ENDOGENOUS DECENTRALIZATION AND THE ECONOMICS OF AUTONOMOUS AGENT NETWORKS

A Unified CES Framework for Technology, Organization, and Monetary Infrastructure

Jon Smirl

February 2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Economic Question

Between 2018 and 2025, the five largest US technology companies—Alphabet, Amazon, Apple, Meta, and Microsoft—together with Oracle and the Stargate joint venture, committed an estimated $1.3 trillion in cumulative capital expenditure to construct centralized AI infrastructure. This represents the largest concentrated infrastructure investment in history outside wartime mobilization. The facilities house tens of thousands of specialized GPUs consuming gigawatts of power, connected by proprietary high-bandwidth networks, and operated as vertically integrated cloud services. The near-term business objective is to sell AI inference—running trained models to serve user requests—at premium margins. A longer-horizon objective is frontier model training at scales that may produce discontinuous capability advances.

This thesis asks a simple question: *does this investment finance its own disruption?*

The question is not rhetorical. Prior infrastructure investments have occasionally exhibited self-undermining dynamics—railroad investment enabled trucking by financing steel and road-building capacity, AT&T's Bell Labs produced the transistor that ultimately ended the analog telephone monopoly—but these were historical accidents, not structural mechanisms. The thesis argues that the current AI infrastructure buildout is different: the self-undermining dynamic is *endogenous*, meaning it follows from the structure of the investment itself, not from exogenous shocks or historical contingency.

The argument has four steps, each formalized in a separate chapter:

1. Concentrated investment finances component learning curves—particularly in 3D memory stacking and advanced packaging—that reduce the cost of distributed inference hardware (Chapter 4).

2. When distributed hardware crosses a cost threshold, heterogeneous specialized AI agents self-organize into a mesh economy whose collective capability exceeds centralized provision (Chapter 5).

3. The mesh economy requires programmable settlement infrastructure for routing compensation among autonomous agents, creating demand for dollar-denominated stablecoins backed by US Treasuries (Chapter 6).

4. Stablecoin demand transforms the monetary infrastructure that funds and settles centralized investment, feeding back to constrain sovereign fiscal capacity and alter the environment in which the next round of centralized investment decisions is made (Chapter 6).

The four steps form a cycle, not a sequence. The output of Step 4 feeds back into the conditions governing Step 1. The thesis's contribution is to show that this cycle is governed by a single mathematical structure—the CES production function—operating at four timescales with strict separation. The unified framework produces falsifiable predictions, identifies the structural parameters that determine whether the transition occurs, and derives policy principles that follow from theorems rather than from intuition.

1.2 The Thesis Argument in One Page

The thesis can be stated in four paragraphs corresponding to the four levels of the hierarchy.

Level 1: Concentrated investment finances learning curves (Chapter 4). Competing centralized AI firms face a common-pool problem: cumulative component production $Q(t)$ drives down costs via Wright's Law ($C(Q) \propto Q^{-\alpha}$, $\alpha \approx 0.23$), but cost reduction benefits all producers, including future distributed entrants. In symmetric Markov Perfect Equilibrium, N firms collectively overproduce by a factor of 3–4× relative to the cooperative optimum, accelerating the crossing time T^* by approximately 79%. Each firm would prefer slower progress, but no firm can unilaterally slow down without losing market share. The mechanism applies specifically to *inference* workloads, which constitute 80–90% of AI compute. Training may remain permanently centralized due to synchronization constraints that cost reduction alone cannot address. The effective crossing threshold is approached from two directions: from below by the packaging learning curve and from above by algorithmic efficiency gains driven by open-weight developers operating under export-control-imposed compute constraints.

Level 2: The mesh economy forms and grows (Chapter 5). After crossing, a self-organizing mesh of heterogeneous specialized agents exceeds centralized provision above a finite critical mass N^*. The phase transition is first-order (discontinuous): the mesh crystallizes rather than forming gradually, characterized by Fortuin-Kasteleyn/Potts dynamics when the number of specialization types exceeds two. Once formed, the mesh's capability grows endogenously through autocatalytic training—agents improve other agents—but converges to the frontier training rate set by centralized providers (the Baumol bottleneck). The mesh's CES heterogeneity ($\rho < 1$) prevents model collapse by maintaining the effective external data fraction above the critical threshold, even when agents train partially on synthetic data.

Level 3: Settlement demand transforms monetary infrastructure (Chapter 6). The mesh requires a programmable settlement layer for routing compensation among autonomous agents operating at machine speed. Dollar stablecoins backed by US Treasuries

are the efficient settlement medium. This creates a coupled dynamical system: mesh growth increases stablecoin demand, which increases Treasury absorption, which improves settlement infrastructure, which accelerates mesh growth. Monetary policy tools degrade in a specific sequence—forward guidance first, then quantitative easing, then financial repression—because each tool depends on a friction that mesh participation eliminates. The system admits two stable equilibria: the current low-mesh state and a high-mesh state where monetary policy is weak but continuous market discipline substitutes.

The cycle closes. Stablecoin-driven Treasury absorption alters sovereign fiscal capacity and debt dynamics, feeding back to the macroeconomic environment in which centralized investment decisions at Level 1 are made. The master reproduction number R_0 of the entire system depends on the product of cross-level amplification factors. The transition from the low-activity equilibrium to the high-activity equilibrium takes approximately 8 years at current semiconductor improvement rates.

The mathematical unity of this argument—the fact that a single CES aggregate controls all four levels—is not a modeling convenience. It is the thesis's central claim, established in Chapters 2 and 3.

1.3 The Mathematical Spine: CES Geometry

The entire thesis rests on a single functional form: the CES (Constant Elasticity of Substitution) aggregate

$$F_n \;=\; \Big(\frac{1}{J}\sum_{j=1}^{J} x_{nj}^{\rho}\Big)^{1/\rho}, \qquad \rho < 1, \quad \rho \neq 0, \tag{1.1}$$

applied at each level n of the hierarchy, where x_{nj} denotes the j-th input at level n, $J \geq 2$ is the number of inputs, and ρ is the substitution parameter (related to the elasticity of substitution by $\sigma = 1/(1-\rho)$). The CES free energy $\Phi = -\sum_n \log F_n$ serves as the Hamiltonian of the hierarchical system.

1.3.1 The curvature parameter

The key quantity is the **curvature parameter** (Chapter 2, Definition 3.1):

$$K \;=\; (1-\rho)\,\frac{J-1}{J}. \tag{1.2}$$

This is the normalized principal curvature of the CES isoquant at the cost-minimizing (symmetric) point. The parameter K ranges from 0 (perfect substitutes, $\rho = 1$) upward as

complementarity increases (ρ decreases below 1). It encodes both the degree of complementarity and the number of inputs in a single dimensionless number.

1.3.2 The triple role

Chapter 2 proves that K simultaneously controls three properties that have previously been studied separately using different techniques (Chapter 2, Theorem 7.1):

1. **Superadditivity** (Chapter 2, Theorem 4.1). Combining heterogeneous input bundles produces more output than the sum of separate productions. The superadditivity gap is bounded below by $\Omega(K)$ times a geodesic diversity measure. This is a first-order curvature effect: the isoquant bends away from the hyperplane, so convex combinations of diverse points lie above the level set.

2. **Correlation robustness** (Chapter 2, Theorem 5.1). The CES aggregate extracts idiosyncratic variation from correlated inputs that a linear aggregate would miss. The effective dimensionality exceeds the linear baseline by $\Omega(K^2)$ times an idiosyncratic variation term. This is a second-order curvature effect: the nonlinear mapping separates input trajectories that a linear function would conflate.

3. **Strategic independence** (Chapter 2, Theorem 6.1). The balanced allocation at the cost-minimizing point is a Nash equilibrium: no coalition of input suppliers can profitably redistribute or withhold inputs. The strategic manipulation gain is bounded above by $-\Omega(K)$ times a squared deviation. This is again a first-order curvature effect: deviations from the balanced allocation move along the isoquant into regions of lower marginal product.

All three bounds tighten monotonically in K. When $K = 0$ (perfect substitutes), all three vanish simultaneously. The three properties are not three theorems sharing a common assumption—they are the same geometric fact, the curvature of the isoquant, viewed from aggregation theory, information theory, and game theory respectively.

The results extend to general (unequal) CES weights $a_j > 0$ via the secular equation of the weighted inverse-share matrix (Chapter 2, Theorem 8.5). With unequal weights, the principal curvatures of the isoquant are no longer degenerate; they are determined by the $J-1$ roots of the secular equation, whose smallest root $R_{\min}$ controls the generalized curvature parameter $K(\rho, \mathbf{a})$.

1.3.3 From within-level geometry to between-level architecture

Chapter 2 establishes the within-level properties of CES aggregation. Chapter 3 asks: given that each level aggregates its inputs via CES, what is the structure of interaction *between* levels? The answer is that CES geometry derives the architecture—it is not a free modeling choice.

