Artificial intelligence and monetary policy:
A framework and perspective on cyclical transmission, structural transition, and financial stability

Simone Lenzu

Federal Reserve Bank of New York

New York University

CEPR

This version: March, 2026

[Link to latest version]

Abstract

I develop a framework analyzing how artificial intelligence (AI) reshapes monetary policy through three interrelated channels: cyclical transmission, structural transition, and financial stability. In the short run, AI can alter inflation dynamics by changing how supply and demand disturbances map into prices—through shifts in production technologies, pricing behavior, cost pass-through, and expectations—even when conventional measures of economic slack are unchanged. Over longer horizons, AI may shift the natural benchmarks around which policy is calibrated, including potential output and the natural rate of interest. For financial stability, AI may improve credit allocation and risk assessment, but can also heighten systemic vulnerabilities through inflated expectation-driven asset valuations and model monocultures. A particular risk arises at the intersection of these channels: if AI initially depresses realized efficiency through adoption frictions while simultaneously fueling elevated asset valuations, the economy may face cost-push inflation and financial fragility at once—an AI-specific stagflation risk that the interest rate instrument alone is ill-suited to address. I argue that AI does not call for a redefinition of central banks' objectives, but it does require a recalibration of existing frameworks: its diffusion blurs the distinction between cyclical fluctuations and structural shifts, raising the value of cost-side diagnostics and robust policy strategies over exclusive reliance on reduced-form inflation-gap relationships.

Keywords: Artificial Intelligence, monetary policy, inflation, financial stability.

JEL codes: O33, E52, E58, E31, E32, E44.

Email: simone.lenzu@ny.frb.org and slenzu@stern.nyu.edu. I am grateful to Nicola Cetorelli, Marco Del Negro, Keshav Dogra, Fulvia Fringuellotti, Luca Gagliardone, Mark Gertler, Danial Lashkari, Paolo Pesenti, Argia Sbordone, Giorgio Topa, and John Williams for useful feedback, suggestions, and discussions. The views expressed in this paper are those of the author and do not necessarily reflect those of the Federal Reserve Bank of New York and the Federal Reserve System, or any other institution with which the author is affiliated. All mistakes are my own.

1 Introduction

Artificial intelligence (AI) is rapidly emerging as a transformative technology with the potential to reshape many, if not most, aspects of economic activity, including how goods are produced, how workers contribute to production, how prices are set, how expectations are formed, and how risks are assessed. Given that central bank mandates center on price stability and financial stability, these developments place AI squarely within the domain of central banking. Indeed, there is a growing consensus that the relevant question is no longer whether the diffusion of AI will matter for monetary policy, but how. A broad-based adoption of AI is likely to alter the behavior of the variables that monetary policy is designed to stabilize—namely inflation dynamics, real economic activity relative to potential, and financial conditions—and, in doing so, may require policymakers to adapt how they interpret and respond to business-cycle fluctuations.

In this paper, I offer a perspective on these issues with a specific focus on the recent wave of advances in artificial intelligence. Although I refer broadly to AI, the analysis is motivated in particular by the rapid development and diffusion of large-scale, general-purpose, and generative AI (GenAI) models. Unlike earlier, more task-specific forms of automation, these technologies can be flexibly applied across a wide range of activities, augment existing AI tools, and be integrated deeply into core production, decision-making, and organizational processes (e.g., Acemoglu and Restrepo 2020; Brynjolfsson et al. 2021; Aghion et al. 2019; Brynjolfsson et al. 2023).[1] It is this breadth and depth that motivates treating the diffusion of AI as a development of direct relevance for central banks—not because of its technological novelty per se, but because of its potential to alter policy conduct and the risks that central banks seek to manage.

Organized around a New Keynesian monetary framework, the paper distinguishes three conceptually distinct but interrelated channels through which AI may affect monetary policy: the short-run transmission of supply and demand shocks into inflation, the long-run transition in economic fundamentals and equilibrium benchmarks, and the implications for financial markets and financial stability.

[1] By lowering the cost of information processing, prediction, and content generation, recent AI innovations increasingly resemble a general-purpose technology with the potential to reshape how goods and services are produced, how tasks are organized within firms, and how firms scale and compete. Emerging evidence suggests that generative AI, in particular, can augment worker productivity across a broad set of tasks rather than simply automating narrow activities, reinforcing its economy-wide scope.

In the short run, taking equilibrium benchmarks as given, AI can affect the **cyclical transmission** of disturbances into inflation. On the supply side, AI reshapes how cyclical fluctuations translate into inflation by altering both the mapping from economic slack to real marginal costs and the pass-through of marginal-cost movements into prices. These effects operate through changes in production technologies, input-market conditions, pricing frictions, and strategic interactions among firms. On the demand side, AI can shift expectations about future productivity, income, and profitability, affecting current expenditure and inflation dynamics even before productivity gains are realized. Together, these channels determine how shocks propagate into inflation over the business cycle and are analyzed in Section 2.

In the long run, AI is likely to drive a **structural transition** affecting both long-run growth trends and the equilibrium benchmarks around which monetary policy is calibrated. Through its impact on technological change, investment opportunities, risk, and market structure, AI influences potential output, the natural rate of interest, and the sensitivity of aggregate demand to interest rates. These effects operate at low frequency and define the evolving benchmarks around which policy must be calibrated, rather than shaping the short-run propagation of shocks. I discuss these long-run effects in Section 3.

Finally, AI bears directly on central banking through its impact on the financial system and **financial stability**. The diffusion of AI is reshaping key segments of financial intermediation—lending, insurance, and asset management—by altering information processing, risk-taking, and intermediary structure. These changes affect monetary transmission, as shifts in intermediation alter how policy rates map into financial conditions faced by households and firms. At the same time, expectations of AI-driven productivity gains may prove overly optimistic or materialize only gradually. In such an environment, elevated asset valuations, leverage, and reliance on nonbank financing can interact with expectations-driven dynamics to generate systemic vulnerabilities, even without conventional macroeconomic imbalances. These concerns are not hypothetical: by late 2025, AI infrastructure investment had evolved from an internally financed technology buildout into a complex, multi-layered financing chain—spanning corporate bond markets, off-balance-sheet project finance, securitizations, and private credit—with the firms at its center transitioning rapidly from self-financing to debt-funded investment. These issues are taken up in Section 4.

Section 5 places the analysis in a policy perspective and concludes. I argue that AI does not call for a redefinition of monetary policy objectives, but it changes the environment in which those objectives are pursued. AI may weaken the informational content of traditional indicators such as the output gap, while simultaneously shifting and making more uncertain the trade-offs between inflation and economic activity. This raises the value of robust policy design and greater reliance on cost-side diagnostics and pricing behavior, rather than exclusive dependence on reduced-form Phillips curve estimates.

A related challenge is that AI-driven structural change can blur the distinction between cyclical fluctuations and shifts in equilibrium benchmarks such as potential output or the natural rate of interest. Transitional frictions and reorganization costs may temporarily depress effective productivity even as the technological frontier expands, complicating the real-time interpretation of inflationary pressures. At the same time, faster information flows and more responsive expectations may make policy both more powerful and more fragile, as errors in assessing costs, benchmarks, or risks propagate more rapidly through prices, expectations, and balance sheets.

2 AI and the transmission of cyclical disturbances

I use a New Keynesian framework to organize how AI affects the short-run mapping from cyclical supply and demand disturbances into inflation. Households choose consumption and saving through an intertemporal Euler equation. Firms set prices under imperfect competition with nominal and real rigidities, so marginal-cost movements translate into inflation only gradually. Factor markets clear, but may be subject to frictions that drive a wedge between factor prices and marginal products.

I use lowercase letters to denote natural logarithms of the corresponding variables; starred variables denote natural (flexible-price) benchmarks that anchor the cyclical model—such as potential output, the natural wage, and the natural rate of interest; and hats denote log deviations from those benchmarks (e.g., $\widehat{y}_t \equiv y_t - y_t^\star$ denotes the output gap). For the purpose of analyzing how AI affects inflation dynamics in response to business-cycle shocks, these equilibrium benchmarks are taken as given. In the long run, however, such natural benchmarks are themselves endogenous outcomes of technological, institutional, and market-structure forces, which may be reshaped by the diffusion of AI

technologies—a point I return to in Section 3.

In this framework, inflation and the output gap are jointly determined by a two-equation forward-looking system consisting of an aggregate supply block—the New Keynesian Phillips curve (NKPC)—and an aggregate demand block—the dynamic investment-saving (IS) relation:

$$\pi_t = \lambda\, \widehat{mc}_t + \beta\, \mathbb{E}_t\{\pi_{t+1}\}, \tag{1}$$

$$\widehat{y}_t = \mathbb{E}_t\{\widehat{y}_{t+1}\} - \frac{1}{\sigma}\big(\widehat{r}_t + \widehat{s}_t\big) + \varepsilon_t^{IS}. \tag{2}$$

The first equation, the NKPC, describes price-setting behavior. In its primitive formulation (Gagliardone et al. 2025b), the NKPC indicates that inflation depends on the real marginal cost gap, $\widehat{mc}_t$, a measure of economic slackness capturing cyclical demand- and supply-driver pressures on firms' production costs relative to their flexible-price benchmark. The parameter λ governs the strength of cost pass-through into inflation, while $\beta \in (0, 1)$ captures the forward-looking nature of price setting: firms set prices today taking into account expected future economic conditions, which gives rise to the expectational term $\mathbb{E}_t\{\pi_{t+1}\}$.

The second equation, the IS, captures intertemporal aggregate demand dynamics. Today's output gap depends on expected future activity and on the real interest rate gap, $\widehat{r}_t := r_t - r_t^\star$, where the realized real rate is given by the Fisher equation, $r_t = i_t - \mathbb{E}_t\{\pi_{t+1}\}$, and $r_t^\star$ denotes the (Wicksellian) natural rate consistent with zero output gap. The parameter σ governs the sensitivity of expenditure to real interest rate movements. Intuitively, when the real rate rises above its natural benchmark, intertemporal substitution leads households and firms to postpone consumption and investment, reducing current aggregate demand. The terms $\widehat{s}_t$ and ε_t^{IS} collect cyclical intermediation wedges and demand disturbances, discussed below.

Taken together, these equations form a tightly linked system. Demand disturbances that move the output gap feed into inflation through their effects on marginal costs in the Phillips curve. Conversely, supply-side forces that alter productivity, input prices, or equilibrium benchmarks affect both the real marginal cost gap and the natural rate of interest, thereby influencing the IS relation. Expectations link the two blocks directly: expected inflation affects the real interest rate, while expected future output affects current demand. Monetary policy operates through this system by influencing the real

rate, thereby affecting aggregate expenditure and, indirectly, inflation dynamics.

In the short run, AI matters insofar as it alters the transmission of shocks through both equations of this system. On the supply side, changes in production technologies, input-market frictions, and pricing complementarities reshape the mapping from activity to marginal cost and from marginal cost to inflation. On the demand side, AI can affect expectations about future productivity, profitability, and income, thereby shifting current expenditure through the IS relation even before realized productivity changes occur. The implications of AI adoption for monetary policy thus arise from its interaction with both pricing and demand dynamics, possibly reshaping the transmission of shocks and the environment in which stabilization policy operates. I now examine how AI affects the supply and demand blocks in turn.

2.1 Cyclical transmission through the supply side

A useful first-order decomposition writes cyclical movements in real marginal cost as the difference between cost pressure and effective productivity. Inflationary pressure loosens when the latter increases faster than the former, even during output expansions:

$$\widehat{mc}_t = \underbrace{\left(\chi \widehat{y}_t + \widehat{w}_t + \widehat{\tau}_t\right)}_{\text{cost pressure}} - \underbrace{\widehat{a}_t}_{\text{effective productivity}}.$$

Here cost pressure can build or relax due to scale effects associated with cyclical expansions in production ($\chi\widehat{y}_t$), the tightness of factor markets that induce changes in real user costs ($\widehat{w}_t$), and cyclical wedges ($\widehat{\tau}_t$). Substituting this expression into the NKPC in (1) yields:[2]

$$\pi_t = \kappa\, \widehat{y}_t + \lambda\left(\widehat{w}_t + \widehat{\tau}_t - \widehat{a}_t\right) + \beta\, \mathbb{E}_t\{\pi_{t+1}\} \qquad \kappa := \lambda\chi \tag{3}$$

This decomposition isolates two margins through which the diffusion of AI can reshape the transmission of shocks into inflation even when the underlying equilibrium benchmarks remain unchanged: (*i*) a cost channel through which demand- and supply-side cyclical disturbances translate into fluctuations in real marginal costs; (*ii*)

[2]Equation (3) nests the textbook NKPC which, under restrictive assumptions (e.g., separable preferences, competitive labor markets, flexible wages), implies $\pi_t = \tilde{\kappa}\, \widehat{y}_t + \beta\, \mathbb{E}_t\{\pi_{t+1}\}$. See Galí (2015) for derivations. Outside these knife-edge cases, input-market frictions and productivity dynamics contribute independently to inflation; output or unemployment gaps are no longer sufficient statistics. See Erceg et al. (2000), Gertler and Trigari (2009), and Gagliardone et al. (2025b).

a pass-through channel, through which cost movements translate into inflation via the slope of the Phillips curve.

2.1.1 Effects of AI on cyclical moments of real marginal cost

Scale effects—When output rises relative to its natural level—i.e., when the output gap $\widehat{y}_t$ is positive—firms must expand production along increasingly costly margins. The elasticity χ governs how strongly marginal cost responds to cyclical activity, reflecting short-run capacity constraints (e.g., returns to scale) and adjustment margins. The diffusion of AI can materially affect the economy's ability to scale production over the business cycle. It can lower χ by altering the feasibility of scaling: better forecasting, scheduling, predictive maintenance can smooth production and relax effective bottlenecks. This would allow firms to accommodate demand expansions through smoother production schedules, reducing the sensitivity of marginal costs to output and lowering χ.

