{AI}}{w_S L_S}. \tag{16}$$

The two sectoral errors are mirror images: M_C^{AI} is missing from the correct numerator in sector C and spuriously present in the observed numerator in sector S.

Aggregate versus sectoral factor shares

The asset-boundary problem primarily affects the *allocation* of capital income across sectors rather than, in the first instance, its aggregate level. To a first approximation the two sectoral errors cancel: M_C^{AI} shifts measured capital income from sector C to sector S but leaves aggregate capital income and aggregate value added unchanged. Aggregate factor shares may therefore appear stable even as AI capital use expands in the using sector. That apparent stability masks a shift in the *location* of capital use: capital appears increasingly concentrated in the producing sector while its role in downstream production is understated.

This invariance breaks down when sector-S firms earn economic rents—from market power, scale economies, or intangible advantages. In that case, the shift toward cloud delivery raises measured operating surplus in sector S without a corresponding increase in measured capital input in sector C, and aggregate capital shares may be biased upward. The stability of aggregate factor shares may therefore be "correct for the wrong reason," obscuring both the extent of AI diffusion across the economy and the location at which AI-related capital services are actually used.

In plain terms: under traditional investment, firms acquire capital goods that appear on their balance sheets and in the national accounts. Under cloud delivery, firms acquire productive capability without taking ownership of the underlying assets. The accounts record investment where assets are owned, not where they are used—appropriate for questions of ownership, but not sufficient for questions of production, substitution, and diffusion.

Table 1 summarizes how the asset-boundary problem propagates through the two diagnostics. The TFP implications are discussed in Appendix A; the net effect on the factor-share

and user-cost diagnostics is a systematic downward bias in sector C—the direction that leads an analyst to understate AI diffusion in downstream production. Even if expenditures were correctly classified, however, the interpretation of user costs and capital deepening depends on accurate price measurement. The next section turns to that issue.

Table 1: **Measurement consequences of shadow capitalization (qualitative)**

Effect	Direction / implication
AI-capital income of sector-C firms understated	Factor shares in using sector biased downward; capital deepening in AI-using industries underestimated.
Operating surplus of sector-S firms overstated	Measured concentration of AI returns in producers; potential overstatement of aggregate capital income if sector S earns rents.
Omitted shadow stock overstates rental rate r_C^{AI}	User-cost calculations overstate r_C^{AI}, understating implied K^{AI} and biasing $\hat{\sigma}_C$ downward.
Misclassification increases intermediate consumption	Measured value added in sector C suppressed; measured TFP in AI-using industries biased downward during periods of rapid expansion of cloud AI expenditure, when the output channel dominates the offsetting input channel.

5 Measuring AI Capital Prices

The AI production stack introduced in Section 4 points to a measurement principle that distinguishes AI price measurement from standard IT-era practice. In earlier generations of digital capital, hardware and software were treated as separable layers—a relationship captured in the familiar adage that "what Intel giveth, Microsoft taketh away." Semiconductor price indices captured chip-level performance gains; software deflators were constructed independently from input costs or matched-model methods. The structure of AI production breaks this separability: capability improvements in frontier systems emerge from the interaction of architectural innovations in model design, scaling of training compute, and the accumulation of training data and fine-tuned weights. A price index capturing only the hardware or only the software component will miss the gains arising from the other layers and their interaction.

The standard response to rapid quality change—hedonic price methods—is also less applicable here. Hedonic methods require a stable set of measurable product characteristics; they work well where improvements are incremental and attributable to specific attributes, but poorly for AI systems, where progress is often discontinuous, reflecting architectural breakthroughs or scaling effects rather than smooth movement along a fixed attribute space.

The appropriate unit of measurement is therefore the integrated system—the deployed model. The capability-adjusted index described below deflates token prices by an aggregate benchmark score reflecting performance on tasks drawing on the full stack, capturing algorithmic progress, scale, and hardware improvement in a single empirical series: the cost per unit of productive AI output delivered by the frontier model.

Two features of the index connect it directly to the user cost framework of Section 3.2. First, the index measures AI capital *services* prices—what downstream firms pay per unit of constant-quality AI output—rather than an acquisition price P_S. As established in Section 3.2, in competitive equilibrium services prices and acquisition prices move in tandem, so the quality-adjusted rate of decline of the services price is a valid empirical proxy for $\dot{P}_S/P_S$. Second, under the shadow capital interpretation of Section 4.3, the cloud and API expenditures that underlie the index are precisely $r_C^{AI} \cdot K_{shadow}^{AI}$—the rental payments downstream firms make for AI capital services they do not own. The index therefore measures the directly observable counterpart to r_C^{AI} for downstream users, quality-adjusted for the capability of the services received.

The subsection that follows develops a proof-of-concept capability-adjusted price index for frontier AI systems. The exercise is necessarily preliminary: the time series covers March 2023 to March 2026, a period of only three years and twenty-one model-release observations. No definitive inference about long-run price trends can rest on a sample of this length, and the WLS trend estimates should be read as indicative rather than structural. That said, the approach is designed to be extensible. As the panel of frontier model releases accumulates, the methodology outlined here—deflating token prices by a standardized capability benchmark, classifying observations by market role, and estimating quality-adjusted price trends by weighted regression—could in principle be operationalized as a satellite price index for AI capital services within the national accounts framework.[3]

5.1 A Price Index for Frontier AI Models

We construct two unit-value price indices for frontier AI systems covering March 2023 to March 2026, using a dataset of 21 model-release observations drawn from the leading frontier AI providers (OpenAI, Anthropic, Google, DeepSeek, and xAI). The base period is the launch of GPT-4 in March 2023, at which both indices are set to 100. Using these observations, two WLS trend estimates are developed, as described below.

Token price index. Large language models charge for AI services on a per-token basis, where a token is a chunk of text (roughly 0.75 words). Token prices serve as a measure of

[3]The Epoch Capabilities Index that anchors the capability adjustment in this exercise is a private-sector product, maintained by Epoch AI as a research output rather than as an official statistical series. Translating it into a national accounts deflator would require agreement on methodology, governance, and update frequency between statistical agencies and the organisations—predominantly private AI labs and research institutes—that produce and maintain frontier model benchmarks. This is an institutional challenge as much as a methodological one, and one that the measurement community will need to address as AI capital becomes an increasingly significant component of national investment. The approach is also broadly consistent with the price measurement guidance in Sakai et al. (2025), which recommends constructing cloud computing price indices from the bill of a representative high-volume user—in effect, a unit-value index of the kind that underlies the token price series here.

"work done" because they directly reflect the computational resources, energy, and infrastructure needed to process inputs and generate outputs, rather than a flat, arbitrary rate. For each observation we compute a blended token price as $p_t = 0.5 \times p_t^{in} + 0.5 \times p_t^{out}$, where p^{in} and p^{out} are the published input and output prices per million tokens at model launch. The index value for each observation is $(p_t/p_{base}) \times 100$, where $p_{base} = \$45$ per million tokens (GPT-4 at launch, March 2023).

Capability-adjusted price index. Each model is assigned an Epoch Capabilities Index (ECI) score from the Epoch AI leaderboard. The ECI is a composite capability metric constructed using Item Response Theory, which aggregates performance across a standardised battery of reasoning, coding, and knowledge benchmarks into a single scale that remains comparable across models evaluated on different benchmark subsets (Lee and Emberson, 2025; Ho et al., 2025). The growth rate of the ECI accelerated at approximately April 2024, rising from 8 to 15 ECI points per year, consistent with the emergence of reasoning models (Edelman and Lee, 2025). The ECI scale is calibrated so that Claude 3.5 Sonnet (June 2024) ≈ 130 and GPT-5 $= 150$; scores for models not directly listed on the leaderboard are interpolated using these published capability growth slopes. The capability-adjusted index value for each observation is $(p_t/\text{ECI}_t) \,/\, (p_{base}/\text{ECI}_{base}) \times 100$, giving the price of a constant-capability token relative to the GPT-4 base.

Observation classification and weighting. The 21 observations are classified into four types on the basis of market role at the time of launch. *Mainstream frontier* observations ($w = 1.0$, $n = 13$) are models that represented the prevailing price-performance frontier for a sustained period following launch—the models that sector-C firms would actually deploy. *Premium launch spike* observations ($w = 0.25$, $n = 6$) are models that launched at prices substantially above the subsequent market rate before competition compressed margins: Claude 3 Opus, OpenAI o1, o1 (full release), Claude 4 Opus, Claude 4.1 Opus, and o3. These are included in the regression but downweighted because their launch prices reflect first-mover rent extraction rather than the equilibrium cost of frontier capability. Two observations are excluded from the trend regression entirely: DeepSeek R1 (January 2025, \$1.37/M blended at ECI ≈ 128) is excluded as a competition-shock event—its price reflected a deliberate market-entry strategy by a new entrant and represents a structural discontinuity rather than a point on the incumbent pricing trend; and GPT-5.2 Pro (\$94.50/M blended) is excluded as an extreme pro-tier outlier priced for a specialist segment outside the mainstream frontier.

Smoothed trends are estimated by weighted least squares (WLS) on log(index) regressed on time (measured in years from the base date), using the weights described above. An acceleration in both capability improvement and price decline is visible at approximately April 2024, coinciding with the GPT-4o launch; the reported trend rates span the full sample and should be understood as averages across the pre- and post-acceleration subperiods. The regressions are shown in table 2 below, and observation-level detail is reported in Appendix B of this paper.