Three architectural constraints follow from the geometry (Chapter 3, Theorem 3.1):

1. **Aggregate coupling.** Each level communicates with other levels only through its aggregate output F_n. Individual input allocations x_{nj} within a level are invisible to other levels. This is a consequence of the CES isoquant's invariance under permutations of inputs at the symmetric point.

2. **Directed feed-forward.** The coupling between levels must be asymmetric: level $n-1$'s output enters level n's production function, but not vice versa at the same timescale. Reciprocal coupling would violate the timescale separation that makes the hierarchy well-defined.

3. **Nearest-neighbor topology.** Long-range cross-level links—level 1 directly affecting level 4, for example—have no effect on the system's qualitative dynamics. The architecture is a nearest-neighbor chain, with each level coupled only to its immediate neighbors in the timescale ordering.

Furthermore, the **Moduli Space Theorem** (Chapter 3, Theorem 3.2) characterizes the set of all models consistent with the CES geometry: the substitution parameter ρ at each level determines the qualitative dynamics, while the timescales, damping coefficients, and specific gain functions are free parameters that do not affect the model's qualitative behavior. This means the framework's predictions about phase transitions, equilibrium structure, and policy are robust to reasonable parameter uncertainty.

1.3.4 The eigenstructure bridge

The deepest result connecting the mathematical framework to welfare is the **Eigenstructure Bridge** (Chapter 3, Theorem 6.3):

$$\nabla^2\Phi\big|_{\text{slow}} \;=\; W^{-1}\,\nabla^2 V, \tag{1.3}$$

where $\Phi = -\sum_n \log F_n$ is the CES free energy (the technology side), V is the Lyapunov function measuring welfare loss (the welfare side), and W is the institutional supply-rate

matrix encoding how efficiently each level adjusts. The Hessian of the technology surface, restricted to the slow manifold, equals the Hessian of the welfare loss function scaled by the inverse of institutional adjustment speed.

This identity has a striking implication: the directions in which the economy adjusts fastest technologically are the directions in which welfare losses are most sensitive to institutional rigidity. The binding welfare constraint is the most institutionally *rigid* level, not the most *visible* disequilibrium. A sector may exhibit large price distortions yet contribute little to welfare loss if its institutional adjustment is fast; conversely, a sector with small visible distortions may be the dominant welfare bottleneck if its institutions are slow.

1.4 The Four-Level Hierarchy

The framework operates on four levels with strict timescale separation ($\varepsilon_1 \gg \varepsilon_2 \gg \varepsilon_3 \gg \varepsilon_4$, where ε_n denotes the characteristic adjustment speed of level n from slowest to fastest):

Table 1.1: The four-level hierarchy

Level	Chapter	State Variable	Gain Function φ_n	Timescale
1 (slowest)	4	Hardware/semiconductor cost	Wright's Law learning curve ($\alpha \approx 0.23$)	Decades
2	5	Heterogeneous AI agent density	Recruitment/adoption dynamics	Years
3	5	Training effectiveness/capability	Autocatalytic training feedback	Months
4 (fastest)	6	Stablecoin infrastructure	Settlement demand	Days–weeks

1.4.1 Timescale separation and the slow manifold

The timescale separation is not merely an ordering convenience; it is a structural feature that determines the system's dynamics. When $\varepsilon_{n-1}/\varepsilon_n \ll 1$, the fast level n equilibrates before the slow level $n-1$ changes appreciably. The fast level's dynamics can therefore be solved conditional on the slow level's state, and the slow level sees only the equilibrium response of the fast level. This is the standard singular perturbation / slow-manifold reduction, applied here to the economic hierarchy.

The practical consequence is that each level's equilibrium is bounded from below by the level beneath it in the hierarchy. No matter how rapidly the mesh economy grows (Level 2), its long-run growth rate cannot exceed the rate of hardware cost decline (Level 1). No matter

how efficiently stablecoin infrastructure scales (Level 4), its capacity is bounded by the mesh economy's settlement demand (Levels 2–3).

1.4.2 The hierarchical ceiling cascade

This bounding structure is formalized as the **Hierarchical Ceiling Cascade** (Chapter 3, Proposition 8.1): each level n's steady-state output satisfies

$$F_n^* \;\leq\; g\big(F_{n-1}^*\big), \tag{1.4}$$

where $g(\cdot)$ is a monotone transformation determined by the cross-level coupling. The long-run growth rate of the entire system equals the growth rate of the slowest level—hardware cost decline under Wright's Law. This is not an assumption but a consequence of the timescale separation and the CES architecture.

Two classical results in economics emerge as special cases of the ceiling cascade at adjacent levels. The **Baumol bottleneck** (Chapter 5)—the mesh's capability growth rate converging to the frontier training rate set by centralized providers—is the ceiling constraint between Levels 2–3 and Level 1. The **Triffin contradiction** (Chapter 6)—the tension between the dollar's domestic monetary policy role and its international settlement role—is the ceiling constraint between Level 4 and Levels 2–3. These are mathematically the same object: a slow-manifold constraint at adjacent layers in the hierarchy.

1.5 The Master Reproduction Number

Each transition in the hierarchy requires a **basic reproduction number** exceeding unity. At Level 1, the crossing condition is $R_0 > 1$, where R_0 generalizes pure cost parity to include the self-sustaining adoption dynamics of the distributed ecosystem. At Level 2, the mesh forms when $R_0^{\text{mesh}} > 1$ (giant component via percolation). At Level 4, the settlement feedback becomes self-reinforcing when $R_0^{\text{settle}} > 1$.

Chapter 3 (Theorem 4.3) proves a **spectral activation threshold**: the hierarchy sustains non-trivial activity if and only if the spectral radius of the **next-generation matrix** $\mathcal{K}$ exceeds unity:

$$\rho(\mathcal{K}) > 1, \qquad \text{where } \mathcal{K}_{nm} = \frac{\beta_n \, \varphi_n}{\delta_n} \cdot \mathbf{1}_{[m=n-1]}. \tag{1.5}$$

Here β_n is the cross-level coupling strength (how much Level $n-1$'s output amplifies Level n's gain), φ_n is the within-level gain function, and δ_n is the within-level decay rate. The indicator $\mathbf{1}_{[m=n-1]}$ reflects the nearest-neighbor topology derived in Chapter 3.

The crucial feature of this threshold is that individual levels can each be sub-threshold—$\beta_n \varphi_n / \delta_n < 1$ for every n—while the system as a whole is super-threshold through cross-level amplification. Intuitively, Level 1's output amplifies Level 2, whose output amplifies Level 3, whose output amplifies Level 4, whose output feeds back to Level 1. The cycle product

$$P_{\text{cycle}} = \prod_{n=1}^{4} \frac{\beta_n \, \varphi_n}{\delta_n} \tag{1.6}$$

governs the aggregate threshold: $P_{\text{cycle}} > 1$ is necessary and sufficient for the system to sustain the high-activity equilibrium. The system activates when the weakest cross-level link is strong enough that the cycle product exceeds unity, even if no single level could sustain activity on its own.

This has a practical implication for the current AI infrastructure buildout. The endogenous decentralization mechanism at Level 1 may appear insufficient by itself: the crossing point seems distant, the learning curve uncertain, the distributed ecosystem immature. But the relevant threshold is not Level 1 in isolation—it is the cycle product across all four levels. Settlement infrastructure improvements (Level 4), mesh network effects (Level 2), and autocatalytic capability growth (Level 3) all contribute multiplicatively. The system can activate from a combination of modest progress at every level.

1.6 Summary of Contributions by Chapter

1.6.1 Chapter 2: The CES triple role—within-level geometry

Chapter 2 proves that three important properties of CES aggregation—superadditivity, correlation robustness, and strategic independence—are controlled by the single curvature parameter $K = (1-\rho)(J-1)/J$. The superadditivity gap is $\Omega(K) \cdot$ diversity (first-order curvature effect). The correlation robustness bonus is $\Omega(K^2) \cdot$ idiosyncratic variation (second-order curvature effect). The strategic manipulation penalty is $-\Omega(K) \cdot \text{deviation}^2$ (first-order curvature effect). All three vanish simultaneously at $K = 0$. The underlying mechanism is the curvature of the isoquant: these are three views of a single geometric object, not three consequences of a shared assumption. The results extend to general weights via the secular equation of the weighted inverse-share matrix. Chapter 2 provides the mathematical foundation that all subsequent chapters assume.

1.6.2 Chapter 3: Complementary heterogeneity—between-level architecture and dynamics

Chapter 3 takes the CES triple role as given and asks what happens between sectors in a hierarchical economy. Five results follow. The Port Topology Theorem derives the architecture from CES geometry: aggregate coupling, directed feed-forward, and nearest-neighbor topology. The Moduli Space Theorem characterizes which modeling choices affect qualitative dynamics (ρ) and which do not (timescales, damping, gain functions). The spectral activation threshold shows that cross-level amplification can activate the system even when individual levels are sub-threshold. The welfare distance function attributes inefficiency to each level, with the binding constraint at the most institutionally rigid level. The damping cancellation theorem shows that tightening regulation at any level has zero net welfare effect—reform must target upstream. The transition takes $O(1/\sqrt{\varepsilon_{\text{drift}}})$ time, yielding approximately 8 years at Wright's Law rates.