Inventory management and production smoothing constitute a particularly important—yet often overlooked—margin of adjustment. AI-enhanced forecasting and dynamic inventory control can decouple sales fluctuations from contemporaneous production, weakening the link between output expansions and marginal cost pressures in the short run. This mechanism effectively flattens the real-side transmission from activity to inflation by lowering χ. By contrast, if AI adoption reinforces just-in-time production, increases reliance on concentrated upstream inputs, or tightens interdependencies across supply chains, inventory buffers may shrink, and marginal costs may respond more sharply to demand-driven expansions, raising χ.

AI can also improve the utilization of existing productive capacity—both capital and labor—through better routing, maintenance, monitoring, and task allocation. To the extent that output can expand along utilization margins rather than through costly factor accumulation, marginal costs rise more slowly with activity. More generally, the presence of utilization, inventories, and other short-run adjustment margins implies that the relevant object for inflation transmission is an effective elasticity χ, shaped by technology, organization, and production planning.

Equilibrium dynamics in factor markets—Cyclical movements in real input prices

further contribute to marginal cost dynamics. The term $\widehat{w}_t$ denotes the deviation of the real user cost index of variable inputs—a bundle of labor, intermediate inputs, and capital services—from its natural benchmark.[3] Movements in $\widehat{w}_t$ reflect how nominal rigidities, scarcity, adjustment costs, and substitution patterns across inputs translate aggregate fluctuations into real input cost pressures, making variable inputs temporarily more expensive than under flexible-price adjustment.

Two margins jointly govern the behavior of $\widehat{w}_t$ over the business cycle. The first operates through nominal rigidities and adjustment frictions. When wages and intermediate input prices are set in staggered contracts, bargained infrequently, indexed imperfectly, or subject to adjustment costs, real user costs respond sluggishly or asymmetrically to demand changes, generating a positive $\widehat{w}_t$ even in the absence of changes in technology or factor scarcity. The second margin operates through the interaction of demand and supply elasticities in factor markets. When aggregate demand rises, the equilibrium response of real input prices depends on the elasticity of factor supply and substitution possibilities across inputs. If labor supply is inelastic, expanding employment requires large increases in real wages, raising the user cost of labor above its natural benchmark. Conversely, when labor supply is elastic, or when firms can substitute toward other inputs or utilization margins, real input prices respond more weakly to cyclical expansions.

AI has the potential to reshape both margins. On the first, by reducing adjustment frictions through improved matching, recruiting, scheduling, and training, AI can increase the speed with which real wages and input prices respond to demand, dampening cyclical deviations in real input costs. Similar mechanisms apply to intermediate inputs and capital: AI-enhanced supply-chain integration and procurement efficiency can mitigate the temporary cost increases—a positive $\widehat{w}_t$—that arise when input prices adjust sluggishly to demand expansions. On the second, AI alters the equilibrium response of factor prices by changing task composition, input substitutability, and the effective elasticity of factor supply. This is particularly relevant for specialized skills, robotics, and digital inputs (software, data, organizational capital). If AI raises returns to scale through automation

[3] As shown in the appendix, the real user-cost index $w_t := w_t^n - p_t$ is a cost-minimizing bundle of the real prices of multiple production inputs—such as different types of labor, capital services, and intermediate goods. Up to a first-order approximation around the natural benchmark, $\widehat{w}_t$ is given by a cost-share-weighted average $\sum_{j=1}^{J} c_j^\star \widehat{w}_{j,t}$ of real user cost gaps across different inputs j.

or allows firms to expand output using slack intangible capital, marginal costs may rise more slowly during expansions. Conversely, if AI adoption increases reliance on scarce complementary inputs—specialized chips, computing infrastructure, proprietary data, or high-skill labor—marginal costs may rise more steeply as economic activity accelerates.

Crucially, the two margins can move in opposite directions. For example, AI may make wages more flexible—dampening the first margin—while simultaneously making them more procyclical, because the labor supply of the scarce complementary inputs AI requires is inelastic—amplifying the second. The net effect on $\widehat{w}_t$ is therefore theoretically ambiguous.

Factor markets' frictions—Distortions in input markets that drive a wedge between real input costs and marginal products also contribute to cyclical marginal cost dynamics. In the equation above, $\widehat{\tau_t} := \tau_t - \tau_t^\star$ captures the deviation of input-market wedges, $\tau_t := w_t - mp_t$, from their flexible-price benchmark.[4] These wedges may reflect cyclical variation in labor bargaining frictions, financing premia, adjustment costs, or markups embedded in intermediate-input user costs. AI can generate cyclical variation in $\widehat{\tau_t}$ by altering the nature, intensity, and state-dependence of frictions in input markets. These effects need not be monotonic and may operate in opposite directions across markets and phases of the cycle.

In labor markets, AI can shift $\widehat{\tau_t}$ through opposing channels. AI-enabled monitoring, performance evaluation, and algorithmic management may reduce informational frictions and weaken workers' bargaining power, compressing wage premia and lowering cyclical wage wedges. At the same time, by lowering the cost of job search and application—through automated resume submission, matching platforms, and remote work—AI may substantially increase the volume of applicants per vacancy, raising screening and selection costs for firms and generating positive cyclical wedges between wages and marginal products. Increased reliance on project-based or platform work may further fragment employment relationships, raising the effective cost of scaling labor despite flexible posted wages.

[4]Up to a first-order approximation around the natural benchmark, τ_t is a cost-share-weighted average $\sum_{j=1}^{J} c_j^\star \tau_{j,t}$ of wedges across input markets. The $\tau_{j,t}$ capture average wedges for production factor j in the spirit of Chari et al. (2007), and are distinct from misallocation wedges, which arise from cross-sectional heterogeneity in the frictions individual firms face accessing factor markets (Hsieh and Klenow 2009); the latter are absorbed into the efficiency wedge $\widehat{a}_t$ discussed below.

In capital markets, AI-driven improvements in credit scoring, risk assessment, and contract enforcement may reduce external finance premia and dampen the cyclicality of user costs. Against this baseline, however, AI may amplify financial cyclicality through the procyclicality of credit supply: algorithmic lending and model monocultures may lead intermediaries to ease credit standards during expansions and tighten them abruptly in downturns, often in a correlated manner across institutions. To the extent that AI-enabled finance raises leverage, accelerates credit booms, or synchronizes lending decisions, financing constraints can become more severe in busts despite higher underlying productivity—so that $\widehat{\tau}_t$ rises even as measured productivity improves. I return to these mechanisms in Section 4.

Finally, AI can affect wedges in intermediate-input markets by reshaping contracting technologies and supply-chain organization. Improved forecasting and inventory management may reduce delivery delays, renegotiation costs, and quantity-adjustment frictions, lowering cyclical wedges. Conversely, increased reliance on complex digital supply chains, proprietary platforms, or concentrated upstream providers may amplify contractual rigidities and raise effective input costs during high-demand periods.

Efficiency wedges—Cyclical movements in production efficiency can either dampen cost pressures, when they reflect temporary productivity gains, or amplify them, when they entail transitory efficiency losses. In Equation (3), the term $\widehat{a}_t \equiv a_t - a_t^\star$ captures this *cyclical efficiency wedge*: the deviation of effective production efficiency from its flexible-price benchmark.[5] This wedge reflects forces that cause realized productivity to differ from its flexible-price counterpart even when the technological frontier is unchanged. A key mechanism is resource misallocation: nominal rigidities, sectoral bottlenecks, and adjustment frictions prevent labor, capital, and intermediate inputs from flowing to their most productive uses, depressing effective productivity relative to the flexible-price allocation. Price and wage dispersion, congestion in factor markets, and imperfect reallocation across tasks or firms can all generate such efficiency losses.

From the perspective of inflation dynamics, a negative efficiency wedge ($\widehat{a}_t < 0$)

[5]Effective productivity is defined as $a_t := (1+\chi)\,\mathrm{tfp}_t + \tilde{a}_t$, where tfp_t denotes Hicks-neutral total factor productivity and χ captures amplification through returns to scale and utilization margins. The benchmark $a_t^\star$ holds fixed the underlying technological frontier and the real frictions that characterize the natural allocation. Derivations are reported in the Appendix.

raises real marginal costs for a given level of activity, amplifying inflationary pressure even without strong demand or input-price growth. Conversely, improvements in allocative efficiency can offset cost pressures by allowing output to expand with smaller increases in marginal cost. Because $\widehat{a}_t$ reflects endogenous, state-dependent distortions rather than exogenous technology, its response to shocks need not be monotonic and can vary over the business cycle.

A common narrative holds that AI, by raising productivity, will act as a disinflationary force.[6] While AI is widely regarded as a general-purpose technology that expands the technological frontier—raising $a_t^\star$ over the long run—the implications for short-run inflation dynamics are far less straightforward. AI is disinflationary at cyclical horizons only insofar as it raises effective productivity relative to its flexible-price benchmark by more than it increases real unit input costs. Productivity gains alone are not sufficient to ensure disinflationary outcomes.

Importantly, AI diffusion need not generate a positive $\widehat{a}_t$ in the short run. The impact on realized efficiency depends on adoption frictions, organizational adjustment, and complementarities with existing inputs, all subject to real and nominal rigidities. When prices and wages adjust sluggishly, firms face distorted relative prices that lead to inefficient input allocation across producers and tasks, even as the technological frontier continues to expand, lowering realized efficiency relative to the flexible-price benchmark. AI adoption can therefore initially generate a negative $\widehat{a}_t$ despite ongoing improvements in frontier productivity—the "productivity J-curve" emphasized by Brynjolfsson et al. (2021), whereby major technological innovations depress measured productivity before delivering sustained gains once complementary investments and organizational adjustments are completed.[7] The scale of current resource diversion

[6]This view has been expressed prominently across policy and industry circles. Sam Altman (OpenAI) suggested at a March 2025 Morgan Stanley investor event that AI's disinflationary effects are underappreciated, while Rick Reider (BlackRock) described the technology as a force that "pushes costs down and output up, leading to lower prices" (Link). Kevin Warsh, former Federal Reserve Governor, in a November 2025 *Wall Street Journal* op-ed, argued that AI "will be a significant disinflationary force, increasing productivity and bolstering American competitiveness," adding that a one-percentage-point increase in annual productivity growth "would double standards of living within a single generation" (*WSJ*, November 2025; see also link). Federal Reserve Governor Lisa D. Cook has offered a more measured assessment, acknowledging that AI's disinflationary potential "could, over time, counter any factors putting upward pressure on inflation," while cautioning that "AI could boost prices in the interim, as adoption of the technology might require a surge in aggregate investment" (Cook, 2025).

[7]During early diffusion, firms incur reorganization costs—learning, integration, data cleaning, workflow redesign—that temporarily divert resources from production. Coordination failures arise as tasks are

is substantial: Hyperscaler capital expenditures have begun to exceed operating cash flows, with external financing needs projected at roughly $1.5 trillion over the next three years, directed overwhelmingly toward data center construction, GPU procurement, and supporting infrastructure whose productivity payoffs remain uncertain in timing and magnitude.[8]

A further note of caution applies to the empirical evidence on AI productivity gains. A growing research effort is devoted to measuring productivity improvements among AI-adopting firms. These are welcome contributions; yet data availability constraints mean that most rely on revenue-based productivity measures—output scaled by sales or revenue, deflated by an industry-level price index—which conflate true technical efficiency gains with changes in firms' pricing behavior and markups (**?**; Lenzu et al. 2025). An observed increase in revenue productivity may therefore reflect higher profitability rather than lower production costs—and the two have very different implications for inflation. This distinction is particularly relevant in the context of AI diffusion: emerging evidence suggests that firms deploy AI disproportionately in marketing, customer support, and revenue-facing functions—activities that raise profitability—and only secondarily in R&D or production processes that translate into genuine efficiency gains (Lenzu et al. 2026). Disentangling cost-reducing efficiency from markup-expanding profitability effects remains an open empirical challenge, and overstating the former risks overconfidence in AI as a disinflationary force.

2.1.2 Effects of AI on the passthrough of cyclical real marginal cost movements

A second channel through which AI can affect inflation dynamics operates through the slope of the Phillips curve, λ, which governs how marginal-cost fluctuations translate into inflation. This slope reflects the interaction between nominal and real rigidities that shape firms' price-setting decisions. Following Gagliardone et al. (2025b), I summarize

reallocated and legacy systems coexist with new technologies. See also Bresnahan and Trajtenberg (1995) on diffusion lags and complementarities in general-purpose technologies.

[8]Morgan Stanley Research (2026) and Hamid et al. (2025) project external financing needs for AI infrastructure of $1.5 trillion or more over the next three years. Noffsinger et al. (2025) estimate that approximately one-third of global AI investment is directed toward data centers and two-thirds toward IT equipment and hardware. Most AI data centers require at least one gigawatt of electric power capacity—5 to 10 times that of existing facilities—with per-site construction costs estimated at $1–2 billion before GPU procurement, which is anticipated to account for more than half of total data center cost.

these forces through the decomposition:[9]

$$\lambda = \frac{(1-\theta)}{\theta}\,(1-\beta\theta)\,(1-\Omega)\,\Theta. \tag{4}$$

Equation (4) highlights four distinct margins governing cost pass-through: the frequency of price adjustment (θ), discounting (β), strategic complementarities in price setting (Ω), and macroeconomic complementarities in production (Θ).