Results. Figure 2 presents the two price indices together with their WLS trend lines. The fitted token price index declines 24.5 percent per year over the full sample period (95% CI: -46% to $+6\%$; $p = 0.096$; $R^2 = 0.15$; see appendix table A1): The blended price of a frontier

Table 2: **WLS Trend Estimates: Frontier AI Model Price Indices, March 2023 – March 2026**

	Token price index	Cap.-adjusted price index
A. Regression coefficients (dep. var.: log index)		
Constant	3.846^{***}	3.839^{***}
	(0.319)	(0.322)
t (years from base)	-0.281^{*}	-0.487^{***}
	(0.160)	(0.161)
Observations	19	19
R^2	0.154	0.350
Adj. R^2	0.104	0.312
F-statistic	3.10	9.16
p-value (F)	0.096	0.008
B. Implied annual rates of change		
Annual rate (%/yr)	−24.5	−38.5
95% CI (%/yr)	[−46.1, +5.7]	[−56.2, −13.7]
C. Quality-adjustment gap		
Gap (pp/yr)	−14.1	

Notes: WLS on log(index) regressed on time in years from base date (GPT-4 launch, March 2023). Observation weights: mainstream frontier $w = 1.0$ ($n = 13$); premium launch spike $w = 0.25$ ($n = 6$; Claude 3 Opus, o1, o1 full release, Claude 4 Opus, Claude 4.1 Opus, o3). Two observations excluded: DeepSeek R1 (competition shock, Jan 2025) and GPT-5.2 Pro (pro-tier outlier). Annual rates derived as $\hat{g} = e^{\hat{\beta}} - 1$. Standard errors in parentheses. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

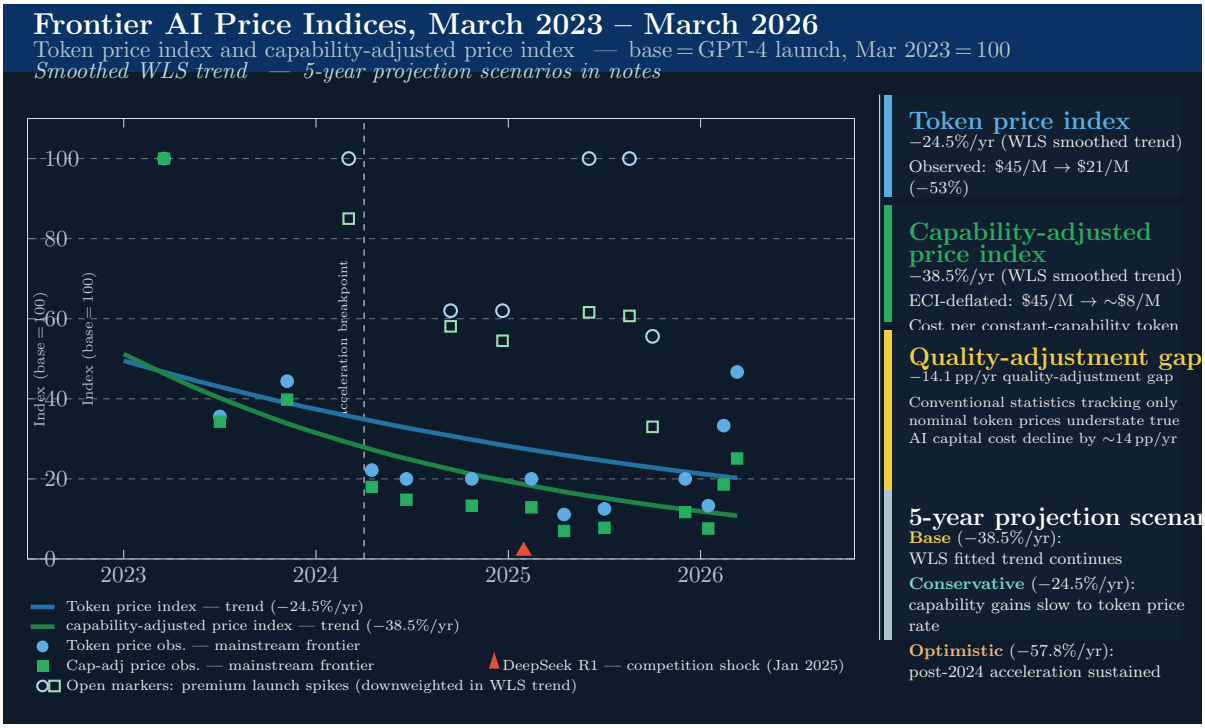
API token fell from $45/M at GPT-4 launch to $21/M by March 2026, a nominal decline of 53 percent in three years. The wide confidence interval and marginal significance reflect the sawtooth pattern generated by alternating mainstream and premium-priced releases; the downward direction is consistent across specifications.

The capability-adjusted price index declines at a substantially faster rate, 38.5 percent per year (95% CI: −56% to −14%; $p = 0.008$; $R^2 = 0.35$): Deflating by the ECI score, the cost per constant-capability token fell from $45/M to approximately $8/M over the same period. This trend is statistically significant and robust to the exclusion of individual observations.

The gap between the two trend rates—approximately 14 percentage points per year—is the empirical estimate of the quality-adjustment wedge between posted token prices and the price of a constant-capability unit of AI service. In the language of equations (9) and (10), what matters for substitution is the relative user cost w/r_C^{AI}, where $r_C^{AI} = P_S(\varrho + \delta - \pi)$ for downstream firms. The capability-adjusted index is the empirical measure of the quality-adjusted services price P_S relevant for calculating r_C^{AI}.

Projection scenarios. Three scenarios for the capability-adjusted price index through 2031 are presented in Figure 2. The *base scenario* continues the WLS fitted trend of −38.5%/yr. The *conservative scenario* (−24.5%/yr) assumes capability gains slow to match

Figure 2: Frontier AI Price Indices, March 2023 – March 2026



Sources: Epoch AI ECI leaderboard (epoch.ai/eci, Apr 2026); Edelman and Lee (2025); Lee and Emberson (2025); OpenAI / Anthropic / Google pricing pages. ECI scale: Claude 3.5 Sonnet ≈ 130; GPT-5 = 150. DeepSeek R1 (ECI 128, Jan 2025, $1.37/M blended) and GPT-5.2 Pro (pro-tier outlier, $94.50/M blended) excluded from trend fit. WLS on log(index); n=19 (13 TREND w=1.0, 6 PREMIUM w=0.25); R^2=0.15 (token), 0.35 (cap-adj). Projection scenarios anchored at trend-implied index at Mar 2026; not shown in chart.

the token price trend—that is, that further ECI improvement stalls and only nominal price competition continues. The *optimistic scenario* (−57.8%/yr, 1.5× the base rate) assumes the post-2024 acceleration is sustained. All three scenarios imply continued and substantial declines in the true user cost of AI capital over the projection horizon, with correspondingly strong pressure toward AI capital deepening in sector C. As noted above, the scenarios should be read as illustrative given the three-year estimation window.

The token and capability-adjusted indices address the price measurement problem for deployed AI models—the layer-two and layer-four components of the stack that dominate current AI expenditure. Let us now situate these expenditures in the context of the entire market for software.

5.2 The Software Market and AI Price Measurement

The capability-adjusted price index developed in Section 5 covers the leading-edge AI services that are the focus of this paper—the layer-two and layer-four components of the production stack that dominate current AI expenditure. But the mismeasurement problem extends across the entire software market, and the aggregate software deflators used in the national accounts blend together components with very different price dynamics. Understanding the

structure of the software market is therefore a prerequisite for assessing how large the total deflation bias is and where it is concentrated.

Figure 3 decomposes the global software market into four components that differ in their economic character, their growth trajectory, and their price measurement requirements. Two features of this decomposition require emphasis before turning to the numbers. First, these are *revenue* figures, not final demand: a substantial share of software expenditure is intermediate consumption by firms rather than final investment or consumption, and the composition of final versus intermediate demand differs across components—AI software at stack layers two and three is classified as almost entirely intermediate under current conventions, although as argued in Section 4.3 a portion of that expenditure is economically the purchase of capital services, while consumer-facing applications at stack layer one and traditional enterprise software at the legacy tier are more heavily weighted toward final demand. The revenue totals must therefore be interpreted with care when used to assess the aggregate investment and productivity implications. Second, the 2035 projections are base-case forecasts subject to substantial uncertainty; they are best read as illustrating the direction and pace of the transition rather than as point estimates.

With those caveats noted, the scale of the transition is striking. In 2025, the market totals approximately $1.24 trillion, with pure legacy software still the dominant component at 83 percent of the total. The three AI components—AI software and services (L1–L4), AI pricing uplift in legacy platforms, and agentic AI embedded in enterprise software—together account for 17 percent today but are projected to account for the entire market by 2035 as the legacy residual migrates into AI-enabled segments. The AI software stack component grows approximately 27-fold over the decade, rising from $72 billion to roughly $1.9 trillion. The agentic AI component, modest at $24 billion today, is projected to reach $482 billion by 2035—the largest single source of absolute growth. The AI pricing uplift component is a transitional phenomenon: large now as vendors extract premiums for AI-featured tiers, but projected to shrink as AI capabilities become baseline and pricing normalizes.[4]

[4]The pure legacy share declining to zero by 2035 does not imply that the underlying platforms exit the market. It reflects a reclassification of revenue across segments as AI content becomes universal: software that today carries no AI features is projected to migrate into the pricing uplift, agentic AI, or AI stack software segments as vendors embed AI capabilities into their products. The value of the platform itself need not fall to zero; rather, it ceases to be recorded in the pure legacy category once an AI tier or AI-agent capability is introduced. Similarly, the pricing uplift segment declines toward 2035 not because the AI premium disappears from the market but because, as AI features become standard, the premium is absorbed into the base subscription price and the revenue is reclassified into the L1–L4 AI stack software segment. The table tracks AI-attributable revenue by economic type, not the survival of specific vendors or platforms.