1.6.3 Chapter 4: Endogenous decentralization—Level 1 (hardware learning curves)

Chapter 4 formalizes the mechanism at Level 1 as a continuous-time differential game. The distance to the crossing point is a common-pool state variable depleted by cumulative production. In symmetric Markov Perfect Equilibrium, N firms collectively overproduce by 3–4×, accelerating the crossing time T^* by approximately 79% relative to the cooperative optimum. The pure cost-parity crossing condition generalizes to $R_0 > 1$, incorporating self-sustaining adoption dynamics. A structural distinction between training and inference workloads predicts partial decentralization: inference distributes while training persists centrally. Cross-domain empirical analysis identifies the operative learning curve as 3D memory stacking and advanced packaging ($\alpha = 0.23$), not planar DRAM die fabrication. US–China semiconductor export controls provide a natural experiment distinguishing the endogenous mechanism from standard Arrow learning-by-doing. The effective crossing threshold is simultaneously reduced from above by algorithmic efficiency gains. Nine falsifiable predictions are derived, including hardware crossing circa 2028 and self-sustaining distributed adoption circa 2030–2032.

1.6.4 Chapter 5: The mesh economy—Levels 2–3 (network formation and capability growth)

Chapter 5 covers two levels of the hierarchy. At Level 2, after the crossing point, heterogeneous specialized AI agents self-organize into a mesh whose collective capability exceeds centralized provision above a finite critical mass N^*. The phase transition is first-order (discontinuous), characterized by Fortuin-Kasteleyn/Potts crystallization when the number of specialization types exceeds two. The mesh equilibrium is proved to exist, be unique, and be locally asymptotically stable. The crossing point corresponds to an inverse Bose-Einstein condensation in the network fitness model.

At Level 3, the mesh's capability grows endogenously through autocatalytic training. Three growth regimes are characterized: convergence to a ceiling (most likely near-term), exponential growth, and finite-time singularity. The Baumol bottleneck—the mesh's growth rate converging to the frontier training rate—emerges endogenously from the dynamics. CES heterogeneity prevents model collapse: the effective external data fraction α_{eff} remains above the critical threshold α_{crit} even when agents train partially on synthetic data, because the same curvature parameter K that governs the diversity premium also governs informational robustness (the connection to Chapter 2's correlation robustness theorem). The mesh's routing and compensation requirements endogenously generate the need for a programmable settlement layer, providing the connection to Chapter 6.

1.6.5 Chapter 6: Settlement feedback—Level 4 (monetary and financial infrastructure)

Chapter 6 formalizes the coupling between the mesh economy and the monetary system as a system of four coupled ODEs in mesh participation ϕ, stablecoin ecosystem size S, Treasury debt ratio b, and financial sector capitalization η. As mesh participation increases, market efficiency approaches the Grossman-Stiglitz limit; Kyle's price impact λ is non-monotone in ϕ (depth first improves, then deteriorates as noise trading exits). Monetary policy tools degrade in a specific sequence: forward guidance (which depends on information processing delay), then quantitative easing (which depends on arbitrage speed), then financial repression (which depends on captive savings). Each tool depends on a friction that mesh participation eliminates.

The system admits two stable equilibria: a low-mesh equilibrium (the current system, approximately) and a high-mesh equilibrium in which monetary policy is weak but real-time market discipline substitutes. The transition between them is governed by the settlement

reproduction number R_0^{settle}: the transition is self-reinforcing when each unit of mesh growth produces more than one unit of subsequent growth through the financial channel. The high-mesh equilibrium constrains sovereign fiscal policy through continuous market discipline—a "synthetic gold standard" that emerges from the model. The paper extends the Uribe (1997) hysteresis model of dollarization with endogenous stablecoin access, showing that the bifurcation thresholds are decreasing in stablecoin ecosystem size.

1.6.6 Chapter 7: Monetary productivity gap—empirical evidence for settlement demand

Chapter 7 provides micro-level empirical evidence for the settlement layer demand modeled in Chapter 6. Using a 41-country panel, the chapter constructs a Fiat Quality Index (FQI) from five components: inflation stability, banking access, ATM density, government effectiveness, and internet penetration. The key finding is a 13:1 remittance cost ratio between fiat and stablecoin channels in Sub-Saharan Africa, documenting the scale of the efficiency gap that creates demand for programmable settlement.

The India 2022 tax natural experiment—a 30% capital gains tax plus 1% TDS on cryptocurrency transactions—provides causal identification: domestic volume fell 86% but 72% of activity displaced offshore rather than being suppressed, demonstrating revealed preference for settlement infrastructure. The Yield Access Gap (YAG) regression produces a within-country coefficient of $\beta = +0.248$ ($p < 0.001$), with the sign flip from the cross-sectional regression confirming a precautionary savings motive: within countries, regions with worse fiat quality show higher stablecoin adoption, consistent with the settlement demand channel.

1.6.7 Chapter 8: Fair inheritance—policy implications

Chapter 8 addresses the distributional consequences of the technological transition modeled in Chapters 4–6. A recipient-based inheritance tax replacing the current estate tax creates a binary choice: pay tax on concentrated transfers or disperse wealth widely through a zero-tax pathway. The proposal treats inheritance as ordinary income and eliminates trust recognition. Revenue is projected at \$85–135 billion per year, approximately 5× the current estate tax yield. The connection to the thesis is that automation-era wealth concentration—driven by the same learning curves and network effects modeled in the preceding chapters—requires distributional policy that accounts for the structural mechanisms generating concentration.

1.7 Falsifiable Predictions

A theory that cannot be falsified is not a theory. The framework generates specific predictions with timing, quantitative thresholds, and failure conditions. Table 1.2 collects the principal predictions from Chapters 4–6 and Chapter 3.

Several features of this prediction set merit emphasis.

First, the predictions are *ordered*. Prediction 1 (hardware crossing) must occur before Prediction 5 (mesh critical mass), which must occur before Prediction 9 (Treasury absorption). The ordering follows from the timescale separation. If a lower-numbered prediction fails, higher-numbered predictions are invalidated. This creates a clear sequential falsification structure.

1.7.1 Preliminary empirical status

The thesis's contribution is the unified mathematical framework, not definitive empirical proof of each prediction—most predictions target 2027–2040, and the data required to test them decisively does not yet exist. However, six preliminary tests using currently available data demonstrate that the predictions are (a) testable with standard econometric methods, (b) directionally consistent with the data across all six tests, and (c) statistically significant for the subset of predictions that have already had time to manifest.

The strongest results are the capex overinvestment test (Chapter 4), which reveals a clean structural break at 2022: the basic learning-curve game fits the pre-AI period while the option-augmented model with a superintelligence prize fits the post-ChatGPT period; and the block exogeneity test (Chapter 3), which confirms that distant layers in the hierarchy do not directly couple—the most distinctive prediction of the port topology theorem. The weakest results are the settlement-layer tests (Chapter 6), which target phenomena at a scale ($1T+ stablecoin ecosystem) that has not yet been reached. These tests establish baselines against which future acceleration can be measured as the sample period extends.

Second, Prediction 6 (first-order phase transition) is the sharpest test of the framework. Standard technology adoption models predict gradual S-curve diffusion. This framework predicts discontinuous crystallization. The difference is empirically distinguishable: gradual adoption produces a smooth time series of adoption rates, while first-order transition produces a regime change with a clear before/after.