Nominal rigidities—Nominal rigidities limit the speed at which firms adjust prices in response to cost shocks, governing how quickly marginal-cost fluctuations translate into inflation. In a standard Calvo (1983) pricing environment, only a fraction $1-\theta$ of firms can reoptimize each period, while the remainder keep prices unchanged. In Equation (4), the term $\frac{1-\theta}{\theta}$ captures this adjustment frequency: as reoptimization becomes more frequent (lower θ), a larger share of firms can respond to cost shocks, increasing pass-through into inflation.[10] Because prices are expected to remain fixed for several periods, firms that do adjust set prices forward-looking: they balance current marginal-cost conditions against expected future costs over the duration of stickiness. The term $(1-\beta\theta)$ captures the role of discounting: when firms weight future profits heavily (higher β) and expect prices to remain fixed longer (higher θ), pricing becomes more sensitive to expected future marginal costs and less responsive to current conditions, dampening the contemporaneous pass-through of cost shocks into inflation.

The diffusion of AI can affect both margins. By reducing the managerial, informational, and computational costs of repricing, AI-enabled automation, real-time demand forecasting, and algorithmic pricing tools can increase the frequency and state-contingency of price adjustment—a decline in effective θ that steepens the Phillips curve. At the same time, improvements in data processing, forecasting accuracy, and real-time monitoring can make pricing decisions more sensitive to expected future conditions, raising the operational relevance of β and making inflation dynamics more expectation-driven and less tightly linked to contemporaneous slack.

[9] See Gagliardone et al. (2025b) for a derivation of the NKPC slope in an environment with oligopolistically competitive firms.

[10] Gagliardone et al. (2025a) develop a state-dependent pricing framework in which the frequency of price adjustment is endogenous and increasing in the magnitude of cyclical shocks (see also Nakamura and Steinsson 2010; Alvarez et al. 2022). The authors show that, in the absence of large aggregate disturbances, the Calvo assumption provides a close approximation to optimal pricing behavior (Gertler and Leahy 2008; Auclert et al. 2022).

The opposing force is uncertainty. AI adoption may increase medium-run uncertainty by accelerating structural change in market structure, technology, and competitive conditions. If firms become less confident about the persistence of future costs or demand, they discount future conditions more heavily, reducing the influence of expected marginal costs on current pricing and making inflation more responsive to current shocks. Thus, even holding deep preferences fixed, AI can reshape inflation dynamics by altering how firms perceive, forecast, and discount future economic conditions.

Strategic price complementarities—Strategic complementarities arise when firms' desired prices depend on competitors' prices, dampening individual firms' pricing responses to changes in their own production costs. The parameter $\Omega \in (0, 1)$ captures the strength of this channel: higher values imply that optimal reset prices weight competitors' prices more heavily and own marginal costs less, so that cost fluctuations translate less into relative price adjustments and, in aggregate, into inflation.[11]

There is growing concern that AI may reshape competitive interactions in ways that strengthen strategic complementarities in pricing. First, AI adoption is skewed toward larger firms, reflecting the importance of complementary intangible assets—data, organizational capital, and specialized skills—that are easier to finance and scale at scale (Calvino and Fontanelli 2023; OECD/BCG/INSEAD 2025; Lenzu et al. 2026).[12] By reinforcing scale advantages, AI may increase market concentration and tilt pricing power toward incumbents controlling compute capacity, data infrastructure, or cloud services (Mihet et al. 2025b). Second, AI improves information acquisition, potentially increasing the extent to which firms condition pricing on competitors' prices.[13] A growing literature on algorithmic pricing shows that learning algorithms can sustain supracompetitive

[11] In models with variable markups, Ω is increasing in the elasticity of desired markups with respect to relative prices. When demand elasticity is constant—as in the benchmark monopolistic competition model with Dixit–Stiglitz demand—markups are fixed, $\Omega = 0$, and strategic complementarities are absent: firms that reoptimize condition only on their own discounted marginal costs, and the Phillips curve is steeper. When markups vary endogenously, pricing decisions become strategic complements and cost pass-through is attenuated. See Amiti et al. (2019) and Gagliardone et al. (2025b).

[12] See also Mihet et al. (2025a), who document that firms integrating data-security expertise into broader R&D teams achieve stronger innovation and growth, underscoring the importance of organizational data complementarities in realizing AI's productive potential.

[13] See also Ramadorai et al. (2025), who document that technically sophisticated firms use AI to extract more consumer data through increasingly opaque privacy agreements, deepening information asymmetries between firms and consumers.

outcomes even without explicit communication (Calvano et al. 2020), a mechanism that maps naturally into stronger strategic complementarities—and that may be amplified when firms or traders share common foundation models that generate homogenized learning biases (Dou et al. 2025).

These forces may be counterbalanced by opposing competitive dynamics. Widespread AI adoption can increase price transparency, reduce search and switching costs, and lower barriers to entry—particularly in platform and digital markets—compressing markups and weakening incumbents' pricing power. In such environments, strategic complementarities may attenuate, increasing cost pass-through into inflation. Whether AI ultimately flattens or steepens the Phillips curve through this channel depends on which force dominates—an outcome likely to vary across sectors, market structures, and stages of AI adoption.

Macroeconomic complementarities—Even when individual firms face identical pricing frictions, aggregate pass-through can be dampened by general-equilibrium feedbacks. These forces are captured by $\Theta \leq 1$, which summarizes the role of macroeconomic complementarities in the transmission of cost fluctuations into prices. They arise from features of the production environment—decreasing returns, shared factor markets, and input-output linkages—that cause firms' costs to co-move in response to aggregate disturbances. When aggregate output expands, higher factor prices, tighter input markets, and rising intermediate costs affect all firms simultaneously. Because marginal costs rise broadly rather than idiosyncratically, relative cost differences across firms are compressed, and the gap between a firm's optimal price and the aggregate price level increases by less, leading to smaller desired relative price adjustments.[14] The result is a parameter $\Theta < 1$ that dampens cost pass-through and flattens the Phillips curve.

AI can reshape these general-equilibrium feedbacks by altering the structure of production. To the extent that AI increases effective returns to scale—enabling

[14]The key point is that firms adjust prices to restore *relative* markups, not to offset changes in the level of marginal costs per se. When cost increases are aggregate and shared, relative marginal costs and markups move little, reducing the marginal benefit of aggressive immediate price adjustment. Macroeconomic complementarities therefore dampen inflation not because marginal costs rise less, but because their commonality weakens incentives to adjust prices. In the appendix, I derive Θ as a function of the elasticity of substitution across producers, the sensitivity of marginal costs to output, and the strength of strategic complementarities (see also Gagliardone et al. 2025b). This mechanism is distinct from χ, which governs how marginal cost responds to the level of economic activity.

firms to expand output using shared platforms, software, or data—macroeconomic complementarities weaken and Θ rises toward one, increasing the sensitivity of inflation to real activity. Conversely, if AI deepens input-output linkages or increases reliance on common upstream inputs and infrastructure, aggregate cost movements become more synchronized, strengthening macroeconomic complementarities and dampening pass-through.

Taken together, AI reshapes—rather than eliminates—the inflation–activity trade-off through multiple margins that need not move in the same direction. Real-side effects on marginal costs, pricing frictions governing pass-through, and expectation-formation channels may offset or reinforce one another, rendering the net impact of AI on the Phillips curve slope ambiguous *ex ante*.

This ambiguity is amplified by heterogeneity. The analysis above treats the aggregate Phillips curve as a representative relationship, but in practice inflation dynamics reflect the aggregation of heterogeneous industry-level curves. AI can shift aggregate inflation dynamics through two channels: within-sector parameter shifts (pricing frictions and pass-through) and across-sector reallocation toward industries with different cost cyclicality and intrinsic pass-through. Because the aggregate relationship depends on sectoral composition, the aggregate cost-price link may shift substantially even if within-industry parameters are stable.[15] Distinguishing within-sector shifts from across-sector reweighting is therefore essential to identifying whether—and why—the Phillips curve is flattening or steepening. Micro price data and sectoral input-cost series are essential to separate the two.

2.2 Cyclical transmission through the demand side

The dynamic IS equation (2) summarizes the demand side of the economy by linking current expenditure to expected future activity and to the real interest rate gap. Mirroring the approach in the previous subsection, and for the purpose of analyzing the short-run

[15] With cross-industry heterogeneity in nominal rigidities and strategic complementarities, the cost-based NKPC can be written as $\pi_t = \lambda\, \widehat{mc}_t + \mathrm{Cov}\big(\lambda_i, \widehat{mc}_{i,t}\big) + \beta\, \mathbb{E}_t\{\pi_{t+1}\}$, where $\lambda := \int \lambda_i\, di$ and $\lambda_i := \frac{(1-\theta_i)(1-\beta\theta_i)}{\theta_i}(1-\Omega_i)\Theta_i$. Inflation depends on average pass-through and on the cross-sectional covariance between marginal-cost fluctuations and industry pricing frictions; reallocation toward sectors with more flexible pricing, stronger complementarities, or tighter cost pressures can shift aggregate dynamics even absent within-industry change. Input-output linkages can further amplify these aggregation effects.

inflationary effects of AI, I abstract from demand-side forces that shift the natural rate $r_t^\star$ or alter preference parameters such as σ, which are equilibrium objects discussed in Section 3. Instead, I focus on two distinct and empirically relevant cyclical channels through which AI can affect aggregate demand: (i) its influence on expectations about future income, productivity, and profitability; and (ii) its effect on financial conditions, which drives a wedge between policy rates and effective borrowing costs.

2.2.1 The expectations channel

The diffusion of AI can alter agents' beliefs about future productivity, income, and profitability even before these changes materialize in measured costs or potential output.[16] Anticipated AI-driven gains raise expected future income and investment returns, stimulating current consumption and investment through intertemporal substitution. As a result, aggregate demand may expand and the output gap may open even if contemporaneous productivity and potential output remain unchanged. This expectations-driven expansion increases utilization and tightens input markets, generating upward pressure on inflation through the Phillips curve.

These considerations raise a natural question: should AI-driven demand news shocks be expected to differ systematically from traditional demand shocks? In reduced form, both operate by shifting the output gap for a given real interest rate gap and can therefore appear similar on impact. The distinction becomes economically relevant once one recognizes that AI-related demand pressure reflects revisions in beliefs about future productivity rather than an exogenous contemporaneous spending impulse. This difference affects the timing and comovement of output, costs, and inflation, complicates real-time identification, and increases the sensitivity of inflation dynamics to how expectations are formed and updated.

A first implication concerns timing and comovement. Because news shocks work through expectations about future conditions, they tend to generate more front-loaded movements in current demand than standard "spot" demand shocks. Anticipated AI-driven gains stimulate consumption and investment today through the entire expected path of future income and returns, even if productivity improvements have not yet

[16] Formally, anticipated AI adoption can be represented as a news shock to future demand conditions, captured either by revisions in expectations of future output gaps $\mathbb{E}_t\{y_{t+1}\}$ or by innovations to the expected future path of IS disturbances ε^{IS}_{t+j} for $j \geq 1$, rather than by a contemporaneous demand shock ε^{IS}_t.

materialized. As a result, output and marginal costs may rise before supply-side efficiency gains are reflected in measured productivity. Whether the resulting inflationary pressure proves temporary or persistent depends on how quickly productivity gains are realized relative to this initial demand expansion.

Recent developments illustrate this mechanism. By the third quarter of 2025, year-over-year capital expenditure growth among the major AI Hyperscalers—Alphabet, Amazon, Meta, Microsoft, and Oracle—exceeded 65 percent, dwarfing the 9 percent growth recorded by the rest of the S&P 500, with AI-related spending accounting for at least one percentage point of real GDP growth.[17] This demand impulse is driven overwhelmingly by expectations of future AI capacity and returns rather than by realized productivity gains—a manifestation of the news-shock channel described above.

This logic connects directly to the efficiency-wedge discussion above. Even when AI is expected to lower production costs in the long run, its diffusion can generate short-run inflationary pressure if expectations-driven demand responds more rapidly than effective productivity. Short-run inflation outcomes therefore depend on the relative timing and strength of two opposing forces: front-loaded demand driven by expectations and the gradual realization of cost-reducing productivity gains on the supply side. When the former dominates, inflation can rise temporarily despite anticipated technological progress.[18]

A second, policy-relevant implication concerns "identification". AI-driven demand news is fundamentally about future supply, but it initially manifests as demand pressure. In real time, this makes it difficult to distinguish from conventional demand overheating, particularly when measured productivity responds sluggishly due to adoption frictions, reorganization costs, or misallocation. As discussed in the context of efficiency wedges, realized productivity may temporarily lag behind its flexible-price benchmark even as expectations improve. In such environments, inflationary pressure may reflect

[17] Capex figures are computed from Hyperscaler quarterly earnings reports (SEC filings). Morgan Stanley Research (2026) project nearly $3 trillion of AI-related infrastructure investment through 2028; Hamid et al. (2025) estimate over $5 trillion in global data center and AI infrastructure spending over five years. Noffsinger et al. (2025) estimate that approximately one-third of global AI investment is directed toward data centers and two-thirds toward IT equipment and hardware.

[18] The distinction between anticipated and unanticipated AI adoption is emphasized in a general-equilibrium setting by Aldasoro et al. (2024a), who show that anticipated AI adoption tends to generate a temporary inflationary response through demand, whereas unanticipated productivity shocks are more likely to be disinflationary on impact.

intertemporal substitution ahead of future capacity rather than excess demand relative to long-run productive potential, posing a challenge for policy calibration.