Figure 3: **The Global Software Market: Composition, 2025 and 2035E**

Component	Definition	2025 Revenue	2025 % of SW	2035E Revenue	2035E % of SW
L1–L4 AI Software Stack	Revenue of AI stack vendors across all four stack layers: applications, foundation models, platforms & tooling, and hyperscale cloud AI increment.	**$72B**	6%	**~$1.9T**	79%
AI Price Uplift in Legacy Software	The premium enterprises pay for AI features embedded in existing platforms — e.g. Microsoft 365 Copilot add-on pricing, Salesforce Einstein tiers, SAP Business AI surcharges.	**$110B**	9%	**$40B**	2%
Agentic AI in Legacy Platforms	AI agents and autonomous workflows embedded in established enterprise platforms: Salesforce Agentforce, ServiceNow Now Assist, SAP Joule, Microsoft Copilot Studio.	**$24B**	2%	**$482B**	20%
Pure Legacy Software (incl. traditional ML)	Residual software revenue: on-premise ERP, legacy CRM, traditional productivity suites, and industry-specific applications, calculated as the residual after the three AI components are subtracted.	**$1,034B**	83%	**$0**	0%
Total software market		**$1,240B**	100%	**$2,434B**	100%

Sources: Total software market: Gartner IT Spending Forecast (Feb 2026). AI price uplift: Gartner (~9% of IT budgets, 2025–26). Agentic AI in legacy: Gartner (2% of enterprise app SW 2025; 30% by 2035). L1–L4 AI software stack: own estimates (ownership basis ≈ global total on residence basis). Pure legacy = residual (incl. est. $150–200B traditional/predictive ML embedded in enterprise platforms — not separately identified in source data). 2035 figures are base-case forecasts.

The aggregate software deflator in national accounts should reflect a weighted average of price change for these four components. As the weights shift toward AI stack software/services and AI agentic software over the projection horizon, the mismeasurement problem identified in this paper becomes increasingly central to the measurement of economy-wide investment and productivity. Even if the projections prove optimistic about the pace of transition, the direction is unambiguous—and it argues for urgency on the part of official statisticians. The window in which quality-adjusted price indices for AI software can be developed and institutionalised before the mismeasurement becomes large enough to materially distort the national accounts is narrow. The remainder of this section sets out what such indices would require, component by component.

Price index approaches by component

The four components of the software market require four distinct approaches to price measurement, ranging from the well-established to the methodologically nascent.

Legacy software (incl. traditional ML). The residual legacy component—on-premise ERP, traditional productivity suites, industry-specific applications—is the component for which existing national accounts practice is best suited. The Bureau of Economic Analysis (BEA) software price deflator, constructed from input costs and matched-model methods, provides a serviceable approximation for this component, though it is known to understate

quality improvements even here (Byrne and Corrado, 2017, 2020; Fleming, 2025). For the purposes of aggregate quality-adjusted price measurement, the BEA deflator can be applied to this component as a lower bound on true price decline.

AI pricing uplift in legacy platforms. The AI pricing uplift component—the premium enterprises pay for AI-featured tiers of existing platforms—is conceptually a margin premium rather than a quality-adjusted price change in the usual sense. The appropriate measurement approach is a matched-model index comparing the price of the same platform with and without AI features at successive points in time. Microsoft 365 provides the cleanest natural experiment: published base subscription and Copilot add-on prices are available at monthly intervals and allow direct construction of an uplift series. Salesforce and SAP offer comparable tier-pricing data. As competition compresses AI pricing premiums into the base subscription—the dynamic projected in Figure 3—this component's contribution to aggregate price change will decline toward zero, making its measurement tractable.

Although the AI premium is identified in this approach, tractability comes at a cost. The AI-featured tier itself improves over time—Microsoft 365 Copilot in 2026 offers substantially greater capability than it did at launch in late 2023—yet the matched-model index will record only the change in the *price differential* between tiers, not the improvement in the capability delivered at the AI tier price. In this respect, the matched-model uplift index shares the limitation of the nominal token price index for AI-native software: it captures price compression but not quality improvement within the tier. A fully quality-adjusted index for this component would require deflating the uplift premium by a measure of AI feature capability over time—analogous to the ECI adjustment for L1–L4 software—but no adequate data currently exist for this purpose. The matched-model approach should therefore be understood as a lower bound on true price decline for the AI pricing uplift component, and its contribution to the aggregate deflation bias will be understated to the extent that intra-tier capability improvement is substantial.

AI software and services (L1–L4, stack layers 1–4). The capability-adjusted price index developed in Section 5 is designed for this component. By deflating a unit-value index of frontier AI services prices by the Epoch Capabilities Index (ECI), the index captures quality change arising from capability improvement across all layers of the production stack—architectural advances, scaling, and hardware improvement—in a single empirical series. The rapid rate of decline in the capability-adjusted index—for which there is no counterpart in official statistics—suggests that prevailing software price deflators substantially understate the true rate of price decline for this component. As the L1–L4 share of the software market rises from 6 percent today toward 79 percent by 2035, this gap will increasingly dominate the aggregate software deflation bias.

Agentic AI in legacy platforms. The agentic AI component poses the most demanding measurement challenge and marks the leading edge of the price index research agenda. Agentic AI is priced in qualitatively new units—per conversation, per task, per outcome—rather than per seat or per token. Salesforce Agentforce, for example, prices at a fixed rate per completed conversation rather than a software license fee. This per-task pricing

structure is in principle well-suited to price index construction: it provides a natural unit of output analogous to the token for AI-software stack APIs. A capability-adjusted price index for agentic AI would deflate the cost per completed task by a measure of task-completion capability—the fraction of real-world tasks an agent can complete autonomously without human intervention. Emerging benchmarks such as METR's time-horizon metric and SWE-bench for software engineering tasks provide candidate capability measures,[5] though none yet offers the breadth and stability required for a production price index. Constructing such an index—a task-completion-adjusted price for agentic work, methodologically consistent with the AI software stack index of Section 5 but anchored to a task-completion benchmark rather than the ECI—is the natural next step in this measurement program and is left for future work.

Aggregate implications

In 2025 the three AI segments account for 17 percent of the software market; by 2035, under the base-case projection, they account for the entire market. The mismeasurement is concentrated in the AI software stack and AI agentic segments, where capability improvement is fastest and existing deflators and methods are least applicable. As their combined weight rises, the gap between quality-adjusted software prices and official measures will widen, generating increasingly large downward biases in measured real investment, real capital services, and productivity—distortions that will reach macroeconomic scale well before the end of the projection horizon. The proof-of-concept AI price index in Section 5 is a step toward addressing this; the broader program sketched here defines the full scope of what remains to be done.

A closely related dimension of the price measurement problem, left for future work, concerns data assets. The 2025 SNA now formally recognizes data as a produced capital good, but the value of a training dataset depends not on its production cost but on its informational content and the AI capabilities available to exploit it: a dataset marginally useful in 2018 may be enormously valuable in 2025 not because the data changed, but because the models trained on it improved. Neither this appreciation dynamic nor its reverse—rapid depreciation if competitors develop synthetic data generation methods that replicate the dataset's informational content—is captured by cost-based deflators, and no adequate methodology currently exists for this class of asset.

[5] METR's time-horizon metric measures autonomous agent capability as the length of task—in hours of human effort—that an agent can complete end-to-end with a given success rate; a rising time horizon indicates that agents can complete longer, more complex sequences of actions without human intervention. SWE-bench is a software engineering benchmark that tests whether agents can autonomously resolve real GitHub issues in open-source repositories; the pass rate provides a direct measure of task-completion capability on a well-defined, verifiable class of knowledge-work tasks. Both produce scalar capability scores that could in principle anchor a quality adjustment analogous to the ECI in the AI-software stack price index.

6 Descriptive Evidence

Even though we are in early days of an AI era and despite the measurement issues discussed in Sections 4.3 and 5, pattens in existing digital capital factor shares warrant scrutiny. This is especially true for the United States in light of the findings in Bontadini et al. (2026 forthcoming), where the share of software capital—defined as software products and software/AI R&D—was found to be rising sharply beginning in 2017. Panel 1 of Figure 4 shows the asset income shares of three primary digital assets: ICT equipment, ICT equipment and software (ICT), and Software/AI R&D. Note that software/AI R&D includes computing, data, and energy input costs.

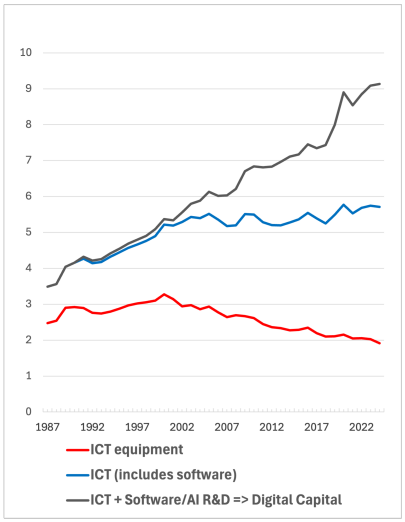
The next two panels of Figure 4 decompose this aggregate pattern to shed light on two questions: how much of the rise is concentrated in the software-producing sector, and what happens when downstream cloud expenditures are partially recaptured as implicit capital payments?

Panel 2 addresses the first question by separating the non-farm business aggregate into the software-producing sector and the rest of the economy. The comparison reveals that a substantial share of the rise visible in Panel 1 is concentrated in the producing sector itself—reflecting both rapid technological progress and the recursive capital deepening that characterises firms at the frontier of AI development. Outside the software-producing sector, the digital capital share rises only modestly over the period, and the trajectory is flatter. The partial exception is a gradual increase driven by the conduct of software R&D in manufacturing and the R&D services industry, which points to slow but detectable diffusion of digital capital intensity into downstream sectors. The overall picture from Panel 2 is therefore one of highly uneven diffusion: the aggregate rise in Panel 1 is, for now, predominantly a phenomenon of the producing sector rather than a broad-based shift in how the rest of the economy employs digital capital.