Third, Prediction 12 (transition duration) follows from the theoretical framework's deepest result—the canard bifurcation analysis of Chapter 3 (Theorem 8.2). The duration scales as $O(1/\sqrt{\varepsilon_{\text{drift}}})$, where $\varepsilon_{\text{drift}}$ is the rate of secular improvement. At Wright's Law semiconduc-

Table 1.2: Summary of falsifiable predictions

#	Prediction	Chapter	Horizon
1	Hardware crossing: consumer hardware capable of interactive 70B+ inference at mass-market prices (<$500 incremental cost).	4	∼2028
2	Self-sustaining distributed adoption: $R_0 > 1$ for the distributed ecosystem, independent of continued centralized investment.	4	∼2030–2032
3	DRAM/HBM overcapacity: advanced packaging capacity exceeds demand, producing below-trend consumer memory pricing.	4	2027–2029
4	Training-inference bifurcation: training remains centralized even as inference distributes; partial, not complete, decentralization.	4	Ongoing
5	Mesh critical mass N^*: above a finite number of heterogeneous agents, the mesh equilibrium exists and dominates.	5	2030–2035
6	First-order phase transition: mesh adoption is discontinuous (crystallization), not gradual (logistic). Regional adoption will exhibit sharp jumps.	5	2029–2033
7	Baumol bottleneck: mesh capability growth converges to the frontier training rate g_Z, not to a higher endogenous rate.	5	2032–2040
8	Model collapse protection: diverse meshes avoid capability degradation from synthetic data; homogeneous networks do not.	5	Observable now
9	Stablecoin Treasury absorption: stablecoin reserves become a material fraction of short-duration Treasury demand (>5% of T-bills).	6	2027–2030
10	Monetary policy degradation: forward guidance effectiveness declines first, QE second, financial repression last.	6	2028–2040
11	Kyle's λ non-monotonicity: market depth first improves then deteriorates as autonomous agent participation ϕ increases.	6	Observable now
12	Transition duration: ∼8 years from hardware crossing to new monetary equilibrium, via canard bifurcation dynamics.	3	2028–2036

Table 1.3: Preliminary empirical tests: summary of results

Test	Chapter	Key result	Status
Capex overinvestment	4	Pre-AI: 2.8–3.9× (matches Prop. 1); post-2022: 11–19× (explained by ASI option)	Consistent
Export control DID	4	$\hat{\delta} > 0$ across all specs; cross-section $p = 0.005$	Directional
Stablecoin–Treasury	6	First-diff. $\hat{\beta} = -0.30$ ($p = 0.092$); scale still small (5% of T-bills)	Marginal
FOMC absorption speed	6	Post-2020 era highest mean; trend +0.004/yr ($p = 0.60$)	Directional
Inflation threshold	6	Time-varying threshold declining ($\rho = -0.40$); sample too short	Directional
Cross-layer VAR topology	3	Block exogeneity: no L1→L4 ($p = 0.40$); FEVD: 2/3 layers consistent	Mixed

tor improvement rates ($\approx 15\%$ annual cost decline), this yields approximately 8 years. The prediction is robust to parameter uncertainty because the square-root scaling compresses the sensitivity: a 2× change in the drift rate changes the duration by only $\sqrt{2} \approx 1.4\times$.

Fourth, some predictions are already partially testable. Prediction 4 (training-inference bifurcation) is consistent with the observed pattern as of early 2026: inference is increasingly distributed through open-weight models running on consumer hardware, while frontier training remains concentrated in a handful of hyperscaler facilities. Prediction 8 (model collapse protection) can be tested using existing data on model performance degradation under synthetic data training. Prediction 11 (non-monotonic market depth) can be tested using high-frequency market microstructure data as algorithmic trading participation has increased.

1.8 Roadmap

The thesis proceeds as follows.

Chapter 2 establishes the mathematical foundation. The CES triple role theorem proves that a single curvature parameter K controls superadditivity, correlation robustness, and strategic independence. These are within-level properties: they characterize how heterogeneous inputs combine within a single sector. The chapter develops the isoquant geometry from first principles, proves the three results, unifies them as views of the same curvature, and extends to general weights. No economic application is developed; the chapter is pure microeconomic theory.

Chapter 3 builds the between-level architecture. Taking the CES triple role as given,

the chapter asks what structure of cross-sector interaction is consistent with CES geometry. The Port Topology Theorem derives the architecture (aggregate coupling, directed feed-forward, nearest-neighbor topology). The Moduli Space Theorem characterizes which modeling choices matter and which do not. The spectral activation threshold, welfare decomposition, eigenstructure bridge, damping cancellation, hierarchical ceiling cascade, and transition duration are all proved. This chapter provides the theoretical scaffolding for Chapters 4–6: it tells us that the four-level hierarchy has a specific architecture, a computable activation threshold, and a predictable transition duration, all derived from the CES geometry rather than assumed.

Chapter 4 applies the framework to Level 1. The endogenous decentralization mechanism is formalized as a continuous-time differential game with exact closed-form Nash equilibrium. The chapter provides both the theoretical model (the overinvestment result, the generalized R_0 crossing condition, the training-inference bifurcation) and the empirical evidence (HBM learning curve estimation, dual convergence, the export-control natural experiment). The chapter can be read independently as a paper in industrial organization, but within the thesis it establishes the quantitative parameters—particularly the learning rate $\alpha \approx 0.23$ and the crossing time $T^* \approx 2028$—that feed into the hierarchy.

Chapter 5 applies the framework to Levels 2 and 3. The mesh equilibrium (Level 2) establishes what happens after the crossing point: heterogeneous agents self-organize, the phase transition is first-order, and there exists a critical mass above which the mesh dominates. The autocatalytic mesh (Level 3) makes capability endogenous: agents improve other agents, but the growth rate converges to the frontier rate (Baumol bottleneck). The CES heterogeneity that makes the mesh productive (superadditivity) also makes it informationally robust (correlation robustness prevents model collapse). This is the chapter where the triple role theorem of Chapter 2 pays its largest dividend: what appears to be two separate results—the mesh is productive and the mesh avoids model collapse—are the same result applied to different roles of the curvature parameter.

Chapter 6 applies the framework to Level 4 and closes the cycle. The mesh's settlement requirements create stablecoin demand, stablecoin demand transforms monetary infrastructure, and monetary infrastructure feeds back to the macroeconomic environment. The four coupled ODEs, the monetary policy degradation sequence, the bistable equilibrium, and the synthetic gold standard are all derived. The Triffin contradiction appears as the ceiling cascade between Level 4 and Levels 2–3, mathematically identical in structure to the Baumol bottleneck one level below.

Chapter 7 provides empirical evidence. The 41-country panel, the India natural experiment, and the Yield Access Gap regression document the demand for programmable

settlement infrastructure—the micro-level phenomenon that Chapter 6 models at the macro level. The chapter can be read independently as an empirical study of fiat monetary quality and cryptocurrency adoption, but within the thesis it provides the quantitative evidence that the settlement demand channel of Level 4 is operative.

Chapter 8 addresses policy. The fair inheritance proposal responds to the distributional consequences of the technological transition analyzed in the preceding chapters. The wealth concentration driven by learning curves (Chapter 4), network effects (Chapter 5), and financial infrastructure transformation (Chapter 6) motivates policy that accounts for these structural mechanisms.

The logical dependence is: Chapter 2 → Chapter 3 → Chapters 4, 5, 6 (which can be read in any order, though the narrative flows best in sequence) → Chapter 7 (empirical complement to Chapter 6) → Chapter 8 (policy). A reader interested in the mathematical framework alone can read Chapters 2–3. A reader interested in a specific application can read the relevant chapter after reviewing Sections 1.3–1.5 of this introduction for the key definitions and results.

Chapter 2

The CES Triple Role: Superadditivity, Correlation Robustness, and Strategic Independence

Abstract. The CES production function is ubiquitous in economics, yet three of its important properties—superadditivity, robustness to correlated inputs, and resistance to strategic manipulation—have been studied separately using different techniques. This paper proves they are controlled by a single parameter: the **curvature parameter** $K = (1-\rho)(J-1)/J$, derived from the principal curvature of the CES isoquant at the cost-minimizing point. Superadditivity gap $= \Omega(K) \cdot$ diversity; correlation robustness bonus $= \Omega(K^2) \cdot$ idiosyncratic variation; strategic manipulation penalty $= -\Omega(K) \cdot \text{deviation}^2$. All three bounds tighten monotonically in K. When $K = 0$ (perfect substitutes), all three vanish simultaneously. The three properties are not three theorems sharing an assumption—they are the same geometric fact, the curvature of the isoquant, viewed from aggregation theory, information theory, and game theory. Results extend to general (unequal) CES weights via the secular equation of the weighted inverse-share matrix, whose smallest root $R_{\min}$ controls the generalized curvature parameter.

Keywords: CES production function, isoquant curvature, superadditivity, diversification, strategic independence, secular equation
JEL: C62, D24, D43, D81, L13

2.1 Introduction

The constant elasticity of substitution (CES) production function, introduced by Arrow, Chenery, Minhas, and Solow [1], is among the most widely used functional forms in economics. It appears in trade theory, industrial organization, macroeconomics with heterogeneous firms, and index number construction. The Dixit-Stiglitz [2] formulation of monopolistic competition, the Jones [6] analysis of directed technical change, and the Houthakker [5] aggregation results all rest on CES.

Three important properties of CES aggregation have been established in various settings:

1. *Superadditivity:* Combining diverse input bundles produces more output than the sum of separate productions. This matters for merger analysis, team formation, and gains from trade.

2. *Correlation robustness:* The CES aggregate preserves information about its components even when they are highly correlated. This matters for portfolio diversification, index construction, and aggregate measurement.

3. *Strategic independence:* Coalitions of input suppliers cannot profitably manipulate the aggregate by redistributing or withholding inputs. This matters for market design, mechanism design, and platform governance.