2.2.2 Savings–investment intermediation

A central insight of modern macro-finance is that monetary policy affects aggregate demand not only through the risk-free real interest rate, but through the process by which private savings are transformed into credit and expenditure. In practice, households' saving decisions and firms' investment decisions are linked by financial intermediaries whose balance sheets, funding conditions, and risk-bearing capacity shape the effective borrowing rates faced by the private sector (Bernanke and Gertler, 1989; Bernanke et al., 1999; Gertler and Karadi, 2011).

In the IS equation, this mechanism is represented by the intermediation wedge s_t, which captures the gap between the policy rate and the effective real borrowing rate relevant for private spending decisions. In equilibrium, this wedge may reflect a gap between the return on savings and the cost of external finance, arising from risk premia, balance-sheet constraints, or intermediary frictions. When s_t rises, the effective cost of funds increases—even if the policy rate is unchanged—tightening financial conditions and depressing aggregate demand.[19]

This representation highlights that monetary policy operates through a two-step process: policy influences market rates r_t, and market rates influence expenditure through the intermediation of savings. The strength of transmission, however, depends on the behavior of s_t, which can materially alter the effective stance of policy as perceived by savers and borrowers.

AI can influence the elasticity of aggregate demand to policy by altering the dynamics of s_t, even if the central bank's reaction function remains unchanged. By improving screening, monitoring, and risk assessment, AI may compress spreads and strengthen the pass-through of policy rates to borrowing costs. At the same time, greater reliance on common models, automated decision-making, or similar risk signals may

[19]The wedge s_t captures distortions between the policy rate and the intertemporal price governing private expenditure decisions. This object is conceptually distinct from the input-cost wedge τ_t entering marginal cost discussed in the previous section. While both may originate in financial frictions, s_t operates through aggregate demand by affecting borrowing conditions, whereas τ_t enters inflation dynamics through firms' production costs.

increase the sensitivity of spreads to news, amplifying fluctuations in effective financial conditions.

For financial stability, the key issue is not only whether AI compresses or amplifies s_t on average, but whether it increases commonality in risk measurement, speed of balance-sheet adjustment, and concentration of exposures. Greater commonality can raise the likelihood of synchronized deleveraging and fire-sale dynamics; greater speed can shorten the window for private and public liquidity provision; and concentration—whether in a small set of large intermediaries or a small set of shared AI/cloud vendors—creates operational and financial single points of failure.[20]

3 Effects of AI on long-run structural transition

In the long run, as nominal rigidities dissipate, AI matters primarily through its effects on the benchmarks around which monetary policy is calibrated—potential output, the natural rate of interest, and the sensitivity of aggregate demand to the policy instrument.

3.1 Potential output and the natural rate of interest

The long-run implications of artificial intelligence operate through its effects on the economy's natural benchmarks—the natural (flexible-price) level of output $y_t^\star$ and the natural real interest rate $r_t^\star$. The natural level of output $y_t^\star$ is the flexible-price level of activity implied by fundamentals (productivity, labor supply, technology, market structure). The natural real rate $r_t^\star$ is the real rate consistent with that allocation. Equivalently, it is the real rate that supports a zero output gap and stable inflation in the sticky-price economy. As nominal rigidities dissipate over time, the economy converges toward this flexible-price equilibrium, in which inflation is stable and real activity is determined solely by preferences, technology, and market structure.

The two benchmarks are linked through intertemporal allocation. In the canonical New Keynesian model, the natural rate is pinned down by the household Euler equation

[20] A growing literature documents a secular shift toward market-based finance and private credit, with important implications for the strength and timing of monetary policy transmission (Fleckenstein et al. 2025; Ivashina 2025).

evaluated at the flexible-price allocation:

$$r_t^\star = -\log\beta + \sigma\,\mathbb{E}_t\{\Delta y_{t+1}^\star\}, \tag{5}$$

where $\mathbb{E}_t\{\Delta y_{t+1}^\star\}$ is expected growth in potential output. Higher expected growth in $y^\star$ raises the real return required to induce households to postpone expenditure, increasing $r^\star$. Conversely, forces that depress investment demand or raise precautionary saving push $r^\star$ downward. Movements in potential output therefore map directly into movements in the natural rate through this intertemporal channel.

Monetary policy stabilizes the economy around these evolving benchmarks rather than attempting to influence them directly. Through conventional operations, policy affects inflation and real activity by influencing deviations of the real policy rate from $r_t^\star$, not through the level of $r_t^\star$ itself—a mechanism made explicit by the IS equation in (2).

Policy errors arise when the central bank mismeasures movements in $r_t^\star$, creating persistent gaps between the policy real rate and the neutral rate. These considerations map directly into standard policy rules. For example, under a conventional Taylor-style rule,

$$i_t = r_t^\star + \bar{\pi} + \phi_\pi(\pi_t - \pi^\star) + \phi_x x_t, \tag{6}$$

changes in $r_t^\star$ require corresponding adjustments to the intercept of the policy rule. If the natural rate rises and policy does not follow, policy becomes unintentionally expansionary; if the natural rate falls and policy does not follow, policy becomes unintentionally contractionary. Reacting to inflation and activity alone is therefore insufficient when benchmark estimates are persistently mismeasured. While theory provides clear guidance on how $r_t^\star$ should enter policy decisions (Woodford 2011; Galí 2015; Clarida et al. 1999), both $y_t^\star$ and $r_t^\star$ are unobserved and subject to substantial real-time uncertainty (Laubach and Williams 2003; Holston et al. 2017), which the rapid diffusion of AI is likely to exacerbate.

The most widely discussed channel operates through productivity. Over time, AI can raise $y_t^\star$ by automating routine cognitive tasks and augmenting complex ones—the displacement and reinstatement effects emphasized by Acemoglu and Restrepo (2020). The net effect on potential output depends on the balance: displacement alone reduces labor demand, whereas reinstatement—the creation of new tasks in which labor retains a comparative advantage—expands production and raises $y_t^\star$. Beyond direct

labor productivity gains, AI can raise total factor productivity by improving logistics, supply-chain coordination, and resource allocation. Most consequentially for long-run growth, AI may accelerate the innovation process itself—by automating aspects of research, experimentation, and knowledge synthesis—effectively raising the productivity of the "production function of ideas" (Aghion et al. 2019). This last channel is critical: it determines whether AI generates a one-time level shift in $y_t^\star$ or a sustained increase in its growth rate.

The Euler equation in (5) makes this distinction operationally precise. A one-time level increase in potential output temporarily raises $\mathbb{E}_t\{\Delta y_{t+1}^\star\}$ during the transition before growth reverts to baseline. In this case, the policy challenge is transitory because $r_t^\star$ rises but eventually falls back. If instead AI permanently accelerates technological change—by augmenting the research process or enabling compounding improvements in AI systems—then $\mathbb{E}_t\{\Delta y_{t+1}^\star\}$ rises persistently and the neutral rate shifts to a structurally higher level, requiring a recalibration of the policy framework. To date, there is a large dispersion in estimates of potential output growth, spanning both scenarios. Acemoglu (2024) argues that AI's aggregate effects will be modest—a few percentage points of GDP over a decade—because most tasks are not easily automatable and cost savings attenuate in general equilibrium. Models in which AI augments the idea-production function generate considerably larger effects, potentially sustaining higher growth rates (Aghion et al. 2019).[21] The range is wide enough to encompass very different paths for $r_t^\star$, and the uncertainty itself generates a real-time inference problem.

Against these productivity-driven forces, which put upward pressure on $y_t^\star$ and $r_t^\star$, several countervailing channels work in the opposite direction. AI may reshape market structure, increase concentration, or shift the distribution of rents toward incumbents in ways that dampen aggregate investment. AI-driven labor displacement risk and structural uncertainty may raise precautionary saving, while income gains concentrated among high-saving households would further depress aggregate consumption demand—both exerting downward pressure on $r_t^\star$, in a manner reminiscent of the secular forces highlighted in the recent literature (Laubach and Williams 2003; Holston et al. 2017). Because these forces work in opposite directions, the net effect of AI on $r^\star$ is *ex ante*

[21] See also Trammell and Korinek (2023) for a systematic analysis of growth dynamics under increasingly capable AI systems, and Nordhaus (2021) for an assessment of whether AI could generate historically unprecedented growth accelerations.

ambiguous. What is unambiguous is the policy implication: the direction and magnitude of any shift in $r^\star$ must be tracked in real time, and policy rules recalibrated accordingly.

The information technology revolution of the late 1990s offers a partial but instructive precedent. IT-driven productivity gains raised trend TFP growth in the United States by roughly one percentage point for about a decade before fading in the mid-2000s (Fernald 2015). The episode illustrates several features likely to recur. First, the gains proved to be a level shift rather than a permanent growth acceleration—a distinction impossible to establish in real time. Second, the real-time identification problem was severe: the Federal Reserve faced conflicting signals about whether rising output reflected expanding potential or demand overheating. Against the views of many contemporaries, Fed chairman Greenspan argued for the former, resisting calls to tighten aggressively on the grounds that conventional estimates of potential output were lagging reality. We now know that Greenspan got it right: unemployment fell well below prevailing natural rate of unemployment ($u_t^\star$) estimates without sustained inflation.[22] Yet, the dot-com crash of 2000–01 demonstrated that even when the supply-side narrative is broadly correct, the expectations channel can generate asset-price dynamics that create independent financial stability risks. The boom was accompanied by financial excess, as optimism about the "new economy" fueled equity valuations that proved unsustainable. The parallels to the current AI cycle—uncertain productivity effects, difficulty distinguishing supply from demand, and expectations-driven asset dynamics—are difficult to ignore.

3.2 Intertemporal elasticity and aggregate demand sensitivity

A second long-run channel connects AI diffusion to monetary policy through the sensitivity of aggregate demand—consumption and investment—to real interest rates.

This channel is captured by σ, the intertemporal elasticity of substitution, in the IS equation (2). In the context of monetary policy transmission, σ is best understood as a reduced-form equilibrium object summarizing all the frictions that limit intertemporal reallocation—borrowing constraints, income risk, and information frictions—rather than as a deep preference parameter.[23] The policy relevance of σ is distinct from that of $y^\star$ and

[22] For a discussion of the real-time policy debate during the IT productivity boom, see Blinder and Reis (2005). Orphanides (2003) documents how real-time mismeasurement of potential output has historically led to substantial policy errors.

[23] In textbook New Keynesian models, σ is derived from household preferences as the curvature of utility

$r^\star$ discussed above. Changes in $y^\star$ and $r^\star$ shift the destination toward which the economy converges and the benchmark around which policy must be calibrated. Changes in σ, by contrast, alter the responsiveness of the economy to the policy instrument itself. Higher σ means that a given deviation of the real rate from $r_t^\star$ generates a larger expenditure response—monetary policy becomes more powerful.

AI may affect σ through several slow-moving structural channels, detailed further in Section 4. On the side of a higher effective σ: AI-enhanced credit scoring and underwriting can relax borrowing constraints for previously constrained households, raising the fraction of agents who can substitute expenditure intertemporally in response to interest rate changes. AI may also compress income risk through better forecasting and risk management, reducing precautionary motives and freeing up intertemporal substitution. And lower investment-planning costs can make investment demand more elastic with respect to expected returns. On the side of a lower effective σ: AI-driven structural uncertainty about future employment and income prospects may increase precautionary saving across a broad swath of households, particularly those at risk of labor displacement; income gains concentrated at the top of the distribution reduce the share of consumption governed by households with high intertemporal substitution; and tighter financial conditions during AI-driven recessions may further constrain borrowers' ability to smooth consumption. The net effect on σ is therefore ambiguous, and likely to vary across the income distribution and the stage of AI adoption.

4 Implications of AI diffusion for financial stability

Beyond price stability, central banks are responsible for safeguarding financial stability. The diffusion of GenAI bears directly on this objective by reshaping financial intermediation, risk assessment, and the propagation of asset-price shocks. Though distinct from the cyclical and structural channels discussed above, these effects interact with them by altering the financial landscape through which monetary policy reaches the real economy.

over consumption; see Galí (2015). The broader interpretation adopted here is more appropriate for the present purposes, as it captures the role of financial frictions and heterogeneity in shaping the aggregate response of demand to interest rates.

4.1 The topology of the financial intermediation and the transmission of monetary policy

AI is already embedded in core segments of the financial system—lending, insurance, and asset management. In lending, AI-enhanced credit scoring and underwriting can expand credit access and improve risk assessment, particularly for thin-file borrowers (Berg et al. 2020; Fuster et al. 2022). The risks, however, are nontrivial. Widespread adoption may embed bias in credit assessments and amplify credit-cycle procyclicality by making supply more sensitive to real-time signals and market sentiment (Gillis et al. (2024); Blattner and Nelson 2024). Consistent with this concern, evidence from Italian banks shows that AI adoption in credit scoring improves credit supply in normal times but amplifies the countercyclical contraction in lending during crises, even after controlling for the length of the bank–firm relationship (Gambacorta et al. 2025). Opaque models trained on historical data can also perform poorly under structural change and generate correlated risk mispricing across institutions (Aldasoro et al. 2024b; International Monetary Fund 2024).

In insurance, AI can improve classification, pricing, and claims processing and reduce fraud. But finer segmentation may erode risk pooling and increase tail-risk exposure precisely in states that are rare or absent in training data, raising questions about long-run insurability and the resilience of insurers' balance sheets under stress (Balasubramanian et al. 2018; European Insurance and Occupational Pensions Authority 2021; International Monetary Fund 2024).

In asset management, AI is used for portfolio construction, algorithmic trading, and risk management (Gensler 2023).[24] Faster information processing and execution can improve efficiency, but may also tighten the link between signals and trading, increasing synchronization of portfolio adjustments and the risk of abrupt price dynamics (Kirilenko et al. 2017).[25]

AI may reshape the topology of intermediation—who bears risk, how concentrated

[24]Beyond public asset management, machine-learning tools are reshaping early-stage capital allocation: Lyonnet and Stern (2025) apply ML to venture capital investment decisions and find that VCs exhibit systematic selection biases—overfitting on observable characteristics such as gender, education, and geography—that algorithmic approaches can identify and potentially correct.