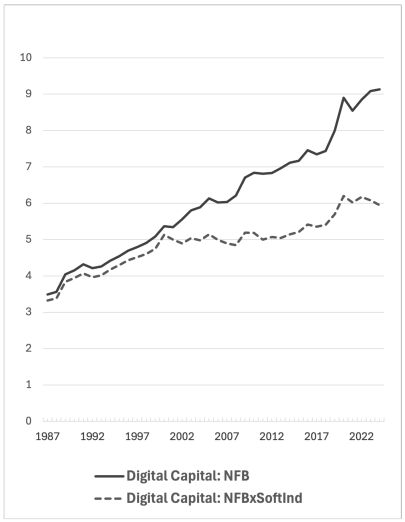
The producing/using decomposition in Panel 2 also speaks directly to the concentration problem noted in Section 2: by isolating software-producing industries, it disentangles frontier-firm effects from the broader diffusion signal, providing a cleaner read on whether AI capital deepening is propagating beyond the handful of firms that dominate aggregate measures. The picture that emerges—modest diffusion outside the producing sector, with slow penetration via software R&D into manufacturing and R&D services (not specifically shown)—suggests that concentration is not merely a statistical artifact but a substantive feature of where AI capital accumulation currently stands.

Panel 3 addresses the second question by incorporating an adjustment for cloud services payments by downstream industries. Because layer-two expenditures—cloud compute, data pipeline access, and foundational model subscriptions—are currently expensed as current costs rather than capitalized, they do not appear in measured capital income flows; the capital stock serving as the implicit denominator of r_C^{AI} is understated, and the measured rental share in downstream sectors is therefore depressed relative to its true economic counterpart. Panel 3 adds these payments back as a shadow capital service flow. The adjustment elevates the trajectory of digital asset rental payments in the non-producing sector appreciably, narrowing the gap with the producing sector visible in Panel 2. Two important caveats attach to this adjustment: first, no attempt is made to separate payments for AI services from payments for generic computing infrastructure, owing to data limitations, so the ad-

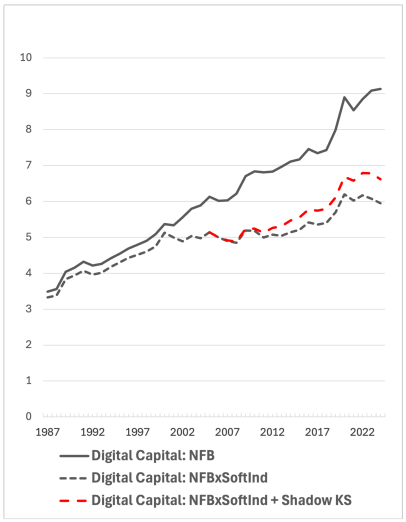
Figure 4: **Nonfarm Business Sector Digital Asset Income Shares, 1987 to 2024**



(a) Panel 1: Digital capital rental shares, NFB aggregate



(b) Panel 2: Decomposition: Software-producing vs. rest of economy (aka downstream sector)



(c) Panel 3: Downstream sector with shadow capital services adjustment

Source: Authors' elaboration of BLS US productivity estimates and software R&D investment developed from US BERD surveys. Shadow capital services adjustment is developed from data on cloud services from *Mordor Intelligence*, adjusted to exclude software-producing industries using information derived from BEA's supply-use tables.

justment likely captures a mixture of AI-specific and non-AI cloud expenditure; second, the adjustment is a lower bound in the sense that it recaptures only observed cloud payments, not the full shadow value of uncapitalized foundational model investment. Taken together, Panels 2 and 3 suggest that the diffusion of digital capital into downstream sectors is further advanced than the official capital data imply, but that confirming this empirically requires the capitalization reforms and price-index improvements discussed in the preceding sections.

7 Empirical Results and Interpretation

The analysis so far has focused on a two-factor setting and drawn on the task approach only briefly. We now turn to the multi-input case, presenting estimates of how the price changes modelled above translate into shifts in factor shares across a broader set of inputs. Doing so requires both a richer empirical framework and a careful mapping of our estimates onto the task literature.

7.1 Tasks, technology and estimating the elasticity of substitution

In the task literature, production involves two distinct stages. In the first, firms carry out a set of individual tasks. In the second, they combine the outputs of those tasks into final goods. To fix ideas, suppose the final good is a video of dancing. This requires two

tasks: 1) performing the dancing itself, and 2) writing the software code needed to turn the live performance into broadcast content. For the first task, the firm can draw on labor, capital, or both — human dancers performing in front of a software-generated backdrop, for instance. For the second, the code can be written by human programmers, by software tools, or by some combination of the two. Assume then that each task is associated with a unit cost function that gives the minimum cost of accomplishing it, given input prices and the available technology. The substitution possibilities within each task cost function are governed by σ_F, the elasticity of substitution between capital and labor *within* a task. When tasks are combined into final output (the video), a second elasticity comes into play: σ_T, the elasticity of substitution *across* tasks. The aggregate cost function therefore depends on *both* σ_F and σ_T. Without task-level data, estimation of that aggregate cost function cannot recover the two elasticities separately; what is identified is some combination of them, the precise nature of which depends on the underlying task structure.

Common in the literature is a special case. Suppose the elasticity of substitution between factors *within* a task, $\sigma_F = \infty$, therefore substitution is infinite. Under this assumption, each task is allocated entirely to whichever factor holds a comparative advantage at prevailing factor prices — for example, the dancing might be performed exclusively by humans, the coding handled entirely by software capital.

In this case the aggregate unit cost of production reduces to a weighted average of the cost of capital tasks and the cost of labor tasks, with weights equal to the share of tasks performed by each factor. The only operative elasticity is therefore σ_T: since every task belongs exclusively to one factor, task boundaries and factor boundaries coincide, and the two-level cost structure collapses to a single one. When within-task substitution is infinite, there is no distinction between a task and the factor that performs it, and the cross-task elasticity alone governs aggregate factor demand.

This special case is particularly useful for illustrating the bottlenecks point. Even if capital becomes extraordinarily cheap and displaces labor entirely *within* every task it enters, the scope for aggregate output growth still depends on how easily those tasks substitute for one another.

The labor share will decline only to the extent that the tasks remaining in human hands are substitutable for those taken over by capital. If the remaining tasks are complements — if, say, coding can be entirely automated but the product still requires human performance — then the cross-task elasticity governs the overall extent of labor displacement, independently of how complete the within-task substitution has become. This is precisely the weak-links mechanism of Jones and Tonetti (2026): the binding constraint on automation operates at the level of task combinations, not at the level of individual factor substitution.

In this special case then, estimation of a cost function might return the cross-task elasticity. But there are two reasons to be cautious. The first is that it would be a very strong assumption.[6]

Second, as we have set out above, production requires organisational capital. In the language of the task literature, *combining the tasks is a task*. The neat separation of within

[6]Or perhaps not if one defined tasks narrowly enough. The task of giving a lecture using slides would seem to involve capital and labor. But the task of projecting the slides is capital only and of giving personal feedback human only.

and across tasks is broken. And changes in the combinations of tasks (bricks and mortar retail versus online) and the technology of such combinations (insourcing versus outsourcing) are likely to have occurred.

The role of Z_C in the above is then that coordination is itself a task, one that must be accomplished before individual task outputs can be assembled into final goods. Z_C represents the expenditure on this coordination task.

7.2 Model

To estimate the evolution of input factor shares we start with a cost function of the usual translog type. We emphasise again that the estimated parameters of the cost function that follow are not structural primitives. Instead they are a reduced-form, so that our factor-share simulations illustrate the future factor-share adjustment under various price trajectories that are consistent with the implied elasticities estimated on current data.

For inputs i and j, the translog cost function is:

$$\ln C = \alpha_0 + \sum_i \alpha_i \ln(p_i) + \frac{1}{2}\sum_i \sum_j \gamma_{ij} \ln(p_i) \ln(p_j) + \sum_i \beta_i \cdot TECH_i \cdot \ln(p_i), \quad \text{with } \gamma_{ij} = \gamma_{ji}. \tag{17}$$

Applying Shephard's lemma yields cost share equations linking the share of each input to input prices:

$$s_i = \alpha_i + \beta_i \, TECH_i + \sum_j \gamma_{ij} \ln(p_j). \tag{18}$$

Estimation of the share equations yields the price elasticities of factor demand and the associated elasticities of substitution. These are summarised in the following system, where all expressions depend only on the estimated coefficients γ_{ij} and the observed cost shares s_i:

$$\eta_{ij} = \frac{\gamma_{ij}}{s_i} + s_j \;\; (i \neq j), \quad \eta_{ii} = \frac{\gamma_{ii}}{s_i} + s_i - 1, \quad AES_{ij} = \frac{\eta_{ij}}{s_j}, \quad MES_{ij} = \eta_{ij} - \eta_{jj}, \tag{19}$$

where η_{ij} is the cross-price elasticity of demand for input i with respect to the price of input j, AES_{ij} is the Allen–Uzawa elasticity of substitution, and MES_{ij} is the Morishima elasticity of substitution. Note that MES_{ij} is asymmetric in general, since the own-price elasticities η_{ii} and η_{jj} are not equal.

7.2.1 The relevant elasticities with many inputs

The system in (19) yields several elasticity concepts, and the choice between them matters for interpretation. The Allen elasticity has been the conventional choice in the literature, and for good reason in the two-factor case: whether it exceeds or falls short of unity is sufficient to determine how the relative cost shares of the two inputs respond to a price change, holding output fixed. In the multi-input case, however, this convenient property breaks down, as we discuss below.

In the multi-input case, as Blackorby and Russell (1989) show, the Allen elasticity is not a behavioral measure of substitution. The reason is this: the cost function gives the

minimum cost of producing a given level of output, and when the price of one input changes the firm reoptimizes over *all* inputs simultaneously. If the price of input 1 rises and the firm substitutes toward input 2, the optimal response will generally also involve an adjustment in input 3. The Allen elasticity between inputs 1 and 2, by implicitly holding all other inputs fixed, therefore fails to trace substitution along the cost function and cannot be given a clean behavioral interpretation.