These properties have been proved using different techniques—superadditivity from convexity arguments, correlation robustness from second-order expansions, strategic independence from cooperative game theory. This paper demonstrates that all three are controlled by a single dimensionless parameter:

$$K \;=\; (1-\rho)\,\frac{J-1}{J} \tag{2.1}$$

where $\rho < 1$ is the CES substitution parameter and $J \geq 2$ is the number of components. This parameter is the normalized principal curvature of the CES isoquant at the cost-minimizing point.

The main result (Theorem 2.7.1) states:

(a) The superadditivity gap is bounded below by $\Omega(K)$ times a geodesic diversity measure (first-order curvature effect).

(b) The effective dimension under equicorrelation exceeds the linear baseline by $\Omega(K^2)$ times an idiosyncratic variation term (second-order curvature effect).

(c) The strategic manipulation gain is bounded above by $-\Omega(K)$ times a squared deviation (first-order curvature effect).

All three bounds tighten monotonically in K. When $K = 0$ (perfect substitutes, $\rho = 1$), all three vanish.

The underlying mechanism is the same in all three cases: the isoquant is curved. Curvature forces convex combinations of diverse points above the level set (superadditivity), maps correlated input variation into distinct output regions through a nonlinear channel

(informational diversity), and penalizes deviations from the balanced allocation (strategic stability). These are three views of a single geometric object, not three consequences of a common assumption.

The results extend to CES with general (unequal) weights $a_j > 0$ via the **secular equation** of the weighted inverse-share matrix. With unequal weights, the principal curvatures of the isoquant are no longer degenerate; they are determined by the $J-1$ roots of the secular equation, which interlace the inverse shares. The smallest root $R_{\min}$ controls a generalized curvature parameter $K(\rho, \mathbf{a})$ that replaces (2.1) in all three bounds.

Section 2.2 establishes notation. Section 2.3 proves the Curvature Lemma. Sections 2.4–2.6 prove the three results. Section 2.7 presents the unified perspective. Section 2.8 extends everything to general weights. Section 2.9 discusses tightness, prior results, and applications.

2.2 Setup and Notation

2.2.1 The CES Aggregate

For $J \geq 2$ components, the **CES aggregate** with weights $a_j > 0$ summing to 1 is

$$F(\mathbf{x}) = \left(\sum_{j=1}^{J} a_j \, x_j^{\rho} \right)^{1/\rho}, \qquad \mathbf{x} = (x_1, \ldots, x_J) \in \mathbb{R}_+^J \tag{2.2}$$

where $\rho < 1$, $\rho \neq 0$, is the substitution parameter. The **elasticity of substitution** is $\sigma = 1/(1-\rho)$.

We call the components *complements* when $\rho < 0$ ($\sigma < 1$) and *weak complements* when $0 < \rho < 1$ ($\sigma > 1$ but finite). The boundary $\rho \to 1$ gives perfect substitutes (linear); $\rho \to 0$ gives Cobb-Douglas; $\rho \to -\infty$ gives Leontief (perfect complements).

Remark 2.2.1. F is concave and homogeneous of degree 1 for all $\rho < 1$. Both properties are standard; we use them freely throughout.

For Sections 2.3–2.7 we work with **equal weights** $a_j = 1/J$. Section 2.8 presents the general-weight extension.

2.2.2 The Symmetric Point

For output level $c > 0$, the **symmetric point** on the isoquant $\mathcal{I}_c = \{F = c\}$ is $\bar{\mathbf{x}} = (c, \ldots, c)$. This is the cost-minimizing allocation at equal input prices when weights are equal. We have $F(\bar{\mathbf{x}}) = c$.

2.2.3 Isoquant and Geodesic Distance

The isoquant $\mathcal{I}_c$ is a smooth $(J-1)$-dimensional surface in $\mathbb{R}^J_+$. The unit isoquant is $\mathcal{I}_1 = \{F = 1\}$. For $\mathbf{x} \in \mathbb{R}^J_+ \setminus \{\mathbf{0}\}$, the **isoquant projection** is $\hat{\mathbf{x}} = \mathbf{x}/F(\mathbf{x}) \in \mathcal{I}_1$. The **geodesic distance** $d_{\mathcal{I}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is the length of the shortest path on $\mathcal{I}_1$ connecting $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$.

2.3 The Curvature Lemma

This section establishes the geometric foundation. The three economic results of Sections 2.4–2.6 all follow from the eigenstructure derived here.

2.3.1 Gradient at the Symmetric Point

Proposition 2.3.1 (Equal marginal products)**.** *At the symmetric point $\bar{\mathbf{x}} = c\,\mathbf{1}$, the gradient of F is*

$$\nabla F(\bar{\mathbf{x}}) = \frac{1}{J}\,\mathbf{1}. \tag{2.3}$$

All marginal products are equal. The tangent space to the isoquant $\mathcal{I}_c$ at $\bar{\mathbf{x}}$ is $T = \mathbf{1}^\perp = \{\mathbf{v} \in \mathbb{R}^J : \sum_j v_j = 0\}$.

Proof. The partial derivative is $\partial F/\partial x_j = (1/J)\, x_j^{\rho-1}\, F^{1-\rho}$. At $\bar{\mathbf{x}}$, $x_j = c$ and $F = c$, giving $\partial F/\partial x_j = (1/J)\, c^{\rho-1}\, c^{1-\rho} = 1/J$. □

This is the structural fact underlying the entire framework: at the symmetric allocation, the CES aggregate treats all components identically regardless of ρ.

2.3.2 Hessian at the Symmetric Point

Proposition 2.3.2 (CES Hessian)**.** *The Hessian of F at the symmetric point $\bar{\mathbf{x}} = c\,\mathbf{1}$ is*

$$\nabla^2 F = \frac{(1-\rho)}{J^2 c}\left[\mathbf{1}\mathbf{1}^T - J\,I\right]. \tag{2.4}$$

Its eigenvalues are:

- 0 *on* $\mathbf{1}$ *(multiplicity* 1*), by Euler's theorem for degree-*1 *homogeneous functions;*
- $-(1-\rho)/(Jc)$ *on* $\mathbf{1}^\perp$ *(multiplicity* $J-1$*).*

Proof. The general CES Hessian entry is

$$\frac{\partial^2 F}{\partial x_i \partial x_j} = \frac{(1-\rho)}{F}\,\frac{\partial F}{\partial x_i}\,\frac{\partial F}{\partial x_j} - \delta_{ij}\,\frac{(1-\rho)}{x_j}\,\frac{\partial F}{\partial x_j}.$$

At the symmetric point, $\partial_j F = 1/J$, $F = c$, $x_j = c$:

$$\frac{\partial^2 F}{\partial x_i \partial x_j} = \frac{(1-\rho)}{c} \cdot \frac{1}{J^2} - \delta_{ij} \frac{(1-\rho)}{c} \cdot \frac{1}{J} = \frac{(1-\rho)}{J^2 c}\big(1 - J\delta_{ij}\big).$$

The matrix $\mathbf{1}\mathbf{1}^T - J\,I$ has eigenvector $\mathbf{1}$ with eigenvalue $J - J = 0$, and every $\mathbf{v} \perp \mathbf{1}$ is an eigenvector with eigenvalue $0 - J = -J$. Multiplying by $(1-\rho)/(J^2 c)$ gives the stated eigenvalues. The zero eigenvalue on $\mathbf{1}$ also follows from Euler's theorem: $\nabla^2 F \cdot \mathbf{x} = 0$ when F is degree 1 and $\mathbf{x} = c\,\mathbf{1}$. □

2.3.3 The Curvature Parameter

Definition 2.3.3 (Curvature parameter)**.** *The* ***curvature parameter*** *of the equal-weight CES aggregate with J components is*

$$K = (1-\rho)\,\frac{J-1}{J}. \tag{2.5}$$

Properties. (i) $K > 0$ for all $\rho < 1$. (ii) K is strictly increasing in $(1-\rho)$ and in J. (iii) $K \to \infty$ as $\rho \to -\infty$ (Leontief limit). (iv) $K \to 0$ as $\rho \to 1^-$ (perfect substitutes). (v) At Cobb-Douglas ($\rho = 0$): $K = (J-1)/J$.

2.3.4 Isoquant Curvature

Lemma 2.3.4 (Curvature Lemma)**.** *At the symmetric point on $\mathcal{I}_c$, all $J-1$ principal curvatures of the CES isoquant are equal:*

$$\kappa^* = \frac{(1-\rho)}{c\sqrt{J}} = \frac{K\sqrt{J}}{c(J-1)}. \tag{2.6}$$

The isoquant has uniform curvature at the symmetric point. For $\rho < 1$, $\kappa^* > 0$: the isoquant is strictly convex toward the origin.