[25]See also Colliard et al. (2026), who show that algorithmic market makers using reinforcement learning tend to learn less competitive pricing strategies due to limited experimentation and noisy feedback, reducing market liquidity in ways that can amplify abrupt price dynamics.

exposures become, and how quickly balance sheets reallocate—with consequences for both the mapping from policy actions into borrower-facing spreads and the likelihood of correlated stress. A simple way to connect financial conditions back to the canonical aggregate demand block is to interpret the relevant real rate in the IS equation (2) as the *effective real rate* faced by households and firms. Such a rate co-moves with the policy rate but is also driven by a wedge (a credit spread, external finance premium, or liquidity premium) that reflects intermediary balance-sheet conditions, funding constraints, and market liquidity. Different topologies of the financial system imply a different reaction of the intermediation wedge to both conventional and unconventional monetary policy, materially altering the effective stance of policy as perceived by the private sector.[26] To fix ideas, consider a central bank that sets the nominal policy rate in response to inflation and economic slack, conditional on its assessment of $r_t^\star$, according to the Taylor rule in Equation (6). For a given setting of the policy rate i_t, different realizations of the spread s_t can therefore imply very different degrees of effective monetary accommodation or tightening, depending on how financial conditions respond to the policy action.

AI can affect monetary transmission by altering the behavior of s_t itself—making spreads more volatile, more state-dependent, or more correlated across institutions, increasing the likelihood that financial conditions tighten abruptly even when inflation is near target and output gaps are small. These effects operate through two channels. First, AI may shift market structure within intermediary sectors—strengthening economies of scale in data and model deployment, increasing reliance on a small set of technology providers, and accelerating diffusion of similar risk signals—thereby altering market power, risk-taking incentives, and the elasticity of credit supply and spreads with respect to policy rates. Second, adoption may be uneven across intermediaries with different funding structures and regulatory constraints. When credit is intermediated through banks, policy tightening primarily transmits through bank funding costs, regulatory constraints, and balance-sheet capacity. When intermediation migrates to NBFI and private credit, transmission may operate more through risk premia, haircuts, margin requirements, and the state-contingent supply of liquidity in secondary markets.[27] As a

[26] More recently, this perspective has been extended to highlight the growing role of market-based finance—securitization, private credit, and CLOs—in transmitting policy through asset prices, leverage, and risk premia (Adrian and Shin 2010; Di Maggio et al. 2020).

[27] Empirical evidence already points to meaningful heterogeneity in monetary policy pass-through across intermediary types. Erel et al. (2023) show that online banks—technology-enabled but still regulated

result, AI-driven reallocation of intermediation can change both the level and cyclicality of the wedge between policy rates and the effective rates faced by households and firms, even holding aggregate fundamentals fixed.

The interaction between AI and the regulatory perimeter raises a macroprudential concern: activity can migrate precisely toward the nodes where supervisory visibility is weaker, where models and data pipelines are less standardized, and where common exposures may build up through shared service providers. In that case, policy tightening could coincide with abrupt nonbank deleveraging—through margin spirals or withdrawal of market-making capacity—producing nonlinear movements in spreads and credit availability that are disproportionate to the initial policy impulse. It is also important to keep in mind that traditional banks are frequently acting as liquidity or credit backstops for these non-bank financial institutions (Acharya et al. Forthcominga; Acharya et al. Forthcomingb), increasing the potential for financial contagion.[28]

The emerging financing ecosystem for AI infrastructure illustrates these dynamics concretely. By late 2025, AI-related debt issuance already spanned multiple segments of the credit market—over $100 billion in investment-grade corporate bonds by Hyperscalers, tens of billions in off-balance-sheet project-finance loans channeled through special purpose vehicles (SPVs), a rapidly growing volume of data center securitizations, and large hybrid private placements involving joint ventures between technology firms and private credit managers.[29] The complexity and layering of this intermediation chain—spanning banks, private credit funds, insurers, and securitization vehicles—materially alters the mapping from policy rates to the effective financial conditions faced by AI-investing firms and their counterparties.

depository institutions—pass through federal funds rate changes to deposit rates roughly 30 basis points more per 100 basis points than traditional banks, with corresponding reallocation of deposit flows. The differential may be even larger for intermediaries outside the traditional bank supervision perimeter.

[28] A growing literature documents the secular rise of market-based finance and private credit and its implications for monetary transmission (Fleckenstein et al. 2025; Ivashina 2025).

[29] In 2025, four Hyperscalers (Alphabet, Amazon, Meta, and Oracle) issued $101 billion of corporate bonds across U.S. dollar and euro investment-grade markets. Separately, five off-balance-sheet project-finance packages tied to Oracle data centers totaled approximately $80 billion, funded by large groups of domestic and foreign banks. Data from public deal announcements, corporate filings, and media reports indicate that data center securitization volumes reached approximately $25 billion in 2025, more than double 2024 levels. These structures involve an array of participants—bank syndicate desks, alternative asset managers, traditional asset managers, insurers, and various institutional investors—whose overlapping exposures complicate monitoring of leverage and interconnectedness.

4.2 Financial infrastructures and systemic interactions

Widespread AI adoption may promote "model monocultures"—many institutions relying on the same datasets, algorithms, or foundation models—turning institution-level efficiencies into system-wide vulnerabilities when technological penetration is high and service provision is concentrated (Financial Stability Board 2017; Financial Stability Board 2024).

Common models can amplify herding and endogenous correlation. If participants condition on the same signals, prices may drift away from fundamentals and then reprice sharply when beliefs shift (International Monetary Fund 2024). This resembles classic crisis amplification mechanisms in which small shocks are magnified through balance-sheet feedbacks and expectations, as in the financial accelerator (Bernanke et al. 1999) and belief-driven boom–bust models (Benigno and Fornaro 2018).

In payment systems, AI-enabled automation, fraud detection, and real-time monitoring can strengthen operational performance. But greater dependence on complex models and a small set of providers can also raise operational and cyber risk, creating potential single points of failure with systemic consequences in the event of outages, cyber incidents, or model breakdowns (McMahon et al. 2024; Kazinnik and Brynjolfsson 2025).[30]

Finally, opacity complicates both risk management and supervision: errors, biases, or regime instability in widely used models can become common shocks rather than idiosyncratic disturbances. Risks that appear diversified at the institution level may therefore be highly correlated systemwide (International Monetary Fund 2024; Basel Committee on Banking Supervision 2020; Gensler 2023), strengthening the case for treating financial stability as integral to the monetary transmission environment (Financial Stability Board 2017; International Monetary Fund 2024).

4.3 Asset valuations and stock-market crash risk

AI diffusion may also generate financial stability risks through its effects on asset valuations and expectations, especially in equity markets. Rapid advances in GenAI have fueled optimism about future productivity and profitability, contributing to elevated

[30] See also Rishabh et al. (2025), who show that AI-intensive firms neutralize cyberrisk's innovation-suppressing effects—but only those with internally developed AI, not mere adopters of external tools—suggesting that cyberrisk acts as a selective tax on data-intensive innovation.

valuations, increased concentration of market capitalization in perceived AI-frontier firms, and large investment flows aimed at expanding AI-related capacity. Evidence for the expectations-driven nature of these valuation effects is direct: firms with high workforce exposure to generative AI earned significant abnormal equity returns immediately following the release of ChatGPT, well before any realized productivity gains had materialized (Eisfeldt et al. 2023). By late 2025, the set of firms most directly exposed to the AI infrastructure buildout—Hyperscalers, advanced semiconductor manufacturers, AI-linked industrials, and AI-exposed utilities—accounted for approximately 16 percent of S&P 500 market capitalization, up from 6 percent at the end of 2022, with aggregate forward valuations rising steeply relative to the broader index.Consistent with a growth-options interpretation, firms investing more in AI exhibit higher equity market betas and more procyclical returns even as the volatility of their operating performance declines—a combination that makes their valuations particularly sensitive to downward revisions in long-run productivity expectations (Babina et al. 2025).

The ongoing structural shift in the financing of AI investment amplifies these vulnerabilities. Historically, the Hyperscalers were asset-light, highly profitable firms that funded capital expenditure almost entirely from retained earnings and large cash reserves, resulting in very low or negative net leverage. This self-financing model insulated AI investment from credit-market conditions. That regime changed abruptly in the final months of 2025. Between September and November 2025 alone, Hyperscalers raised over $100 billion of new investment-grade debt, including some of the largest corporate bond issues in years and a $27 billion off-balance-sheet hybrid placement.[31]

The macroeconomic implication is that AI investment—previously exogenous to the financial cycle—is becoming endogenous to financial conditions. A tightening in credit spreads or a deterioration in risk appetite now feeds back into the pace and scale of AI infrastructure buildout through higher borrowing costs, tighter covenant constraints, and reduced investor demand for new issuance. This channel strengthens the interaction between the intermediation wedge $\widehat{s_t}$ in the IS equation and the real investment decisions

[31]The Hyperscalers' plans to continue returning significant capital to shareholders through dividends and share buybacks appear to be a key factor driving the need for external financing, implying that the leverage-up is partly a choice to preserve equity payouts rather than a pure reflection of investment needs. It should be noted, however, that to date most Hyperscalers retain very low gross debt-to-EBITDA ratios, with the notable exception of Oracle, which has approximately $95 billion of investment-grade bonds outstanding and carries a BBB rating with two negative outlooks from major rating agencies. Issuance figures are based on public deal announcements and Bloomberg data.

that drive AI-related demand, making monetary policy transmission both more potent and more uncertain in sectors at the center of the AI capex cycle.

To the extent these valuations reflect rational expectations, they can be interpreted as part of the transition to a higher-productivity long-run equilibrium (Sections 2–3). Even then, the adjustment path can pose challenges for price stability and raise concerns about concentration and rents. More importantly, the transition is uncertain in timing and magnitude: productivity gains may arrive more slowly, differ across sectors, or be offset by increased market power among AI leaders. When valuations are supported by optimistic but fragile beliefs about an uncertain transition, news that revises the expected path of AI profits can produce disproportionate repricing.

Abrupt revisions to AI expectations can therefore trigger sharp asset-price corrections with broad macro-financial consequences—tighter financial conditions, impaired intermediary balance sheets, and disrupted credit provision—even when inflation is near target and output gaps are small (Bank for International Settlements 2025). Downside risks are amplified when optimism is intermediated through leverage and risk-taking, particularly via nonbank channels (International Monetary Fund 2024; Gopinath 2025; The Economist 2025). The fragility of expectations-driven AI valuations was vividly illustrated in February 2026, when the publication of a fictional scenario describing AI-driven mass unemployment and a 38 percent equity correction triggered a sharp market selloff—with the Dow Jones Industrial Average falling over 800 points in a single session—before any underlying economic fundamental had changed.[32]

The expansion of private credit and other NBFI funding—often more opaque and reliant on maturity transformation—can be a key propagation margin, with banks frequently providing liquidity or credit backstops (International Monetary Fund 2024; Acharya et al. Forthcominga; Acharya et al. Forthcomingb). Consistent with the financing patterns of intangible-intensive firms, AI producers and adopters may rely less on traditional bank loans and more on equity and nonbank finance.[33] This makes it more likely that AI-related exposures accumulate in private credit and

[32]The scenario, titled "The 2028 Global Intelligence Crisis," was published by Citrini Research on February 22, 2026 (link). The market reaction was widely covered; see, e.g., *Fortune*, February 28, 2026, and *Business Insider*, February 2026. The episode was compounded by Block CEO Jack Dorsey's concurrent announcement of a 40 percent workforce reduction attributed to AI tools.

[33]See Falato et al. (2022) and Jang et al. (2025) on the limited role of bank debt for intangible-intensive firms.

other NBFI balance sheets—and that a repricing of AI cash flows propagates through funding conditions, margins, and liquidity rather than through bank loan quantities alone. While more evidence is needed, the AI investment wave may already have generated material private-credit exposures—directly or indirectly—to AI-related firms across both banks and NBFIs. Opacity is compounded by the Hyperscalers' growing reliance on off-balance-sheet structures and forward lease commitments—which reached approximately $500 billion by late 2025—that will appear on balance sheets only once data centers become operational in three to five years.[34] These structures—designed in part to insulate credit ratings and balance-sheet metrics—create a pipeline of liabilities that is largely invisible to investors and regulators during the construction period, increasing the difficulty of tracking leverage and interconnectedness across the AI financing chain.

The contractual architecture of AI data center financing introduces an additional source of fragility: a cascade channel running from Hyperscaler demand to the solvency of the developers and lenders that finance construction. In the typical arrangement, a data center developer creates a special purpose vehicle that raises debt—often 80 percent of the capital structure—to fund construction, with repayment predicated on cash flows from a long-term lease signed by a Hyperscaler.[35] The Hyperscaler commits to use the facility but typically provides limited guarantees beyond the lease itself. If AI demand disappoints—because expected returns on AI investment fail to materialize, because technological obsolescence renders current-generation data centers inadequate, or because the Hyperscaler faces its own financial stress—lease non-renewal or termination would leave the SPV with a highly specialized, capital-intensive asset whose alternative-use value is uncertain. The resulting debt distress would propagate to the banks, private credit funds, and institutional investors that hold the project-finance exposure. Given the scale of these commitments and the concentration of counterparty risk—a single Hyperscaler appears as lessee across multiple large project-finance packages—a pullback by one or two major players could trigger correlated losses across

[34]Hyperscalers increasingly lease data centers from developers rather than building to own, to quickly secure available electric power, minimize up-front costs, and manage accounting preferences. The volume of future leases not yet commenced has grown rapidly.