The appropriate elasticity in this case is the Morishima elasticity, which measures how the relative use of two inputs changes when the price of one of them changes, holding all other prices constant. Its key property is that $MES_{ij} \gtrless 1$ is a sufficient statistic for whether the relative cost share of i with respect to j rises or falls in response to a change in p_j. To see this, note that for any two inputs i and j, the change in their relative cost shares when p_j changes is:

$$\frac{\partial \ln(s_i/s_j)}{\partial \ln p_j} = \frac{1}{s_i}\frac{\partial s_i}{\partial \ln p_j} - \frac{1}{s_j}\frac{\partial s_j}{\partial \ln p_j} = \frac{\gamma_{ij}}{s_i} - \frac{\gamma_{jj}}{s_j}. \tag{20}$$

Using the definition of the Morishima elasticity, this becomes:

$$\frac{\partial \ln(s_i/s_j)}{\partial \ln p_j} = MES_{ij} - 1, \tag{21}$$

so that the relative share of i rises or falls as $MES_{ij} \gtrless 1$. Note also that MES_{ij} is asymmetric — $MES_{ij} \neq MES_{ji}$ in general — since the adjustment of all other inputs that accompanies a change in p_j differs from the adjustment that accompanies a change in p_i. Asymmetry is therefore a feature: it correctly reflects that replacing input j with input i need not be as easy as replacing input i with input j.

The Morishima elasticity is therefore the appropriate summary statistic for assessing whether AI-driven price declines will raise or lower the labor share in a multi-input setting, and it is this measure we report below.

7.3 The transition to econometric work

Our dataset is drawn from the EU KLEMS & INTANProd cross-country industry-level database (LUISS 2025). To sidestep the complications of adjustment dynamics—exacerbated by the financial crisis and COVID-19—we pool annual observations into two cross-sections, 1998–2007 and 2011–2019, replacing annual data with period averages. We work throughout in first differences, since cross-sectional variation in rental prices is uninformative and the meaningful signal is in how prices change over time. The labour price is constructed as growth in compensation per hour less a labour composition index controlling for changing worker skills. We impose symmetry and homogeneity restrictions, normalize on the labour price, and estimate the system by Seemingly Unrelated Regression (SURE), with the labour equation omitted to avoid singularity and its coefficients recovered from the cross-equation restrictions. To guard against outliers we trim the baseline dataset.[7] Our estimating equation

[7] Outlier detection is robust to choice of numéraire: we demean within each observation and remove the 1% of highest deviations from the mean.

system is:

$$\Delta s_i^{C,I,T} = \beta_i + \sum_j \gamma_{ij} \, \Delta \ln \left(\frac{p_j}{p_L} \right)^{C,I,T}, \qquad i = 1 \ldots j, \; i \neq L, \tag{22}$$

where β_i captures input-i-biased technical progress. We carry out robustness checks for sensitivity to outliers, pooling periods, and variation in β.

7.4 Share simulation

Armed with these estimated coefficients we can then simulate the future trend of the factor shares. To do this, we start with the level of shares in the second pooled period, 2011–2019. We then apply the estimated change-in-share equation, simulating a new share for the next year and then updating that new share for 40 periods. To apply the change equation we need to assume values for input-specific technical change, β, and the evolution of relative prices.

Two points of interpretation bear emphasis. First, if input-specific technical change $\hat{\beta}_i$ is non-zero, shares will evolve even under zero price change; any hypothetical price change should therefore be read as a marginal effect on top of the path implied by biased technical change alone. Second, the coefficients $\hat{\gamma}_{ij}$ are held fixed across the simulation, but the Morishima elasticities of substitution are not: because they depend on the input shares themselves, they vary endogenously as shares evolve over the simulation horizon.

In sum, our simulated shares are constructed as:

$$s_{i,t+1}^{C,I} = s_{i,t}^{C,I} + \hat{\beta}_i + \sum_j \hat{\gamma}_{ij} \, \Delta \ln p_j^*, \qquad i = 1 \ldots j, \tag{23}$$

where if a share falls to zero, we renormalize the vector so that the five shares continue to sum to one in every simulation year. Notice that in this approach we do not keep the same elasticities of substitution over time, but the same coefficients. In principle, since the elasticities of substitution depend on the shares, they will vary over time as the shares change.

7.5 Data and sample choices

Regarding our data we have the following. First, we add a period dummy for our pooled results across the two periods in case the bias of technical change has changed.

Second, we use one-letter industries in NACE Rev2 classification since more narrow industry data is likely noisy. That said, we drop industry B (mining), D–E (gas/electricity), and F (construction), since they are also very noisy as they have an omitted input, namely land. We also drop industry J since the pattern of substitution might vary in the software industry itself. This has an additional benefit: software spending in the course of R&D is counted as R&D spend in the US but as software spend in Europe. To the extent that such spending is confined to industry J, dropping this will help alleviate this problem.

This then gives 9 industries in 12 countries, with pooling over the years 1998–2007 and

2011–2019.[8]

7.6 Results

7.6.1 Capital and labor

To fix ideas, we begin by estimating the elasticity of substitution with just two inputs. Table 3 reports the results. In Panel A, combining all capital inputs with labor yields own-price elasticities of −1.04 for capital and −0.28 for labor, and a cross-price elasticity of substitution of 0.54. Panel B groups "ICT" capital (software, hardware and communications equipment) — against all other inputs (labor and non-ICT); in Panel C we contrast software capital alone against all other inputs. In all three cases the elasticity of substitution is well below unity, consistent with the weak-links structure set out in Section 7.1.

Table 3: **Estimated elasticities of substitution, two inputs**

Panel A: Capital and labor		
	Capital	labor
Capital	−1.04	0.54
labor	0.54	−0.28
Panel B: ICT capital and other inputs		
	ICT capital	Other inputs
ICT capital	−15.31	0.82
Other inputs	0.82	−0.04
Panel C: Software and other inputs		
	Software	Other inputs
Software	−17.31	0.39
Other inputs	0.39	−0.01

Notes: The table reports Allen elasticities of substitution estimated from a two-input translog cost function for three alternative input groupings: capital and labor; ICT capital (software, communications equipment, and computer hardware) and all other inputs; and software and all other inputs. Off-diagonal elements are cross-price elasticities of substitution; diagonal elements are own-price elasticities.

[8]To be precise, the countries are: Austria, Czechia, Denmark, Finland, France, Italy, Germany, the Netherlands, Spain, Sweden, UK, and US. Industries include: C = Manufacturing, G = Wholesale and retail trade; repair of motor vehicles and motorcycles, H = Transportation and storage, I = Accommodation and food service activities, K = Financial and insurance activities, M = Professional, scientific and technical activities, N = Administrative and support service activities, R = Arts, entertainment and recreation, and S = Other service activities.

7.6.2 Four capital inputs and labor

Table 4 reports results for four capital inputs and labor, estimated on the trimmed sample. In the table ""ICT" capital is software, hardware and communications equipment. Panel A reports the matrix of Allen elasticities of substitution. All own-price elasticities are negative. The cross-price elasticities are positive in most cases, indicating substitutability across input pairs, with particularly large values between ICT and intangibles excluding software (1.74) and between ICT and tangible non-ICT capital (0.76).

Table 4: **Estimated elasticities of substitution, multiple inputs**

	ICT	TangNICT	IntangXsw	labor
Panel A: Allen elasticity of substitution				
ICT	−13.62	0.76	1.74	0.36
Tang. non-ICT	0.76	−2.19	1.16	0.35
Intang. excl. SW	1.74	1.16	−5.79	0.61
labor	0.36	0.35	0.61	−0.23
Panel B: Cross-price elasticities				
ICT	−0.58	0.14	0.20	0.24
Tang. non-ICT	0.03	−0.40	0.14	0.23
Intang. excl. SW	0.07	0.21	−0.68	0.40
labor	0.02	0.06	0.07	−0.15
Panel C: Morishima elasticity of substitution				
ICT	·	0.54	0.89	0.39
Tang. non-ICT	0.61	·	0.82	0.38
Intang. excl. SW	0.65	0.61	·	0.55
labor	0.59	0.46	0.75	·

Notes: Panel A reports Allen–Uzawa elasticities of substitution; Panel B reports conditional cross-price elasticities of factor demand; Panel C reports Morishima elasticities of substitution. In Panel A, off-diagonal elements are cross elasticities and diagonal elements are own elasticities. In Panel C, diagonal entries are undefined and marked with a dot. A Morishima elasticity above (below) unity implies that the relative row cost share rises (falls) when the column input price increases. All estimates are obtained from a translog cost function estimated by SURE on pooled cross-sections 1998–2007 and 2011–2019.

Panel B reports the conditional cross-price elasticities of factor demand. These confirm the pattern from Panel A: substitution is widespread, with the largest responses found between ICT and intangibles excluding software, and between intangibles and labor. Own-price elasticities are again negative throughout, with ICT exhibiting the largest response (−0.58) and labor the smallest (−0.15).

Panel C reports the Morishima elasticities of substitution, which are the relevant measure for the multi-input case as discussed in Section 7.2.1. All results are below unity, confirming of any input rises, the ratio of shares of other inputs adjusts less than proportionally. The

highest Morishima elasticity is between intangibles excluding software and labor (0.75), and the lowest is between tangible non-ICT and labor (0.46). Taken together, the results are consistent with the weak-links structure: no single input can be cheaply and fully replaced by another, and the substitution impulse from any price change is only partially transmitted to factor proportions.

7.6.3 Share projections

We use the estimated coefficients to simulate the evolution of factor shares under three scenarios, whose assumed price changes are set out in Table 5. The first scenario uses observed average price changes from the second pooled period (2011–2019) and serves as the baseline: any evolution of shares under this scenario reflects the estimated input-biased technical change, $\hat{\beta}_i$, rather than a counterfactual price path. The second scenario holds all prices at their observed values except for ICT, whose price is set exogenously to -10% per year — more than ten times the observed rate of ICT price decline in the data. The third scenario goes further: ICT prices fall at -20% per year and intangible (excl. software) prices also fall at -10% per year, while labor and tangible nonICT prices remain at their observed values. This third scenario is designed to capture the joint effect of rapidly falling AI-related capital prices and a relaxation of the complementary organizational investment constraint, in the spirit of the Z_C mechanism discussed in Section 7.1.