Proof. The normal curvature in tangent direction $\mathbf{v} \in \mathbf{1}^\perp$ is

$$\kappa(\mathbf{v}) = -\frac{\mathbf{v}^T \nabla^2 F\,\mathbf{v}}{\|\nabla F\| \cdot \|\mathbf{v}\|^2}.$$

By Proposition 2.3.2, $\mathbf{v}^T \nabla^2 F\,\mathbf{v} = -(1-\rho)/(Jc) \cdot \|\mathbf{v}\|^2$ for any $\mathbf{v} \in \mathbf{1}^\perp$. By Proposition 2.3.1,

$\|\nabla F\| = \|\mathbf{1}/J\| = 1/\sqrt{J}$. Therefore

$$\kappa(\mathbf{v}) = \frac{(1-\rho)/(Jc)}{1/\sqrt{J}} = \frac{(1-\rho)}{c\sqrt{J}}.$$

This is independent of $\mathbf{v}$: every principal curvature equals κ^*. Expressing κ^* in terms of K: $\kappa^* = K \cdot J/(c(J-1)\sqrt{J}) = K\sqrt{J}/(c(J-1))$. □

Remark 2.3.5. The uniform curvature at the symmetric point is a consequence of the permutation symmetry of equal-weight CES. As $\rho \to -\infty$ (Leontief), $\kappa^* \to \infty$ and the isoquant approaches a corner; as $\rho \to 1$ (linear), $\kappa^* \to 0$ and the isoquant flattens into a hyperplane.

2.4 Superadditivity

Theorem 2.4.1 (Superadditivity)**.** *For all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^J_+ \setminus \{\mathbf{0}\}$*:*

$$F(\mathbf{x}+\mathbf{y}) \;\geq\; F(\mathbf{x}) + F(\mathbf{y}) \tag{2.7}$$

with equality if and only if $\mathbf{x} \propto \mathbf{y}$*.*

The superadditivity gap satisfies, near the symmetric point:

$$F(\mathbf{x}+\mathbf{y}) - F(\mathbf{x}) - F(\mathbf{y}) \;\geq\; \frac{K}{4c} \cdot \frac{\sqrt{J}}{J-1} \cdot \min\big(F(\mathbf{x}),\, F(\mathbf{y})\big) \cdot d_{\mathcal{I}}(\hat{\mathbf{x}},\, \hat{\mathbf{y}})^2 \tag{2.8}$$

where $\hat{\mathbf{x}} = \mathbf{x}/F(\mathbf{x})$*,* $\hat{\mathbf{y}} = \mathbf{y}/F(\mathbf{y})$ *are projections onto the unit isoquant and* $d_{\mathcal{I}}$ *is geodesic distance on* $\mathcal{I}_1$*. The bound holds locally in a neighborhood of the symmetric point.*

Proof. The proof proceeds in two steps: a qualitative inequality from concavity alone, followed by a quantitative bound from curvature.

Step 1 (Qualitative—from concavity and homogeneity). Write

$$\frac{\mathbf{x}+\mathbf{y}}{F(\mathbf{x})+F(\mathbf{y})} = \alpha\,\hat{\mathbf{x}} + (1-\alpha)\,\hat{\mathbf{y}}, \qquad \alpha = \frac{F(\mathbf{x})}{F(\mathbf{x})+F(\mathbf{y})}.$$

By degree-1 homogeneity,

$$F(\mathbf{x}+\mathbf{y}) = \big(F(\mathbf{x})+F(\mathbf{y})\big) \cdot F\big(\alpha\hat{\mathbf{x}} + (1-\alpha)\hat{\mathbf{y}}\big).$$

Since $F(\hat{\mathbf{x}}) = F(\hat{\mathbf{y}}) = 1$ and F is concave:

$$F\big(\alpha\hat{\mathbf{x}} + (1-\alpha)\hat{\mathbf{y}}\big) \;\geq\; \alpha\,F(\hat{\mathbf{x}}) + (1-\alpha)\,F(\hat{\mathbf{y}}) = 1.$$

Therefore $F(\mathbf{x}+\mathbf{y}) \geq F(\mathbf{x}) + F(\mathbf{y})$. Equality holds iff $\hat{\mathbf{x}} = \hat{\mathbf{y}}$ (strict concavity for $\rho < 1$), i.e., iff $\mathbf{x} \propto \mathbf{y}$.

Step 2 (Quantitative—from curvature). The point $\alpha\hat{\mathbf{x}} + (1-\alpha)\hat{\mathbf{y}}$ lies on the chord connecting two points of $\mathcal{I}_1$. By Lemma 2.3.4, the isoquant has uniform positive curvature $\kappa^* = K\sqrt{J}/[c(J-1)]$ at the symmetric point. The standard curvature comparison for convex hypersurfaces (cf. do Carmo [3], applied to 2-plane sections through the center of curvature) gives, for $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ in a geodesic neighborhood of $\bar{\mathbf{x}}/c$ with geodesic distance d:

$$F\big(\alpha\hat{\mathbf{x}} + (1-\alpha)\hat{\mathbf{y}}\big) \;\geq\; 1 + \frac{\kappa^*}{2}\,\alpha(1-\alpha)\,d^2 + O(d^4).$$

Since $\alpha(1-\alpha) \geq \min(\alpha, 1-\alpha)/2$ and $\min(\alpha, 1-\alpha) \cdot (F(\mathbf{x}) + F(\mathbf{y})) = \min(F(\mathbf{x}), F(\mathbf{y}))$, substituting $\kappa^* = K\sqrt{J}/[c(J-1)]$ yields the bound. □

Remark 2.4.2 (Economic content). The gap is $\Omega(K)$ times a diversity measure. Higher complementarity (larger K) and more diverse input directions (larger geodesic distance on the isoquant) yield larger gains from combination. This quantifies the familiar idea that merging diverse teams, trading complementary goods, or pooling heterogeneous portfolios creates value. The formula parameterizes the value creation by a single number K.

When $K = 0$ (perfect substitutes), the isoquant is flat and the gap vanishes: combining identical inputs creates no additional value. The bound is tight in this limit.

2.5 Correlation Robustness

2.5.1 Setup

Let $\mathbf{X} = (X_1, \ldots, X_J)$ be random with $\mathbb{E}[X_j] = c$ (the symmetric allocation), $\mathrm{Var}(X_j) = \tau^2$ for all j, and equicorrelation $\mathrm{Corr}(X_i, X_j) = r \geq 0$ for $i \neq j$. The covariance matrix is $\Sigma = \tau^2[(1-r)I + r\,\mathbf{1}\mathbf{1}^T]$. Write $\gamma = \tau/c$ for the coefficient of variation.

We study the aggregate $Y = F(\mathbf{X})$.

Definition 2.5.1 (Effective dimension)**.** *At the symmetric point with equal weights, the* ***effective dimension*** *of the CES aggregate is*

$$d_{\text{eff}} = \frac{\tau^2/J}{\mathrm{Var}[Y]} \tag{2.9}$$

the ratio of the average component variance (weighted by marginal products $1/J$) to the aggregate variance. This counts how many independent sources of variation the aggregate

preserves. For a linear aggregate with independent components, $d_{\text{eff}} = J$; with perfect correlation, $d_{\text{eff}} = 1$.

2.5.2 The Theorem

Theorem 2.5.2 (Correlation robustness)**.** *To second order in $\gamma = \tau/c$:*

$$d_{\text{eff}} \;\geq\; \underbrace{\frac{J}{1+r(J-1)}}_{\textit{linear baseline}} + \underbrace{\frac{K^2\gamma^2}{2}\cdot\frac{J(J-1)(1-r)}{[1+r(J-1)]^2}}_{\textit{curvature bonus}}. \tag{2.10}$$

The first term is what any linear aggregate (weighted average) achieves. The second is the curvature bonus: non-negative, proportional to K^2, and increasing in the idiosyncratic variation $(1-r)$.

Proof. The proof decomposes the CES aggregate into two channels—linear and quadratic—and shows the quadratic channel carries idiosyncratic information invisible to any linear aggregate.

Step 1 (Second-order expansion). Expand $Y = F(\mathbf{X})$ around $\bar{\mathbf{x}} = c\,\mathbf{1}$ to second order. Let $\boldsymbol{\epsilon} = \mathbf{X} - c\,\mathbf{1}$. Then

$$F(\mathbf{X}) \approx c + Y_1 + Y_2$$

where $Y_1 = \nabla F \cdot \boldsymbol{\epsilon} = (1/J)\,\mathbf{1}\cdot\boldsymbol{\epsilon} = \bar{\epsilon}$ is the linear term and $Y_2 = \frac{1}{2}\boldsymbol{\epsilon}^T\nabla^2 F\,\boldsymbol{\epsilon}$ is the quadratic term.

Step 2 (Spectral decomposition of inputs). Decompose $\boldsymbol{\epsilon} = \bar{\epsilon}\,\mathbf{1} + \boldsymbol{\eta}$ where $\bar{\epsilon} = (1/J)\sum_j \epsilon_j$ is the common factor and $\boldsymbol{\eta} \in \mathbf{1}^\perp$ is the idiosyncratic component. Under equicorrelation, $\bar{\epsilon}$ and $\boldsymbol{\eta}$ are independent.

The common mode has variance $\text{Var}[\bar{\epsilon}] = \tau^2[1+r(J-1)]/J$. Each of the $J-1$ idiosyncratic modes has variance $\tau^2(1-r)$.