[35]Project finance loans for data center construction are typically structured as interest-only during the construction period, with maturities of four years and extension options. Based on media reports and market outreach, pricing on these facilities is approximately SOFR + 250 basis points, with commitment fees of 75–100 basis points—a spread of 50–100 basis points above the Hyperscaler's own senior unsecured obligations, reflecting SPV-level lending and construction risk.

a broad set of intermediaries, amplifying the kind of nonlinear financial-conditions tightening discussed above.

Similar fragilities may reside in insurers' balance sheets. Over the past decade, insurers have shifted toward privately placed debt and securitized corporate loans (CLOs), which tend to be less liquid and more opaque than public bonds (Fringuellotti and Santos 2025; Fournier et al. 2024). Although systematic evidence on sectoral exposures is limited, it is plausible that a nontrivial share is linked to firms at the frontier of AI production and adoption, whose financing needs match insurers' long-horizon return targets. Importantly, the investor bases for AI-related corporate bond issuance, data center securitizations, and private placements can overlap—insurers, for instance, participate across all three channels—creating wrong-way risk if a broad deterioration in data center valuations or cash flows were to materialize simultaneously across instruments and vehicles.

5 Policy implications and concluding remarks

Artificial intelligence is reshaping the economy that central banks are tasked to stabilize. Its implications for monetary policy are neither speculative nor distant: AI is already altering how costs respond to activity, shifting equilibrium benchmarks, and changing the structure of the financial system through which policy reaches households and firms. For central banks, the relevant question is not whether AI matters, but how. The analysis organizes these implications around three interrelated dimensions: cyclical transmission, structural transition, and financial stability.

The canonical New Keynesian framework remains a useful organizing device for these purposes. Viewed through this lens, the diffusion of AI does not call for a redefinition of monetary policy objectives, nor does it imply that central banks should respond mechanically to technological innovation or asset-price movements—since such movements may partly reflect rational revisions to long-run productivity expectations beyond the central bank's mandate. Instead, by altering the mapping from economic conditions to inflation and financial risks, AI complicates the interpretation of familiar indicators and the application of standard policy rules, requiring greater caution and judgment in real-time assessment.

AI can alter inflation dynamics even when traditional measures of economic slack appear unchanged. By reshaping the elasticity of marginal costs with respect to output, the cyclical behavior of real input prices, and the degree of cost pass-through, AI may weaken—or, in some sectors, strengthen—the link between inflation and standard indicators such as the output gap or labor market tightness. Inflation may thus become a less reliable cyclical signal, and a given movement in activity may generate cost pressure that departs substantially from historical experience.

More fundamentally, AI may alter the inflation–real activity trade-off itself. If AI allows output to expand with smaller increases in real marginal cost—by lowering the elasticity of production costs with respect to activity—the effective slope of the Phillips curve becomes flatter. In such an environment, inflation is a less reliable contemporaneous signal of demand overheating: large output expansions can occur with modest price pressure, while reducing inflation once it has risen requires larger output contractions than historical experience would suggest. This two-sided ambiguity increases the value of robust policy strategies and calls for greater emphasis on cost-side diagnostics—real marginal cost proxies, input-market conditions, and pricing behavior—rather than exclusive reliance on reduced-form Phillips curve relationships.

A related policy challenge concerns real-time inference about equilibrium benchmarks. Rapid AI-driven structural change can induce persistent uncertainty about potential output and the natural rate of interest, increasing the risk of policy miscalibration. Short-run productivity dynamics need not align with long-run technological gains: transitional frictions, reorganization costs, and misallocation can temporarily depress effective productivity even as the frontier expands. This compresses the space for policy maneuver, making it harder to distinguish cyclical inflationary pressures—which call for a stabilization response—from shifts in equilibrium benchmarks, which do not.

AI may also alter the timing structure of monetary policy transmission in ways that echo—and recast—Friedman's dictum that policy operates with "long and variable lags." AI compresses some lags: faster information flows, quicker pricing responses, and more elastic expectations accelerate the transmission of policy to inflation and activity. But AI may lengthen others. Adoption frictions, reorganization costs, and possible productivity J-curves mean that supply-side adjustments can be protracted and non-monotonic. The net effect is not a uniform shortening of transmission lags but a restructuring of their

timing and shape—some channels accelerating, others elongating—making historically calibrated policy prone to systematic rather than random error. In such an environment, policy errors are less likely to be absorbed gradually and more likely to be amplified.

AI also raises financial stability considerations that interact closely with monetary policy. By accelerating information processing, encouraging reliance on common models and technologies, and interacting with leveraged balance sheets—often outside the traditional banking system—AI may increase the likelihood that financial stress originates from expectations-driven valuation dynamics rather than conventional macroeconomic imbalances. Sudden asset-price corrections can therefore tighten financial conditions rapidly, impair intermediary balance sheets, and disrupt credit provision even when inflation is near target. Precisely because price stability and financial stability may interact more tightly in an AI-intensive economy, having well-specified and distinct tools for each objective becomes more—not less—important: conflating them risks achieving neither.

A particularly acute risk arises from the interaction of the supply-side and financial stability channels. If AI adoption initially depresses realized efficiency while simultaneously fueling elevated asset valuations built on expectations of long-run gains, there is a window during which the economy faces cost pressures and financial fragility at once. On the supply side, reorganization costs and misallocation raise real marginal costs, generating inflationary pressure even without strong demand. On the financial side, the asset valuations and leverage that sustain demand are predicated on productivity gains that have not yet materialized. The combination of cost-push inflation and expectations-driven financial fragility represents a form of AI-specific stagflation risk: the central bank faces pressure on both its price-stability and financial stability mandates, with limited room to address both through the interest rate alone. The current AI investment cycle exhibits precisely this configuration: Hyperscaler capital expenditures have begun to exceed operating cash flows, with trillions of dollars in projected investment directed toward infrastructure whose productivity payoffs remain uncertain, while the share of S&P 500 market capitalization accounted for by AI-exposed firms has nearly tripled since 2022—a combination in which supply-side cost pressures and expectations-driven financial fragility can coexist. Managing this risk calls precisely for the combination of price-stability tools and macroprudential instruments whose joint deployment existing frameworks are not always designed to coordinate.

An analogy synthesizes these insights. If the economy is a car, may AI upgrade the engine—raising potential speed through higher productivity, expanded capacity, and improved information—while making the steering more sensitive by reshaping inflation transmission, shifting policy benchmarks, and amplifying financial feedbacks. The task of central banks is not to slow or accelerate the engine, but to adjust the steering: calibrating policy to maintain macroeconomic stability as the structure of the economy evolves. From this perspective, successful monetary policy hinges not on reacting to AI per se, but on maintaining clarity about what policy can and cannot control, improving real-time inference about costs and benchmarks, and designing robust strategies that perform well under heightened structural uncertainty. In this sense, AI reinforces—rather than overturns—a central lesson of modern monetary policy: effective stabilization requires a deep and continuously updated understanding of the economy to which policy is applied.

References

Daron Acemoglu. The simple macroeconomics of AI. Working Paper 32487, National Bureau of Economic Research, May 2024.

Daron Acemoglu and Pascual Restrepo. Artificial intelligence and jobs. *Journal of Economic Perspectives*, 34(3):30–50, 2020.

Viral V Acharya, Nicola Cetorelli, and Bruce Tuckman. Where do banks end and nbfis begin? *Review of Corporate Finance Studies*, Forthcominga.

Viral V Acharya, Nicola Cetorelli, and Bruce Tuckman. Transformed intermediation: Credit risk to nbfis, liquidity risk to banks. *Journal of Finance Insights and Perspectivess*, Forthcomingb.

Tobias Adrian and Hyun Song Shin. Liquidity and leverage. *Journal of Financial Intermediation*, 19(3):418–437, 2010.

Philippe Aghion, Benjamin Jones, and Charles I. Jones. Artificial intelligence and economic growth. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*, pages 237–282. University of Chicago Press, 2019.

Inaki Aldasoro, Sebastian Doerr, Leonardo Gambacorta, and Daniel Rees. The impact of artificial intelligence on output and inflation. BIS Working Paper 1179, Bank for International Settlements, 2024a.

Iñaki Aldasoro, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti, and Merlin Stein. Intelligent financial system: how ai is transforming finance. Technical report, Bank for International Settlements, 2024b.

Fernando Alvarez, Francesco Lippi, and Aleksei Oskolkov. The macroeconomics of sticky prices with generalized hazard functions. *The Quarterly Journal of Economics*, 137(2): 989–1038, 2022.

Mary Amiti, Oleg Itskhoki, and Jozef Konings. International shocks, variable markups, and domestic prices. *The Review of Economic Studies*, 86(6):2356–2402, 2019.

Andrew Atkeson and Ariel Burstein. Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, 98(5):1998–2031, 2008.

Adrien Auclert, Rodolfo D Rigato, Matthew Rognlie, and Ludwig Straub. New pricing models, same old phillips curves? Technical report, National Bureau of Economic Research, 2022.

Tania Babina, Anastassia Fedyk, Alex Xi He, and James Hodson. Artificial intelligence and firms' systematic risk. Working Paper, SSRN 4868770, October 2025.

Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. Insurance 2030—the impact of ai on the future of insurance. *McKinsey & Company*, pages 1–10, 2018.

Bank for International Settlements. Annual economic report 2025: Chapter 3. financial conditions in a changing global financial system. Technical report, Bank for International Settlements, June 2025.

David R Baqaee, Emmanuel Farhi, and Kunal Sangani. The supply-side effects of monetary policy. *Journal of Political Economy*, 132(4):1065–1112, 2024.

Basel Committee on Banking Supervision. Sound practices: Model risk management. Technical report, Bank for International Settlements, 2020.

Gianluca Benigno and Luca Fornaro. Stagnation traps. *Review of Economic Studies*, 85(3): 1425–1470, 2018.

Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7):2845–2897, 2020.

Ben S. Bernanke and Mark Gertler. Agency costs, net worth, and business fluctuations. *American Economic Review*, 79(1):14–31, 1989.

Ben S. Bernanke, Mark Gertler, and Simon Gilchrist. The financial accelerator in a quantitative business cycle framework. *Handbook of Macroeconomics*, 1:1341–1393, 1999.

Laura Blattner and Scott Nelson. How costly is noise? data and disparities in consumer credit. 2024.

Alan S. Blinder and Ricardo Reis. Understanding the Greenspan standard. In *The Greenspan Era: Lessons for the Future*, pages 11–96. Federal Reserve Bank of Kansas City, 2005. Proceedings of the Jackson Hole Economic Policy Symposium.

Timothy F Bresnahan and Manuel Trajtenberg. General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108, 1995.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. NBER Working Paper 31161, National Bureau of Economic Research, 2023.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10): 3267–3297, 2020.

Flavio Calvino and Luca Fontanelli. A portrait of ai adopters across countries. *Documents de travail de l'OCDE sur la science, la technologie et l'industrie*, 2023.

Guillermo A Calvo. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12(3):383–398, 1983.

Varadarajan V Chari, Patrick J Kehoe, and Ellen R McGrattan. Business cycle accounting. *Econometrica*, 75(3):781–836, 2007.

Richard Clarida, Jordi Gali, and Mark Gertler. The science of monetary policy: A new keynesian perspective. *Journal of Economic Literature*, 37(4):1661–1707, 1999.

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo. Algorithmic pricing and liquidity in securities markets. *The Review of Financial Studies*, 2026. Advance article. DOI: 10.1093/rfs/hhag010.

Lisa D. Cook. Opening remarks on productivity dynamics, May 2025. Speech delivered at "Finishing the Job and New Challenges," Hoover Institution, Stanford University.

Marco Di Maggio, Amir Kermani, and Christopher J. Palmer. How quantitative easing works: Evidence on the refinancing channel. *The Review of Economic Studies*, 87(3): 1498–1528, 2020.

Winston Wei Dou, Itay Goldstein, and Yan Ji. AI-powered trading, algorithmic collusion, and price efficiency. Working Paper 34054, National Bureau of Economic Research, 2025.

Andrea L. Eisfeldt, Gregor Schubert, and Miao Ben Zhang. Generative AI and firm values. Working Paper 31222, National Bureau of Economic Research, 2023.

Christopher J. Erceg, Dale W. Henderson, and Andrew T. Levin. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics*, 46(2):281–313, 2000.

Isil Erel, Jack Liebersohn, Constantine Yannelis, and Samuel Earnest. Monetary policy transmission through online banks. Working Paper 31380, National Bureau of Economic Research, June 2023. Revised March 2025.

European Insurance and Occupational Pensions Authority. Artificial intelligence governance principles: Towards ethical and trustworthy artificial intelligence in the european insurance sector. Technical report, EIOPA, 2021.

Antonio Falato, Dalida Kadyrzhanova, Jae Sim, and Roberto Steri. Rising intangible capital, shrinking debt capacity, and the us corporate savings glut. *The Journal of Finance*, 77(5):2799–2852, 2022.

John G. Fernald. Productivity and potential output before, during, and after the Great Recession. *NBER Macroeconomics Annual*, 29(1):1–51, 2015.

Financial Stability Board. Artificial intelligence and machine learning in financial services. Technical report, FSB, 2017.

Financial Stability Board. The financial stability implications of artificial intelligence. Technical report, Financial Stability Board, 2024.

Quirin Fleckenstein, Manasa Gopal, Germán Gutiérrez, and Sebastian Hillenbrand. Nonbank lending and credit cyclicality. *The Review of Financial Studies*, page hhaf024, 2025.