Table 5: **Price changes under simulation scenarios, percent**

	Observed price changes	Exogenous ICT prices	Exogenous ICT and intang. prices
ICT	−0.6	−10.0	−20.0
Tang. non-ICT	2.2	2.2	2.2
Intang. excl. SW	1.9	1.9	−10.0
labor	1.8	1.8	1.8

Notes: The table reports annual log price changes, expressed as percentages, used in the three simulation scenarios. The first column reproduces observed average price changes from the second pooled period (2011–2019) and serves as the baseline. The second column sets the ICT price change exogenously to -10% per year, holding all other prices at their observed values. The third column sets the ICT price change to -20% per year and additionally allows intangible (excl. software) prices to fall at -10% per year, with labor and tangible nonICT prices unchanged.

The simulated factor shares are set out in Figure 5. Under averaged price changes, the labor share declines slightly, non-software intangibles rise, and other shares are broadly flat—largely reflecting the underlying constants in the estimated equations, which capture technical progress, outsourcing, and other forces inducing cost-share changes net of price effects over the sample period.

When ICT prices are instead projected to fall at 10% per period, the ICT share declines sharply, reaching zero within 30 periods—reminiscent of what Jones and Tonetti (2026) describe for agriculture, which becomes so productive that its economy-wide share shrinks

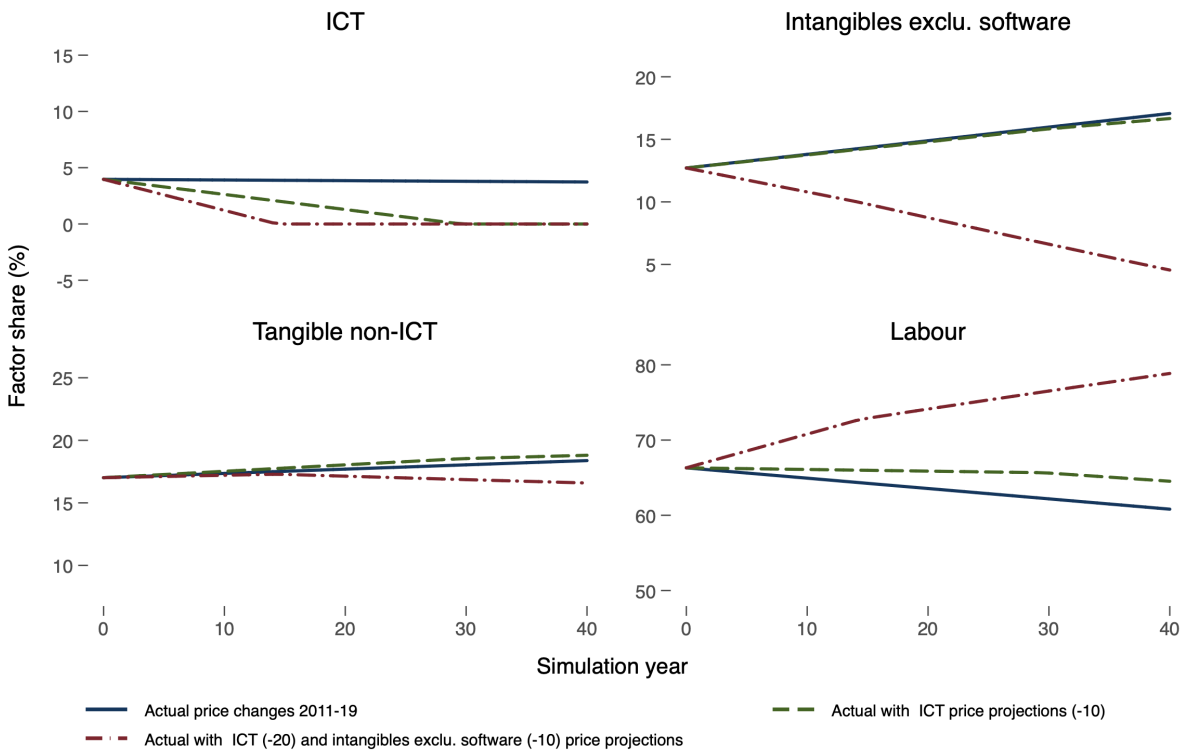


Figure 5: **Simulated factor shares under three price scenarios**

toward zero. Crucially, the labor share *rises*. This follows directly from the Morishima elasticities being uniformly below unity: when an input becomes relatively cheaper and $M_{ij} < 1$, the input's share increases. ICT does not displace labor sufficiently to reduce its cost share; the quantity of labor demanded does not fall proportionally to the ICT price decline.

When both ICT and non-software intangible prices fall jointly, the shares of both these inputs decline, while the labor share rises more substantially than under the ICT-only scenario.

Taken together, our estimates support the "bottlenecks" conjecture: factor shares are ultimately governed by limited elasticity of substitution *across* tasks, even if within-task substitution is considerable. The important caveat is that what these estimates cannot capture is a change in the technology of *combining* tasks. The estimated elasticities vary with input shares, but in a deeper sense AI may alter the meta-technology of task combination itself. If AI automates tasks previously requiring human judgement, the Morishima elasticities could shift above unity, reversing the labor-friendly results simulated here and driving the labor share down. Conversely, if AI primarily lowers the cost of producing intangible assets while leaving human-intangible complementarities intact, the simulations remain a reasonable guide to the direction of factor-share changes, even if not their magnitude. The Amazon example illustrates the more radical possibility: new organizational forms can reshape the entire technology of task combination, changing not just input prices but the

production possibility set itself. Static elasticity estimates, by construction, cannot capture such structural change—which is precisely why the measurement agenda set out above remains indispensable.

8 Conclusion

AI can be measured through the joint behavior of factor shares and capital costs. The two-sector framework organizes the key mechanisms: upstream productivity growth in AI production lowers the price of AI capital; through the user cost, this decline induces downstream substitution; and because AI also raises the productivity of the upstream sector itself, the mechanism is self-amplifying.

Three systematic measurement challenges qualify this picture. The cloud computing asset boundary problem—in which layer-two enabling infrastructure is expensed rather than capitalized—implies that measured capital deepening understates true AI investment and that measured user costs overstate the true economic cost of AI capital, biasing estimated elasticities of substitution downward. The quality-change problem for AI capital asset prices amplifies this bias further. And the concentration of AI investment in a small number of frontier firms means that aggregate diagnostics may obscure broad-based substitution. Progress on all three fronts requires better empirical foundations: quality-adjusted price indices for AI systems, capitalization methods for cloud and foundational model expenditures, and firm-level data linking AI investment to factor shares and market structure. These are the priorities that follow most directly from the framework developed here.

Having constructed a proof-of-concept price index for leading edge AI—one suggesting substantial declines in the rental price of the AI software/hardware bundle, and possibly in intangible capital more broadly—we offer a preliminary empirical indication of what such declines might imply for factor input shares. We stress that this empirical work is necessarily conducted with existing ICT capital measures rather than the AI-specific, quality-adjusted data the framework calls for; the estimates are therefore best read as an order-of-magnitude diagnostic rather than a structural characterization of AI's factor-market effects.

The analysis draws on the modern literature on tasks and input-augmenting technical change (Autor et al., 2003; Acemoglu and Restrepo, 2018a, 2019). The central intuition is that substitution within tasks may be considerable, but the aggregate output effect depends critically on substitution *across* tasks—and the prevailing theoretical conjecture is that cross-task substitution is limited, a "bottlenecks" view. That conjecture motivates using factor input shares as a diagnostic: limited substitution implies that a fall in an input's price reduces rather than raises its rental share.

Our empirical contribution is novel in two respects: rather than treating capital as a single aggregate combined with labor—the conventional approach—we look at digital capital as software and compute and then at intangibles (other than software capital) separately, better targeting the major asset groups whose prices AI will affect; and we estimate Morishima elasticities, which are the conceptually appropriate measure in the multi-input case. Our findings support the bottlenecks story. Morishima elasticities are uniformly below unity, indicating that firms substitute toward cheaper inputs but not by enough to raise their cost shares. In our ICT price simulations (Figure 5), the rental shares of ICT inputs in total

costs fall—and the overall labor share rises. These results, while based on pre-AI capital data, are consistent with the theoretical prior that the aggregate factor-share consequences of AI cost declines will be shaped more by cross-task complementarities than by within-task substitution.

Two extensions of the framework are left for future work. First, the treatment of upstream TFP A_S as exogenous is a maintained simplification. If A_S is endogenous—so that past AI deployment feeds back into current research productivity—the self-amplifying mechanism is strengthened and the path of P_S becomes a function of the entire history of upstream AI investment, with implications for both the speed of capital cost decline and the long-run distribution of factor income. Second, the measurement program sketched in Section 5 for AI-stack software must ultimately be extended to agentic AI, where task-completion benchmarks will be needed to anchor capability adjustment.

Third, the paper has flagged but not developed the financial implications of the framework. Announced commitments to data center construction and foundational model R&D—running to hundreds of billions of dollars in the United States alone—are leading indicators of the upstream investment flows that will eventually appear in national accounts and drive the rental rate dynamics the framework analyses. Connecting these announcements to the two diagnostics requires resolving two prior questions: how much of announced spending will be capitalized rather than expensed under current and prospective accounting standards, and how quickly announced capacity translates into productive capital services. Asset valuations provide a complementary signal, encoding market beliefs about both the pace of capital cost declines and the degree to which AI—combined with complementary organizational investment—will dissolve the bottlenecks that currently moderate its output effects. Developing a coherent framework that integrates announced investment, capitalization rules, and forward-looking valuations into the factor-share diagnostics is a natural and urgent extension of this work.

Addressing these measurement problems is not merely a technical exercise but a prerequisite for the quantitative policy analysis that AI's arrival makes urgent.