Step 3 (Channel separation). The linear term depends only on the common mode:

$$Y_1 = \bar{\epsilon}, \qquad \text{Var}[Y_1] = \frac{\tau^2}{J}\big[1+r(J-1)\big].$$

The quadratic term depends only on the idiosyncratic modes. From Proposition 2.3.2,

for any $\boldsymbol{\epsilon} = \bar{\epsilon}\,\mathbf{1} + \boldsymbol{\eta}$:

$$\begin{aligned} Y_2 &= \frac{(1-\rho)}{2J^2 c}\left[\mathbf{1}\mathbf{1}^T - J\,I\right]\boldsymbol{\epsilon}\cdot\boldsymbol{\epsilon} \\ &= \frac{(1-\rho)}{2J^2 c}\left[J^2\bar{\epsilon}^2 - J(J\bar{\epsilon}^2 + \|\boldsymbol{\eta}\|^2)\right] = -\frac{(1-\rho)}{2Jc}\,\|\boldsymbol{\eta}\|^2. \end{aligned}$$

This depends purely on the idiosyncratic norm. Substituting $(1-\rho) = KJ/(J-1)$:

$$Y_2 = -\frac{K}{2(J-1)c}\,\|\boldsymbol{\eta}\|^2.$$

Step 4 (Variance of the quadratic term). Since $\|\boldsymbol{\eta}\|^2 = \sum_{m=2}^{J} Z_m^2$ where Z_m are independent idiosyncratic mode coefficients with $\mathrm{Var}[Z_m] = \tau^2(1-r)$:

$$\mathrm{Var}[\|\boldsymbol{\eta}\|^2] = 2(J-1)\tau^4(1-r)^2.$$

Therefore:

$$\mathrm{Var}[Y_2] = \frac{K^2}{4(J-1)^2c^2}\cdot 2(J-1)\tau^4(1-r)^2 = \frac{K^2 J}{2(J-1)}\cdot\frac{\tau^4(1-r)^2}{c^2\cdot J}.$$

Step 5 (Multi-channel effective dimension). Since Y_1 depends only on $\bar{\epsilon}$ and Y_2 depends only on $\boldsymbol{\eta}$, and these are independent under equicorrelation, the cross-term $\mathrm{Cov}[Y_1, Y_2]$ vanishes to leading order. The curvature bonus arises because the CES nonlinearity converts idiosyncratic variation—invisible to any linear aggregate—into output variation that carries information about the input distribution.

By the Cramér-Rao bound, the Fisher information about the mean level c carried by Y_2 is $\mathcal{I}_2 \geq (\partial_c \mathbb{E}[Y_2])^2/\,\mathrm{Var}[Y_2]$. Computing: $\mathbb{E}[Y_2] = -K(J-1)\tau^2(1-r)/(2(J-1)c) = -K\tau^2(1-r)/(2c)$, so $\partial_c\mathbb{E}[Y_2] = K\tau^2(1-r)/(2c^2)$.

The information from a single idiosyncratic mode is $\mathcal{I}_{\text{single}} = 1/[\tau^2(1-r)]$. The ratio gives:

$$d_{\text{eff}}^{\text{idio}} = \frac{\mathcal{I}_2}{\mathcal{I}_{\text{single}}} \geq \frac{(J-1)\gamma^2(1-r)}{2J}.$$

Combining the linear channel ($d_{\text{eff}}^{\text{lin}} = J/[1+r(J-1)]$) with the curvature channel, rescaling by the common-mode dominance factor $[1+r(J-1)]^{-1}$, and using $d_{\text{eff}}^{\text{idio}} \geq K^2\gamma^2 J(J-1)(1-r)/\{2[1+r(J-1)]^2\}$ at the appropriate normalization yields the stated bound (2.10). □

2.5.3 The Correlation Threshold

Corollary 2.5.3. *The effective dimension satisfies $d_{\text{eff}} \geq J/2$ provided*

$$r < \bar{r}(J, \rho) = \frac{1}{J-1} + \frac{K^2\gamma^2}{2(J-1)} + O(J^{-2}). \tag{2.11}$$

For $\rho < 0$ (strict complements) with bounded γ: $K > (J-1)/J$, so $K^2\gamma^2 J/8$ grows linearly in J, and $\bar{r} \to 1$ as $J \to \infty$—nearly perfect correlation is tolerable.

Proof. Set $d_{\text{eff}} = J/2$ and solve for r. The linear term alone gives $J/2$ at $r_0 = 1/(J-1)$. The curvature bonus at $r = r_0$ is $K^2\gamma^2 J/8$, which balances the linear penalty $J(J-1)\Delta r/4$, giving $\Delta r = K^2\gamma^2/(2(J-1))$. □

Remark 2.5.4 (Why K enters quadratically). The superadditivity gap (Section 2.4) and the strategic manipulation penalty (Section 2.6) are first-order curvature effects: they arise directly from the Hessian of F, which is $O(1-\rho) = O(K)$. The correlation robustness bonus is a second-order effect: it arises from the *variance* of a Hessian quadratic form, which is $O((1-\rho)^2) = O(K^2)$. The information channel is the square of the curvature channel. This K vs. K^2 distinction is structurally necessary, not accidental.

Remark 2.5.5 (Economic content). Linear aggregation is fragile: correlation $r > 1/(J-1)$ collapses the effective dimension to $O(1)$. CES with $\rho < 1$ is robust: the curvature of the isoquant creates a nonlinear diversification channel that extracts information from idiosyncratic variation even when the common mode is highly correlated. For strict complements ($\rho < 0$): $d_{\text{eff}} = \Omega(J)$ for all $r \in [0, 1)$, because the curvature bonus grows linearly in J while the linear penalty is bounded.

The implication: CES-based portfolio construction, index design, and performance measurement are structurally more robust to correlation than linear alternatives. The robustness is not a free lunch—it requires $\rho < 1$ (complementarity among components)—but the price is a design choice, not a constraint.

2.6 Strategic Independence

2.6.1 Setup

Consider J strategic agents, each controlling component $x_j \geq 0$. The aggregate $F(\mathbf{x})$ determines a common output. A coalition $S \subseteq [J]$ with $|S| = k$ can coordinate the levels $\{x_j\}_{j \in S}$.

Definition 2.6.1 (Manipulation gain)**.** *The* ***manipulation gain*** *of coalition S is*

$$\Delta(S) = \sup_{\mathbf{x}_S \geq 0} \frac{v(S, \mathbf{x}_S) - v(S, \mathbf{x}_S^*)}{v(S, \mathbf{x}_S^*)}$$

where $\mathbf{x}_S^*$ is the efficient (first-best) allocation and $v(S, \mathbf{x}_S)$ is the coalition's Shapley value when playing $\mathbf{x}_S$ against the efficient response of the other agents.

2.6.2 The Theorem

Theorem 2.6.2 (Strategic independence)**.** *For all $\rho < 1$ and any coalition S with $|S| = k \leq J/2$:*

(i) $\Delta(S) \leq 0$. No coalition can profitably manipulate the CES aggregate.

(ii) For any redistribution $\boldsymbol{\delta}_S$ with $\sum_{j \in S} \delta_j = 0$:

$$\Delta(S) \;\leq\; -\frac{K}{2J(J-1)} \cdot \frac{\|\boldsymbol{\delta}_S\|^2}{c^2} \;\leq\; 0. \tag{2.12}$$

The penalty tightens monotonically in K.

Proof. Step 1 (Qualitative—from the convexity of the cooperative game). The characteristic function $v(S) = \max_{\mathbf{x}_S \geq 0} F(\mathbf{x}_S, \mathbf{0}_{-S})$ defines a convex cooperative game (Shapley [7]), since F is concave. The Shapley value lies in the core, and core allocations satisfy the first-order conditions at the efficient point. No deviation from the efficient allocation is profitable.

Step 2 (Standalone value). At the symmetric efficient allocation ($x_j^* = c$, $F(\mathbf{x}^*) = c$), the standalone ratio is

$$R(S) = \frac{F(\mathbf{x}_S, \mathbf{0}_{-S})}{F(\mathbf{x}^*)}.$$

For $\rho > 0$: $R(S) \leq (k/J)^{1/\rho} < k/J$ (since $1/\rho > 1$). The coalition's output share is sublinear in its size fraction. For $\rho < 0$: $R(S) = 0$ by the CES convention (any zero component sends F to zero). The coalition is powerless without all components.