Anne Fournier, Ralf Meisenzahl, and Andy Polacek. Privately placed debt on life insurers' balance sheets: Part 1—a primer. Chicago Fed Letter 493, Federal Reserve Bank of Chicago, May 2024.

Fulvia Fringuellotti and João A. C. Santos. Insurance companies and the growth of corporate loans' securitization. *FRB of New York Staff Report*, 975, 2025.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance*, 77(1):5–47, 2022.

Luca Gagliardone, Mark Gertler, Simone Lenzu, and Joris Tielens. Micro and macro cost-price dynamics in normal times and during inflation surges. Technical report, National Bureau of Economic Research, 2025a.

Luca Gagliardone, Mark Gertler, Simone Lenzu, and Joris Tielens. Anatomy of the phillips curve: micro evidence and macro implications. *American Economic Review*, 115(11): 3941–3974, 2025b.

Jordi Galí. *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, 2015.

Leonardo Gambacorta, Fabiana Sabatini, and Stefano Schiaffi. Artificial intelligence and relationship lending. Working Paper 1244, Bank for International Settlements, 2025.

Gary Gensler. Isaac newton to ai. Technical report, U.S. Securities and Exchange Commission, 2023. Speech at the National Press Club.

Mark Gertler and Peter Karadi. A model of unconventional monetary policy. *Journal of Monetary Economics*, 58(1):17–34, 2011.

Mark Gertler and John Leahy. A phillips curve with an ss foundation. *Journal of Political Economy*, 116(3):533–572, 2008.

Mark Gertler and Antonella Trigari. Unemployment fluctuations with staggered nash wage bargaining. *Journal of Political Economy*, 117(1):38–86, 2009.

Talia Gillis, Scott Nelson, and Jann Spiess. Regulating algorithms: What and when. Technical report, National Bureau of Economic Research, 2024.

Gita Gopinath. Statement by imf first deputy managing director gita gopinath at the conclusion of the third meeting of g20 finance ministers and central bank governors. Technical report, International Monetary Fund, July 2025.

Tarek Hamid et al. AI data center financing: Infrastructure investment and capital market implications. Technical report, J.P. Morgan Chase & Co., November 2025. J.P. Morgan Research report. Summarized in Bloomberg, November 10, 2025.

Kathryn Holston, Thomas Laubach, and John C. Williams. Measuring the natural rate of interest: International trends and determinants. *Journal of International Economics*, 108:S39–S75, 2017.

Chang-Tai Hsieh and Peter J Klenow. Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, 124(4):1403–1448, 2009.

International Monetary Fund. Global financial stability report, october 2024: Chapter 3 – advances in artificial intelligence and financial stability. Technical report, International Monetary Fund, 2024.

Victoria Ivashina. Private credit: What do we know? Technical report, Harvard Business School and NBER, October 2025. Working paper.

Young Soo Jang, Dasol Kim, and Amir Sufi. The lending technology of direct lenders in private credit. Technical report, Working Paper, November 2025.

Sophia Kazinnik and Erik Brynjolfsson. Ai and the fed. Technical report, National Bureau of Economic Research, 2025.

Miles S Kimball. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking*, 27(4):1241–1277, 1995.

Andrei Kirilenko, Albert Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3):967–998, 2017.

Thomas Laubach and John C. Williams. Measuring the natural rate of interest. *Review of Economics and Statistics*, 85(4):1063–1070, 2003.

Simone Lenzu, David A Rivers, Joris Tielens, and Shi Hu. *Financial shocks, productivity, and prices*. National Bank of Belgium, 2025.

Simone Lenzu, Filippo Mezzanotti, and Joris Tielens. A early assessment of the adoption, use, and impact of genai on firms' production, productivity, and profitability. 2026.

Victor Lyonnet and Léa H. Stern. Machine learning about venture capital choices. Working Paper, SSRN 4035930, 2025.

Christopher McMahon, Donald McGillivray, Ajit Desai, Francisco Rivadeneyra, Jean-Paul Lam, Thomas Lo, Danica Marsden, and Vladimir Skavysh. Improving the efficiency of payments systems using quantum computing. *Management Science*, 70(10):7325–7341, 2024.

Roxana Mihet, Anastassia Fedyk, Orlando Gomes, and Kumar Rishabh. Data innovation complementarity and firm growth. Working Paper, SSRN 4559921, 2025a.

Roxana Mihet, Kumar Rishabh, and Orlando Gomes. Is it AI or data that drives firm market power? *Journal of Monetary Economics*, 2025b.

Morgan Stanley Research. AI is now a macro variable. Are you positioned? Technical report, Morgan Stanley, March 2026.

Emi Nakamura and Jón Steinsson. Monetary non-neutrality in a multisector menu cost model. *The Quarterly Journal of Economics*, 125(3):961–1013, 2010.

Jesse Noffsinger, Mark Patel, Pankaj Sachdeva, Arjita Bhan, Haley Chang, and Maria Goodpaster. The cost of compute: A $7 trillion race to scale data centers. Technical report, McKinsey & Company, April 2025.

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

OECD/BCG/INSEAD. The adoption of artificial intelligence in firms: New evidence for policymaking, 2025.

Athanasios Orphanides. The quest for prosperity without inflation. *Journal of Monetary Economics*, 50(3):633–663, 2003.

Tarun Ramadorai, Antoine Uettwiller, and Ansgar Walther. Privacy policies and consumer data extraction: Evidence from U.S. firms. *Review of Finance*, 29(5):1337–1367, 2025.

Kumar Rishabh, Roxana Mihet, and Julian Jang-Jaccard. Cyberrisk and AI firms. Working Paper 25-39, Swiss Finance Institute, April 2025.

The Economist. What if the $3trn ai investment boom goes wrong? *The Economist*, September 2025.

Philip Trammell and Anton Korinek. Economic growth under transformative AI. Working Paper 31815, National Bureau of Economic Research, October 2023. Revised 2024.

Olivier Wang and Iván Werning. Dynamic oligopoly and price stickiness. *American Economic Review*, 112(8):2815–49, 2022.

Michael Woodford. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press, 2011.

A Model appendix

A.1 Setup

The economy is populated by heterogeneous producers (or firms), denoted by f, each operating in an industry $i \in \mathcal{I} = [0, 1]$. We denote by $\mathcal{F}_i$ the set of firms competing in industry i. Each firm is measure zero relative to the economy as a whole but may be large relative to its industry. Hence, it takes the aggregate expenditure as given but internalizes the effect of its pricing decisions on the consumption and price index of its industry. Time is discrete.

Let P_{ft} denote the price charged by each firm for a unit of its output, P_{it} the industry price index, φ_{ft} a firm-specific relative demand shifter, and Y_{it} the real industry output. For any industry i, we consider an arbitrary, invertible demand system that generates a residual demand function of the following form:

$$\mathcal{D}_{ft} := d(P_{ft}, P_{it}, \varphi_{ft}) Y_{it} \qquad \forall f \in \mathcal{F}_i. \tag{A.1}$$

Below we follow closely the steps in Gagliardone et al. (2025b) to characterize the firms' pricing problem and derive the cost-based NKPC. We extend the framework by allowing for a richer characterization of real marginal cost that accounts for cyclical variation in factor market frictions and technology.

A.2 The firm pricing problem

Firms adjust their prices during each period in order to maximize expected profits facing nominal rigidities as in Calvo (1983). Each period they face a probability $(1 - \theta) \in [0, 1]$ of being able to change their price, independent across time and across firms. Thus, the price P_{ft} paid by consumers to buy goods produced by firm f is either the optimal reset price if the firm is able to adjust, denoted by P^o_{ft}, or the price it charged in the previous period, P_{ft-1}.

When choosing P^o_{ft}, firms consider both their own costs, the pricing choices made by competitors, as well as the impact of their own price adjustments on their residual demand and on the industry-wide price index. Let $\Lambda_{t,\tau}$ denote the stochastic discount factor between time t and $t + \tau$, $TC_{ft} := TC(\mathcal{D}_{ft})$ the real total costs, and MC^n_{ft} the nominal marginal cost of firm f (which we characterize below). Then the optimal reset

price P^o_{ft} solves the following profit maximization problem:

$$\max_{P^o_{ft},\{Y_{ft+\tau}\}_{\tau\geq 0}} \mathbb{E}_t \left\{ \sum_{\tau=0}^{\infty} \theta^\tau \left[\Lambda_{t,\tau} \left(\frac{P^o_{ft}}{P_{t+\tau}} \mathcal{D}_{ft+\tau} - TC(\mathcal{D}_{ft+\tau}) \right) \right] \right\},$$

subject to the sequence of expected demand functions $\{\mathcal{D}_{ft+\tau}\}_{\tau\geq 0}$ in Equation (A.1). Nominal rigidities generate forward-looking pricing behavior, as firms take into account that it might not be possible to adjust prices every period. As a result, the optimal reset price is a weighted average of current and expected future nominal marginal costs and markups. Denoting by μ_{ft} the desired log markup, the FOC of the problem is:

$$\mathbb{E}_t \left\{ \sum_{\tau=0}^{\infty} \theta^\tau \Lambda_{t,\tau} \mathcal{D}_{ft+\tau} \left[\frac{P^o_{ft}}{P_{t+\tau}} - (1+\mu_{ft+\tau}) \frac{MC^n_{ft+\tau}}{P_{t+\tau}} \right] \right\} = 0. \tag{A.2}$$

Thus, the optimal reset price depends on the expected path of marginal cost and desired markups over the period the firm expects its price to be fixed, where θ^τ is the probability the firm expects its price to be fixed τ periods from now.

We log-linearize the FOC in Equation (A.2) around the symmetric steady state with zero inflation.[36] Denoting the variables in logs with lower-case letters, we obtain that the reset price satisfies:

$$p^o_{ft} = (1-\beta\theta)\mathbb{E}_t \left\{ \sum_{\tau=0}^{\infty} (\beta\theta)^\tau \left(\mu_{ft+\tau} + mc^n_{ft+\tau} \right) \right\}. \tag{A.3}$$

The log-linearized desired markup (in deviation from steady state markup μ_f) is a function that depends inversely on the log-difference between a firm's own reset price and its competitors' prices (p^{-f}_{it}):

$$\mu_{ft} - \mu_f = -\Gamma \left(p^o_{ft} - p^{-f}_{it} \right) + u^\mu_{ft}, \tag{A.4}$$

where $\Gamma > 0$ denotes the markup elasticity with respect to prices and u^μ_{ft} is a firm-specific demand shock to the desired markup that depends on the demand shifter φ_{ft}. Gagliardone et al. (2025b) show that, under weak assumptions, the expression in Equation (A.4) holds for standard frameworks with imperfectly competitive firms, including monopolistic competition with variable elasticity of demand (Kimball 1995), static oligopoly (Atkeson and Burstein 2008) and dynamic oligopoly (Wang and Werning 2022). These frameworks share the property that, in equilibrium, a firm's elasticity of demand declines as its market

[36] The choice of the zero-inflation steady state permits simpler notation; but is largely immaterial for our purposes.

share increases. Thus, the presence of strategic complementarities in price setting implies that a relative price increase lowers a firm's desired markup, dampening the response of prices to marginal cost.

Substituting the expression for $\mu_{ft+\tau}$ in the log-linearized first-order condition, we obtain the following forward-looking pricing equation:

$$p^o_{ft} = (1-\beta\theta)\Theta\mathbb{E}_t\left\{\sum_{\tau=0}^{\infty}(\beta\theta)^\tau\left((1-\Omega)(mc^n_{ft+\tau}+\mu_f)+\Omega p^{-f}_{it+\tau}\right)\right\}+u_{ft}, \tag{A.5}$$

where u_{ft} captures residual variation in the markup that depends on the aggregation of firms' demand shifters and the changes in the slope of competitors' reaction function. For the purposes of this paper, we ignore this term and set it to zero.

The parameter $\Omega := \frac{\Gamma}{1+\Gamma}$ captures the strength of strategic complementarities and impacts the firm's pricing policy by muting the price response to changes in marginal costs. If the demand elasticity is constant, as in the textbook New Keynesian model with monopolistically competitive firms, the desired markup is a constant. In this case, $\Omega = 0$ and the optimal pricing equation simplifies to the familiar formulation where the reset price exclusively depends on the current and future stream of marginal costs. Competitors' prices are then irrelevant.

The parameter $\Theta \leq 1$ captures macroeconomic complementarities due to aggregate returns to scale in production. For example, under CES demand with elasticity of substitution γ, we have that $\Theta := \frac{1}{1+\gamma(1-\alpha)(1-\Omega)}$. A higher elasticity of substitution increases competitive pressure and magnifies the aggregate response of costs. Similarly, a higher elasticity of cost to output—reflecting stronger decreasing returns or tighter capacity constraints—amplifies aggregate marginal-cost pressures as output expands, strengthening macroeconomic complementarities and lowering Θ. Strategic complementarities interact with these forces in a subtle way. A higher degree of complementarities in price setting weakens macroeconomic complementarities by dampening the amplification of marginal costs through output.

A.3 Aggregation and the cost-based New Keynesian Phillips curve

To obtain closed form expressions, suppose there are $N < \infty$ firms in each industry i competing a la Bertrand, and order firms in each industry from 1 to N.[37] The aggregate price index (in log-linear terms) is:

$$p_t = \int_{i \in I} \left(\frac{1}{N} \sum_{f=1}^{N} p_{fit} \right) di,$$

(In the paper, we dropped the industry subscript for ease of notation.) Denote by $B^{\star}_{ft}$ for $f \in \{1, \ldots, N\}$ the set of industries in which the f-th firm can adjust. The price index can then be rewritten as:

$$p_t = \frac{1}{N} \sum_{f=1}^{N} \left(\int_{i \in I \setminus B^{\star}_{ft}} p_{fit-1} di + \int_{i \in B^{\star}_{ft}} p^{o}_{fit} di \right),$$

where we are using the fact that firms that cannot adjust set their price to their $t-1$ level, whereas firms that can adjust set it to the optimal reset price.