Appendix A: The TFP bias—two channels

The downward bias in measured TFP for AI-using industries (sector C) indicated in table 1 of the main text operates through two distinct channels that reinforce each other. Both apply to *value-added* TFP—the residual from a value-added production function in which intermediate inputs have already been netted out of output, and the measure that enters standard growth-accounting decompositions of labor productivity.

The first is the *output channel*. In the national accounts, value added equals gross output minus intermediate consumption. Misclassifying M_C^{AI}—the rental payment for shadow AI capital services—as intermediate consumption inflates measured intermediate inputs, suppresses measured value added in sector C, and reduces the growth rate of value-added TFP by a share-weighted amount of the misclassified expenditure.[9]

The second is the *input channel*. Because the shadow capital stock K_{shadow}^{AI} is unrecorded, the capital input term in the growth-accounting decomposition is understated. Omitting a capital input that is growing rapidly causes its contribution to output growth to be attributed to TFP instead. This effect runs in the *opposite* direction to the output channel: in isolation, the input omission would *inflate* measured TFP. The net bias therefore depends on the relative magnitudes of the two channels. During periods of rapid cloud AI adoption, the output channel is likely to dominate: the misclassified intermediate share M_C^{AI}/Y_C grows faster than the imputed capital contribution, and the net effect is a downward bias in measured TFP.[10]

[9] In fact, the share is greater than one since value added is a fraction of gross output.

[10] The net value-added TFP bias is $-\Delta(M_C^{AI}/Y_C) \cdot (GO_C/Y_C) + s_K^{AI}\hat{K}_{shadow}^{AI}$. The first term is negative (value added suppressed, scaled by the gross-output-to-value-added ratio); the second is positive (omitted capital input inflates the residual). The output channel dominates when the misclassified intermediate share is growing rapidly relative to the shadow capital share — precisely the condition that holds during periods of accelerating cloud AI adoption.

Appendix B: WLS Regression Data

Table A1: Observation-Level Detail: Model Releases and Index Values

Date	Model	Type	w	Tok	Cap	Tok fit	Cap fit	Cap res
2023-03-14	GPT-4 (8K)	TREND	1.00	100.0	100.0	46.8	46.6	53.4
2023-07-01	Claude 2.0	TREND	1.00	35.6	34.3	43.1	40.3	-6.0
2023-11-06	GPT-4 Turbo	TREND	1.00	44.4	39.8	39.0	33.9	5.9
2024-03-04	Claude 3 Opus	PREMIUM	0.25	100.0	85.0	35.6	29.0	56.0
2024-04-09	GPT-4o (launch)	TREND	1.00	22.2	18.0	34.6	27.6	-9.6
2024-06-20	Claude 3.5 Sonnet v1	TREND	1.00	20.0	14.8	32.8	25.1	-10.3
2024-09-12	OpenAI o1	PREMIUM	0.25	83.3	58.1	30.7	22.4	35.7
2024-10-22	Claude 3.5 Sonnet v2	TREND	1.00	20.0	13.3	29.8	21.3	-8.0
2024-12-20	o1 (full release)	PREMIUM	0.25	83.3	54.5	28.5	19.7	34.8
2025-02-27	Claude 3.7 Sonnet	TREND	1.00	20.0	12.9	27.0	17.9	-5.0
2025-04-16	GPT-4.1	TREND	1.00	11.1	7.0	26.0	16.8	-9.8
2025-06-01	Claude 4 Opus	PREMIUM	0.25	100.0	61.6	25.1	15.8	45.8
2025-07-01	Gemini 2.5 Pro	TREND	1.00	12.5	7.8	24.6	15.2	-7.4
2025-08-15	Claude 4.1 Opus	PREMIUM	0.25	100.0	60.7	23.7	14.3	46.4
2025-10-01	o3	PREMIUM	0.25	55.6	33.0	22.9	13.5	19.5
2025-12-01	Claude 4.5	TREND	1.00	20.0	11.7	21.8	12.4	-0.7
2026-01-15	GPT-5	TREND	1.00	13.3	7.6	21.1	11.7	-4.1
2026-02-10	Claude Opus 4.6	TREND	1.00	33.3	18.6	20.7	11.3	7.3
2026-03-09	GPT-5.4 Pro	TREND	1.00	46.7	25.1	20.2	10.9	14.2
Excluded from regression:								
2025-01-31	DeepSeek R1	COMP_SHOCK	—	3.0	2.0	—	—	—
2026-03-01	GPT-5.2 Pro	OUTLIER_PRO	—	210.0	115.2	—	—	—

Notes: Tok = token price index; Cap = capability-adjusted price index (base = 100 at GPT-4 launch, March 2023); Tok fit / Cap fit = WLS fitted values; Cap res = residual from capability-adjusted regression. Index values and ECI scores from Epoch AI leaderboard (Edelman and Lee, 2025; Lee and Emberson, 2025) and provider pricing pages.

Appendix C: Tasks and the Production Function

This Appendix provides a formal analysis of the relation between the "tasks" in a firm and the resulting production function that links inputs and outputs. We use this analysis to clarify the statements about the task approach in the main text, and to help inform the analysis of elasticities of factor substitution that we undertake.

1 Models

We wish to distinguish between X types of technical change. First, factor-augmenting technical progress, when inputs get better at the tasks they currently do (in the older literature, intensive technical change). Second, "automation", which is when capital gets better at tasks done by labor (extensive technical change). Third, new tasks.

1.1 Within tasks

Within task z, output is a CES aggregate of effective capital and effective labor:

$$y(z) = \left[\omega(z)^{1/\sigma_F}\, q_K(z)^{\rho_F} + \big(1-\omega(z)\big)^{1/\sigma_F}\, q_L(z)^{\rho_F}\right]^{1/\rho_F}, \tag{24}$$

where $\rho_F \equiv (\sigma_F - 1)/\sigma_F$ and $\sigma_F > 1$. Effective inputs are defined as:

$$q_K(z) = A^K \gamma^K(z)\, k(z), \qquad q_L(z) = A^L \gamma^L(z)\, l(z). \tag{25}$$

The parameters are:

- σ_F: elasticity of substitution *within a task* between capital and labor;
- $\omega(z) \in (0,1)$: task-specific distribution parameter;
- $\gamma^K(z)$, $\gamma^L(z)$: task-specific factor productivities;
- A^K, A^L: factor-augmenting productivities that boost efficiency uniformly across all tasks K and L perform.

An improvement in all tasks done by K and L is a change in A. An improvement in the relative ability of K to perform tasks formerly done by L is a rise in γ^K/γ^L.

1.1.1 Within-task unit cost

Let R denote the rental rate of capital and W the wage. Effective factor prices — the cost per unit of effective input — are:

$$p_K(z) \equiv \frac{R}{A^K \gamma^K(z)}, \qquad p_L(z) \equiv \frac{W}{A^L \gamma^L(z)}. \tag{26}$$

The dual unit cost of producing one unit of task output is:

$$c(z) = \Big[\omega(z)\, p_K(z)^{1-\sigma_F} + \big(1-\omega(z)\big)\, p_L(z)^{1-\sigma_F}\Big]^{1/(1-\sigma_F)}. \tag{27}$$

The relevant elasticity here is the within-task elasticity σ_F.

1.2 Aggregate cost function

We turn now to the aggregate terms. Let the final good be a CES aggregate of task outputs:

$$Y = \left(\int_{N-1}^{N} y(z)^{\rho_T}\, dz\right)^{1/\rho_T}, \tag{28}$$

where $\rho_T \equiv (\sigma_T - 1)/\sigma_T$ and $\sigma_T > 1$ is the elasticity of substitution *across tasks*. Profit maximisation by firms yields the task-demand system:

$$y(z) = Y\left(\frac{c(z)}{P}\right)^{-\sigma_T}. \tag{29}$$

The dual cost index for final demand is therefore:

$$P(R,W) = \left(\int_{N-1}^{N} c(z)^{1-\sigma_T}\, dz\right)^{1/(1-\sigma_T)}. \tag{30}$$

Substituting the unit task cost function (27) into (30) yields the explicit aggregate unit cost function:

$$P(R,W) = \left\{\int_{N-1}^{N}\left[\omega(z)\left(\frac{R}{A^K\gamma^K(z)}\right)^{1-\sigma_F} + \big(1-\omega(z)\big)\left(\frac{W}{A^L\gamma^L(z)}\right)^{1-\sigma_F}\right]^{\frac{1-\sigma_T}{1-\sigma_F}} dz\right\}^{1/(1-\sigma_T)}. \tag{31}$$

This shows that the aggregate cost function is not itself a simple CES in R and W: it nests one CES inside another. Unless the task heterogeneity takes a very special form, there is no closed-form aggregate production function $Y = F(K,L)$ that depends only on total capital and total labor.

1.3 Special cases

We are now in a position to understand better technical change in the various cases that are popular in the literature.

1.3.1 Hicks-neutral task heterogeneity

Suppose task productivity is "Hicks-neutral", meaning $\gamma^K(z) = \gamma^L(z) \equiv a(z)$ for all z. In this case the task production function becomes:

$$y(z) = a(z)\left[\omega^{1/\sigma_F}\left(A^K k(z)\right)^{\rho_F} + (1-\omega)^{1/\sigma_F}\left(A^L l(z)\right)^{\rho_F}\right]^{1/\rho_F}, \tag{32}$$

which says that within tasks the same CES combination holds, and the only productivity difference across tasks is absolute advantage. The within-task unit cost simplifies to:

$$c(z) = \frac{1}{a(z)}\left[\omega\left(\frac{R}{A^K}\right)^{1-\sigma_F} + (1-\omega)\left(\frac{W}{A^L}\right)^{1-\sigma_F}\right]^{1/(1-\sigma_F)}, \tag{33}$$

and the aggregate unit cost index is:

$$P(R,W) = \frac{1}{\Pi_T}\left[\omega\left(\frac{R}{A^K}\right)^{1-\sigma_F} + (1-\omega)\left(\frac{W}{A^L}\right)^{1-\sigma_F}\right]^{1/(1-\sigma_F)}, \tag{34}$$

where the TFP aggregator Π_T is defined as:

$$\Pi_T \equiv \left(\int_{N-1}^{N} a(z)^{\sigma_T - 1}\, dz\right)^{1/(\sigma_T - 1)}. \tag{35}$$

The corresponding primal aggregate production function is:

$$Y = \Pi_T\left[\omega^{1/\sigma_F}\left(A^K K\right)^{\rho_F} + (1-\omega)^{1/\sigma_F}\left(A^L L\right)^{\rho_F}\right]^{1/\rho_F}. \tag{36}$$

Thus Hicks-neutral task heterogeneity is Hicks-neutral at the aggregate level. Changes in $a(z)$ shift aggregate TFP through Π_T, but there is no task-specific comparative advantage of capital versus labor. The reason is that task heterogeneity affects only the overall productivity level of each task, not the relative productivity of capital and labor within a task. This can be seen by noting that the within-task capital–labor ratio is identical across all tasks z:

$$\frac{q_K(z)}{q_L(z)} = \left(\frac{\omega}{1-\omega}\right)^{\sigma_F}\left(\frac{R/A^K}{W/A^L}\right)^{-\sigma_F}. \tag{37}$$

1.3.2 Special case: the complete specialisation "cut off" model

We now impose the polar case in which within-task substitution is infinite: $\sigma_F \to \infty$. In this limit each task is allocated entirely to the cheaper factor. This is the assignment structure used in the Acemoglu and Restrepo (2018b) cutoff model.

1.3.3 Dual derivation of the cutoff rule

From the within-task cost function (27), the limit $\sigma_F \to \infty$ gives:

$$c(z) = \min\left\{ \frac{R}{A^K \gamma^K(z)},\ \frac{W}{A^L \gamma^L(z)} \right\}. \tag{38}$$

Suppose relative comparative advantage is strictly increasing in z, so there exists a cutoff I such that:

$$\frac{R}{A^K \gamma^K(z)} \leq \frac{W}{A^L \gamma^L(z)} \qquad \text{for } z \in [N-1,\ I], \tag{39}$$

$$\frac{R}{A^K \gamma^K(z)} > \frac{W}{A^L \gamma^L(z)} \qquad \text{for } z \in (I,\ N]. \tag{40}$$

Tasks at or below the cutoff are performed entirely by capital; tasks above the cutoff entirely by labor.

1.3.4 Aggregate unit cost function in the cutoff case

In the cutoff case the aggregate cost index separates into capital tasks and labor tasks:

$$P(R,W) = \left[\int_{N-1}^{I} \left(\frac{R}{A^K \gamma^K(z)} \right)^{1-\sigma_T} dz + \int_{I}^{N} \left(\frac{W}{A^L \gamma^L(z)} \right)^{1-\sigma_T} dz \right]^{1/(1-\sigma_T)}. \tag{41}$$

Defining the task-productivity aggregates for capital and labor:

$$B_K(I,N) \equiv \int_{N-1}^{I} \left[\gamma^K(z)\right]^{\sigma_T - 1} dz, \tag{42}$$

$$B_L(I,N) \equiv \int_{I}^{N} \left[\gamma^L(z)\right]^{\sigma_T - 1} dz, \tag{43}$$

the aggregate cost index takes the simple form:

$$P(R,W) = \left[B_K(I,N) \left(\frac{R}{A^K} \right)^{1-\sigma_T} + B_L(I,N) \left(\frac{W}{A^L} \right)^{1-\sigma_T} \right]^{1/(1-\sigma_T)}. \tag{44}$$

The cutoff model is therefore the special case in which the aggregate dual collapses to an ordinary CES cost function in R/A^K and W/A^L. Only σ_T appears in this expression since $\sigma_F = \infty$.

Now consider the case where a share β_K of tasks are performed by capital and $1-\beta_K$ by labor, with γ_K and γ_L constant within their respective task sets. Then $B_K = \beta_K \gamma_K^{\sigma_T - 1}$ and

$B_L = (1-\beta_K)\gamma_L^{\sigma_T - 1}$, and the cost index reduces to the weighted sum of costs per task:

$$P(R,W) = \left[\beta_K \left(\frac{R}{\gamma^K A^K}\right)^{1-\sigma_T} + (1-\beta_K)\left(\frac{W}{\gamma^L A^L}\right)^{1-\sigma_T}\right]^{1/(1-\sigma_T)}. \tag{45}$$

1.4 Recovering the primal reduced form

The input demand equations are:

$$R = Y^{1/\sigma_T} K^{-1/\sigma_T} \left(\gamma^K A^K\right)^{(\sigma_T-1)/\sigma_T} \beta_K^{1/\sigma_T}, \tag{46}$$

$$W = Y^{1/\sigma_T} L^{-1/\sigma_T} \left(\gamma^L A^L\right)^{(\sigma_T-1)/\sigma_T} (1-\beta_K)^{1/\sigma_T}. \tag{47}$$

Normalising aggregate prices to unity and substituting factor prices into the aggregate unit cost index yields the primal production function in factor quantities. The first form is:

$$Y = \left[\beta_K^{1/\sigma_T}\left(\gamma^K A^K K\right)^{(\sigma_T-1)/\sigma_T} + (1-\beta_K)^{1/\sigma_T}\left(\gamma^L A^L L\right)^{(\sigma_T-1)/\sigma_T}\right]^{\sigma_T/(\sigma_T-1)}, \tag{48}$$

or equivalently, rearranging to make the thinning-out effect explicit:

$$Y = \left[\beta_K \left(\frac{\gamma^K A^K K}{\beta_K}\right)^{(\sigma_T-1)/\sigma_T} + (1-\beta_K)\left(\frac{\gamma^L A^L L}{1-\beta_K}\right)^{(\sigma_T-1)/\sigma_T}\right]^{\sigma_T/(\sigma_T-1)}. \tag{49}$$

This follows Jones and Tonetti (2026). The second form gives intuition: a rise in automation β_K means more output for given K, but if K is spread across more tasks — a "thinning out" effect — it becomes less productive per task. Correspondingly, labor becomes more productive when it concentrates on the tasks for which it has a comparative advantage.

References

Acemoglu, Daron and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "The race between man and machine: Implications of technology for growth, factor shares, and employment," *American economic review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Blackorby, Charles and R Robert Russell, "Will the real elasticity of substitution please stand up?(A comparison of the Allen/Uzawa and Morishima elasticities)," *The American economic review*, 1989, *79* (4), 882–888.

Bontadini, Filippo, Carol Corrado, Jonathan Haskel, and Cecilia Jona-Lasinio, "AI as an Innovation in the Method of Innovation: Implications for Productivity Growth," *AEA Papers and Proceedings*, 2026 forthcoming, *116*.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–372.

Byrne, David M. and Carol A. Corrado, "ICT Asset Prices: Marshaling Evidence into New Measures," Finance and Economics Discussion Series 2017-016, Board of Governors of the Federal Reserve System 2017.

_ **and** _ , "The Increasing Deflationary Influence of Consumer Digital Services," *Economics Letters*, 2020, *196*, 109447.

Corrado, Carol, Charles Hulten, and Daniel Sichel, "Measuring Capital and Technology: An Expanded Framework," in Carol Corrado, John Haltiwanger, and Daniel Sichel, eds., *Measuring Capital in the New Economy*, Vol. 65 of *Studies in Income and Wealth*, Chicago: University of Chicago Press, 2005, pp. 11–45.

_ **,** _ **, and** _ , "Intangible Capital and U.S. Economic Growth," *Review of Income and Wealth*, 2009, *55* (3), 661–685.

Edelman, Yafah and Jaeho Lee, "AI Capabilities Progress Has Sped Up," December 2025. Epoch AI Data Insight. Accessed April 2026.

Fleming, Martin, "Enterprise Information and Communications Technology Software Pricing and Developer Productivity Measurement," *Review of Income and Wealth*, 2025, *71*, e12711.

Ho, Anson, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah, "A Rosetta Stone for AI Benchmarks," 2025. arXiv preprint arXiv:2512.00193v1, November 28, 2025.

Jones, Charles I., "A.I. and Our Economic Future," Working Paper 34779, National Bureau of Economic Research January 2026.

_ **and Christopher Tonetti**, "Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion," January 2026. Unpublished manuscript, Stanford GSB.

Jorgenson, Dale W., "Capital Theory and Investment Behavior," *American Economic Review*, 1963, *53* (2), 247–259.

Lee, Jaeho and Luke Emberson, "Epoch's Capabilities Index Stitches Together Benchmarks Across a Wide Range of Difficulties," November 2025. Epoch AI Data Insight. Accessed April 2026.

Milgrom, Paul and John Roberts, "The Economics of Modern Manufacturing: Technology, Strategy, and Organization," *American Economic Review*, 1990, *80* (3), 511–528.

Oulton, Nicholas, "Long Term Implications of the ICT Revolution: Applying the Lessons of Growth Theory and Growth Accounting," *Economic Modelling*, 2012, *29* (5), 1722–1736.

Reinsdorf, Marshall, "Measurement of Cloud Computing in National Accounts," Guidance Note DZ.8, Digitalization Task Team, Advisory Expert Group on National Accounts 2023. Prepared as consultant to the Bureau of Economic Analysis. Version of March 7, 2023.

Sakai, Haruko, Venkateswarlu Josyula, and Andrew Baer, "Compilation Guidance Note on Cloud Computing," Advisory Expert Group on National Accounts, 28th Meeting SNA/M1.25/9, Inter-secretariat Working Group on National Accounts October 2025. Luxembourg, 21–23 October 2025. With contributions from M. Ludwig, R. Leisch, F. Correa, M. Tamburro (Eurostat); J. Bruner (BEA); T. Mori and A. Kobashi (Bank of Japan).