Since $B^{\star}_{ft}$ has measure $1-\theta$, and the identity of firms that adjust is an i.i.d. draw from the total population of firms, using the law of large numbers for each $f = \{1, \ldots, N\}$ across industries we have that:[38]

$$\frac{1}{N} \sum_{f=1}^{N} \int_{i \in I \setminus B^{\star}_{ft}} p_{fit-1} di = \theta \int_{i \in I} \left(\frac{1}{N} \sum_{f=1}^{N} p_{fit-1} \right) di = \theta p_{t-1}$$

and

$$\frac{1}{N} \sum_{f=1}^{N} \int_{i \in B^{\star}_{ft}} p^{o}_{fit} di = (1-\theta) \int_{i \in I} \left(\frac{1}{N} \sum_{f=1}^{N} p^{o}_{fit} \right) di.$$

Defining the average reset price in the economy:

$$p^{o}_{t} := \int_{i \in I} \left(\frac{1}{N} \sum_{f=1}^{N} p^{o}_{fit} \right) di,$$

we obtain an equation characterizing the log-linear aggregate price index:

$$p_t = (1-\theta) p^{o}_{t} + \theta p_{t-1}, \tag{A.6}$$

[37] Letting $N \to \infty$, all results hold under Kimball preferences. Note also that the same argument goes through with minor modifications, but heavier notation, for $N_i \neq N$ for a non-zero measure of industries.

[38] The i.i.d. assumption implies that: $\int_{i \in B \subseteq [0,1]} p_{fit} di = Pr(B) \int_{i \in I} p_{fit} di$. Notice also that $\int_{i \in [0,1]} \left(\frac{1}{N} \sum_{f=1}^{N} p_{it}^{-f} \right) di = \int_{i \in [0,1]} \left(\frac{1}{N} \sum_{f=1}^{N} \left[\frac{N}{N-1} p_{it} - \frac{1}{N-1} p_{fit} \right] \right) di = p_t$.

with p_t and p_t^o denoting the aggregate price indices implied by the demand system. Next, we replace the aggregate reset price, p_t^o, with an expression that depends on aggregate marginal costs and prices.

Let $mc_t = mc_t^n - p_t$ denote aggregate real marginal cost (characterized below) and define aggregate inflation as $\pi_t = p_t - p_{t-1}$. Following the steps in Gagliardone et al. (2025b), we average across firms and industries to obtain an expression for the aggregate reset price:

$$p_t^o = (1 - \beta\theta)\left((1 - \Omega)\Theta\widehat{mc}_t + p_t\right) + \beta\theta\mathbb{E}_t p_{t+1}^o$$

Subtracting p_t from both sides and using the log-linearized price index:

$$p_t^o - p_t = (1 - \beta\theta)(1 - \Omega)\Theta\widehat{mc}_t + \beta\theta(\mathbb{E}_t p_{t+1}^o - p_t)$$

Rearranging and combining the equation for the log-linear aggregate price index in A.6 with the equation above gives the primitive formulation of the NKPC curve:

$$\pi_t = \lambda\, \widehat{mc}_t + \beta\, \mathbb{E}_t\{\pi_{t+1}\}, \tag{A.7}$$

with the slope given by Equation (4) in the text:

$$\lambda := \frac{(1 - \theta)(1 - \beta\theta)}{\theta}(1 - \Omega)\Theta.$$

A.4 Derivation of the real marginal cost gap $\widehat{mc}_t$

To derive the aggregate real marginal cost gap we start from the derivation of real marginal cost at the firm-level, $mc_{fi,t}$, and aggregate by averaging across firms and industries. In doing so, we omit the firm and industry subscript (f, i) for ease of notation.

Each firm chooses a bundle of variable inputs $X_{j,t}$, $j = 1, \ldots, J$ (e.g., different types of labor, intermediate inputs, or capital services), with nominal user costs collected in the vector $W_t^n = (W_{1,t}^n, \ldots, W_{J,t}^n)$. Output is produced according to the technology

$$Y_t = A_t\, H(X_t)^{\alpha},$$

where A_t is Hicks-neutral productivity, $H(\cdot)$ is an aggregate-input index, and $\alpha > 0$ governs returns to scale in the mapping from aggregate inputs to output.

Throughout, a superscript $\star$ denotes the flexible-price benchmark allocation, holding fixed the real frictions that characterize the natural allocation. Hence, real wedges need not vanish in the flexible-price equilibrium. Lowercase letters denote natural

logarithms, and for any variable z_t we define $\widehat{z}_t \equiv z_t - z_t^\star$.

Cost function and nominal unit cost. Define the nominal cost function

$$C(Y_t, A_t, W_t^n, \Xi_t) \;\equiv\; \min_{X_t} \sum_{j=1}^{J} W_{j,t}^n X_{j,t} \quad \text{s.t.} \quad Y_t \le A_t \big(H(X_t)\big)^{\alpha}.$$

The term Ξ_t is a reduced-form placeholder for real input-market frictions (e.g., bargaining wedges, markups embedded in intermediate-input prices, or financing premia embedded in user costs). These frictions are not assumed away under flexible prices, so $\Xi_t^\star$ is generally nonzero.

Assume that the aggregate-input index $H(\cdot)$ is homogeneous of degree one. Then producing Y_t units of output requires an aggregate input level satisfying $H(X_t) \;\ge\; (Y_t/A_t)^{1/\alpha}$. By homogeneity, the cost-minimization problem admits a two-stage representation:

$$C(Y_t, A_t, W_t^n, \Xi_t) = \left(\frac{Y_t}{A_t}\right)^{1/\alpha} \cdot \Gamma(W_t^n, \Xi_t),$$

where $\Gamma(W_t^n, \Xi_t)$ is the minimum nominal cost of producing one unit of the aggregate input H.

Nominal marginal cost. Nominal marginal cost is defined as:

$$MC_t^n \;\equiv\; \frac{\partial C(Y_t, A_t, W_t^n, \Xi_t)}{\partial Y_t}.$$

Differentiating with respect to Y_t yields:

$$MC_t^n = \Gamma(W_t^n, \Xi_t) \cdot \frac{\partial}{\partial Y_t}\left(\frac{Y_t}{A_t}\right)^{1/\alpha} = \Gamma(W_t^n, \Xi_t) \cdot \frac{1}{\alpha}\left(\frac{Y_t}{A_t}\right)^{1/\alpha} \frac{1}{Y_t}.$$

Equivalently,

$$MC_t^n = \frac{1}{\alpha}\,\Gamma(W_t^n, \Xi_t)\, Y_t^{\frac{1}{\alpha}-1}\, A_t^{-\frac{1}{\alpha}}.$$

It is convenient to define $\chi \equiv (1/\alpha) - 1$, so that $\chi > 0$ corresponds to decreasing returns to scale ($\alpha < 1$), while $\chi < 0$ corresponds to increasing returns to scale ($\alpha > 1$). We then have that:

$$MC_t^n = \frac{1}{\alpha}\,\Gamma(W_t^n, \Xi_t)\, Y_t^{\chi}\, A_t^{-(1+\chi)}.$$

Taking logs, we obtain

$$mc_t^n \;\equiv\; \log MC_t^n = \log \Gamma(W_t^n, \Xi_t) + \chi\, y_t - (1+\chi)\, a_t^{tfp} - \log \alpha,$$

where $a_t^{tfp} \equiv \log A_t$ denotes log technical efficiency.

Real user-cost index and wedges. We decompose the real unit cost implied by the cost function into a component reflecting equilibrium factor prices and wedges:

$$\log \Gamma(W_t^n, \Xi_t) - p_t = \left(\log \Gamma(W_t^n, 0) - p_t\right) + \tau_t = (w_t^n - p_t) + \tau_t,$$

where $w_t^n \equiv \log \Gamma(W_t^n, 0)$ denotes the (optimized) nominal unit cost implied by factor prices in the absence of the additional real frictions; τ_t captures distortions summarized by Ξ_t, that shift effective user costs independently of equilibrium factor prices. Such wedges may reflect, among others, bargaining frictions in labor markets, financing premia embedded in user costs, markups in intermediate-input prices, or other input-market distortions. Importantly, τ_t represents an aggregate (or average) factor-market wedge in the spirit of Chari et al. (2007). These aggregate wedges are conceptually distinct from misallocation wedges arising from cross-sectional heterogeneity in the frictions faced by individual firms when accessing factor markets (Hsieh and Klenow 2009; Baqaee et al. 2024). We capture such misallocation effects instead through the productivity disturbance $\tilde{a}_t$, defined below.

Up to first order, we can express the user-cost index w_t^n and the factor market's wedge τ_t as a function of individual factors' prices and wedges. To see this, recall that the nominal cost function admits the representation $C(Y_t, A_t, W_t^n, \Xi_t) = (Y_t/A_t)^{1/\alpha}\, \Gamma(W_t^n, \Xi_t)$. Because the scale term $(Y_t/A_t)^{1/\alpha}$ does not depend on individual input prices, Shephard's lemma applies directly to $\Gamma(W_t^n, \Xi_t)$, the minimum nominal cost of producing one unit of the aggregate input $H(\cdot)$. Let $c_{j,t}$ denote the cost share of input j:

$$c_{j,t} \;\equiv\; \frac{W_{j,t}^n X_{j,t}}{C(Y_t, A_t, W_t^n, \Xi_t)} = \frac{\partial \log \Gamma(W_t^n, \Xi_t)}{\partial \log W_{j,t}^n}.$$

Evaluating this expression at the flexible-price benchmark yields the reference cost shares $c_j^\star$, which are time-invariant to a first-order approximation because they are computed at the natural allocation. Taking a first-order Taylor expansion of $\log \Gamma(W_t^n, \Xi_t)$ around the flexible-price allocation $(W_t^{n\star}, \Xi_t^\star)$ and subtracting $(p_t - p_t^\star)$, we obtain the first-order approximation:

$$\widehat{w}_t \;\equiv\; w_t - w_t^\star \;\approx\; \sum_{j=1}^{J} c_j^\star\, \widehat{w}_{j,t}, \qquad \widehat{\tau}_t \;\equiv\; \tau_t - \tau_t^\star \;\approx\; \sum_{j=1}^{J} c_j^\star\, \widehat{\tau}_{j,t},$$

This decomposition clarifies the distinction between real equilibrium factor-price

pressures, summarized by $\widehat{w}_t$, and cyclical distortions in input markets that shift marginal costs independently of factor prices, summarized by $\widehat{\tau}_t$, both of which enter the marginal-cost gap that drives inflation dynamics.

Real marginal cost. Define real marginal cost as nominal marginal cost divided by the aggregate price level. In logs, $mc_t \equiv mc_t^n - p_t$. Substituting the expressions above yields:

$$mc_t = w_t + \tau_t + \chi\, y_t - a_t - \log \alpha,$$

where we define the *real unit cost index* as $w_t \equiv w_t^n - p_t$ and *effective productivity*—the component of efficiency relevant for marginal cost—as:

$$a_t \equiv (1+\chi)\, a_t^{tfp} + \tilde{a}_t.$$

The term $\tilde{a}_t$ captures cyclical distortions in technical efficiency arising from price dispersion, wage dispersion, misallocation of inputs (Hsieh and Klenow 2009; Baqaee et al. 2024), or congestion in factor markets. Under flexible prices, real marginal cost is $mc_t^\star = w_t^\star + \tau_t^\star + \chi\, y_t^\star - a_t^\star - \log \alpha$, where by construction, the productivity distortions vanish in the flexible-price allocation, so that $\tilde{a}_t^\star = 0$.

Define the *cost pressure index* $q_t \equiv \chi\, y_t + w_t + \tau_t$. Subtracting term by term the flexible-price benchmark from the time-t values, we obtain the decomposition of the primitive NKPC in terms of the real marginal cost gap $\widehat{mc}_t$ used in Equation (1) in the paper:

$$\begin{aligned} \widehat{mc}_t &\equiv mc_t - mc_t^\star \\ &= \widehat{q}_t - \widehat{a}_t \\ &= \chi\, \widehat{y}_t + \widehat{w}_t + \widehat{\tau}_t - \widehat{a}_t \end{aligned} \tag{A.8}$$

This expression shows that deviations of real marginal cost from its flexible-price benchmark reflect four distinct forces: scale effects associated with deviations of output from potential ($\chi\, \widehat{y}_t$), cyclical movements in real input prices ($\widehat{w}_t$), cyclical input-market wedges ($\widehat{\tau}_t$), and deviations in effective productivity relative to the flexible-price allocation ($\widehat{a}_t$). The latter captures efficiency losses due to misallocation and dispersion induced by nominal rigidities.

Aggregation. Aggregating across firms and industries, we obtain the aggregate (i.e., average) real marginal cost gap that enters the primitive NKPC as a forcing variable.

$$\widehat{mc}_t \equiv \int_i \frac{1}{N} \sum_{f=1}^{N} mc_{fit} \, di - \int_i \frac{1}{N} \sum_{f=1}^{N} mc^{\star}_{fit} \, di$$

Recall that we assumed that each firm faces the same input prices and that the wedges τ_t are aggregate wedges, which implies that $w_{fit} = w_t$ and $\tau_{fit} = \tau_t$. Thus, the construction of aggregate real marginal cost requires averaging only across firms' output and realized productivity (y_{fit}, a_{fit}).