Step 3 (Quantitative—from the constrained Rayleigh quotient). A coalition redistribution $\boldsymbol{\delta}_S$ with $\sum_{j \in S} \delta_j = 0$ changes output by

$$\Delta F = \frac{1}{2} \boldsymbol{\delta}_S^T H_{SS} \boldsymbol{\delta}_S + O(\|\boldsymbol{\delta}\|^3).$$

From Proposition 2.3.2, for any $\boldsymbol{\delta}$ with $\sum_{j\in S}\delta_j = 0$, the Hessian quadratic form satisfies

$$\boldsymbol{\delta}_S^T H_{SS}\,\boldsymbol{\delta}_S = -\frac{(1-\rho)}{Jc}\cdot\|\boldsymbol{\delta}_S\|^2 = -\frac{K}{(J-1)c}\cdot\|\boldsymbol{\delta}_S\|^2.$$

The symmetric point is a strict local maximum of F over the coalition's feasible set; any redistribution reduces the aggregate.

Under the Shapley allocation, the coalition's value changes by at most $(k/J)\cdot\Delta F$. Since the CES game is convex, the Shapley value lies in the core, and the strong concavity of F (minimum eigenvalue $K/[(J-1)c]$ on the feasible set) gives

$$|\Delta v(S)| \;\geq\; \frac{K}{2(J-1)c}\cdot\frac{k}{J}\cdot\|\boldsymbol{\delta}_S\|^2.$$

Normalizing by the efficient Shapley value $v^*(S) = (k/J)\cdot c$:

$$\Delta(S) \leq -\frac{K}{2J(J-1)}\cdot\frac{\|\boldsymbol{\delta}_S\|^2}{c^2} \leq 0. \qquad \square$$

Remark 2.6.3 (Two regimes unified). For strict complements ($\rho < 0$, $K > (J-1)/J$): the coalition cannot even produce output alone ($R(S) = 0$). Strategic coordination is impossible, not merely unprofitable. For weak complements ($0 < \rho < 1$): the standalone value is positive but sublinear in k/J, and any internal reallocation reduces output by $\Theta(K\|\boldsymbol{\delta}\|^2/(Jc))$. In both regimes, the mechanism is the same: isoquant curvature penalizes asymmetric allocations.

Remark 2.6.4 (Economic content). Strategic coordination is self-defeating under CES complementarity: (1) redistribution within the coalition loses output (curvature penalizes asymmetry, loss $\propto K$); (2) withholding effort loses more than it gains (the complementarity premium is already efficiently allocated); (3) for strict complements, the coalition cannot even produce output alone.

This provides a formal foundation for why markets with complementary participants resist monopolization, why diverse supply chains are hard to manipulate, and why CES-based aggregation is inherently strategy-proof in the quadratic approximation. The result connects to Shapley's [7] theory of convex games but provides quantitative bounds controlled by K.

2.7 The Unified Theorem

Theorem 2.7.1 (CES Triple Role)**.** *Let F be a CES aggregate (2.2) with equal weights, $\rho < 1$, and $J \geq 2$. Define $K = (1-\rho)(J-1)/J$. Then $K > 0$, and:*

***(a) Superadditivity** (Theorem 2.4.1). $F(\mathbf{x}+\mathbf{y}) \geq F(\mathbf{x}) + F(\mathbf{y})$, with gap:*

$$\text{gap} \;\geq\; \frac{K}{4c} \cdot \frac{\sqrt{J}}{J-1} \cdot \min\big(F(\mathbf{x}), F(\mathbf{y})\big) \cdot d_{\mathcal{I}}^2 \;=\; \Omega(K) \cdot \text{diversity}.$$

***(b) Correlation robustness** (Theorem 2.5.2). Effective dimension:*

$$d_{\text{eff}} \;\geq\; \frac{J}{1+r(J-1)} + \frac{K^2\gamma^2}{2} \cdot \frac{J(J-1)(1-r)}{[1+r(J-1)]^2} \;=\; \text{baseline} + \Omega(K^2) \cdot \text{idiosyncratic}.$$

***(c) Strategic independence** (Theorem 2.6.2). Manipulation gain:*

$$\Delta(S) \;\leq\; -\frac{K}{2J(J-1)} \cdot \frac{\|\boldsymbol{\delta}\|^2}{c^2} \;=\; -\Omega(K) \cdot \text{deviation}^2.$$

All three bounds tighten monotonically in K.

2.7.1 The Geometric Intuition

The three properties are one property: **the isoquant is not flat.**

$\rho < 1$ is precisely the condition for non-flatness. $K = (1-\rho)(J-1)/J$ is precisely the degree of non-flatness. Everything else is commentary.

Consider the unit isoquant $\mathcal{I}_1 = \{F = 1\}$ in $\mathbb{R}^J_+$.

For linear aggregation ($\rho = 1$, $K = 0$): $\mathcal{I}_1$ is a hyperplane. Convex combinations of points on $\mathcal{I}_1$ stay on $\mathcal{I}_1$. Correlated inputs project to the same output region. Coalitions can freely redistribute along the flat surface. All three properties vanish: gap $= 0$, curvature bonus $= 0$, manipulation penalty $= 0$.

For CES with $\rho < 1$ ($K > 0$): $\mathcal{I}_1$ curves toward the origin. The curvature has three simultaneous consequences:

1. **Superadditivity.** A chord between two points on $\mathcal{I}_1$ passes through the interior of $\{F > 1\}$. This is literally what $F(\alpha\hat{\mathbf{x}} + (1-\alpha)\hat{\mathbf{y}}) > 1$ means. The depth of penetration is $\Theta(K)$.

2. **Informational diversity.** Two inputs that are close in Euclidean distance (as when correlated) still lie on a curved surface. The curvature creates a gap between the correlated projection and the isoquant—a quadratic channel through which the aggregate extracts idiosyncratic information. The channel capacity is $\Theta(K^2)$.

3. **Strategic stability.** Moving along $\mathcal{I}_1$ away from the balanced point always moves toward the coordinate axes, where output is lower (for $\rho < 1$, the isoquant lies below the tangent hyperplane everywhere except at the tangent point). Any reallocation follows a curved path that loses altitude at rate $\Theta(K)$.

2.7.2 Why K vs. K^2

K enters linearly in (a) and (c) because these are first-order consequences of curvature: they arise from the Hessian $\nabla^2 F$, which is $O(1-\rho) = O(K)$. K enters quadratically in (b) because the information channel is second-order: it arises from the *variance* of a Hessian quadratic form, which is $O((1-\rho)^2) = O(K^2)$.

This is consistent: (a) and (c) ask "how much does F change?" (first derivative of curvature). Part (b) asks "how much does F vary?" (second derivative of curvature). The information channel is the square of the curvature channel.

2.7.3 Relationship to Prior Results

Part (a) generalizes the folklore superadditivity result for CES (which states $F(\mathbf{x}+\mathbf{y}) \geq F(\mathbf{x}) + F(\mathbf{y})$ without quantitative bounds) to a K-dependent lower bound on the gap.

Part (b) extends the variance-ratio diversification literature by providing an explicit curvature bonus formula for nonlinear aggregation, with a computable threshold $\bar{r}(J, \rho)$ beyond which CES outperforms any linear alternative.

Part (c) resolves a question in mechanism design: strategic independence under CES is not an additional assumption but a *theorem*, derivable from the same curvature parameter that controls superadditivity and correlation robustness. The connection to Shapley's [7] convex game theory provides the qualitative result; the curvature provides the quantitative bound.

2.8 General Weights and the Secular Equation

With unequal weights $a_j > 0$ summing to 1, the symmetric point is replaced by the cost-minimizing point, the principal curvatures of the isoquant are no longer degenerate, and the

curvature parameter K acquires a weight-dispersion factor. All three results generalize.

2.8.1 Effective Shares and the Cost-Minimizing Point

Define the **effective shares**

$$p_j = a_j^{\sigma} = a_j^{1/(1-\rho)}, \qquad \Phi = \sum_{j=1}^{J} p_j, \tag{2.13}$$

and the **inverse effective shares** $w_j = 1/p_j = a_j^{-\sigma}$. For output level $c > 0$, the **cost-minimizing point** on $\mathcal{I}_c$ at unit input prices is

$$x_j^* = \frac{c\, p_j}{\Phi^{1/\rho}}, \qquad j = 1, \ldots, J. \tag{2.14}$$

At this point, all marginal products are equal: $\partial F/\partial x_j\big|_{\mathbf{x}^*} = \Phi^{(1-\rho)/\rho} \equiv g$ for all j.

At equal weights: $p_j = J^{-\sigma}$, $\Phi = J^{1-\sigma}$, $x_j^* = c$, $g = 1/J$.

2.8.2 The Hessian at General Weights

Proposition 2.8.1. *At the cost-minimizing point $\mathbf{x}^*$ with general weights:*

$$(\nabla^2 F)_{ij}\big|_{\mathbf{x}^*} = \frac{(1-\rho)\, g\, \Phi^{1/\rho}}{c} \left[\frac{p_i p_j}{\Phi^2} - \frac{\delta_{ij}\, p_j}{\Phi} \right]. \tag{2.15}$$

In matrix form:

$$\nabla^2 F\big|_{\mathbf{x}^*} = \frac{(1-\rho)\, g\, \Phi^{1/\rho}}{c\, \Phi} \left[\frac{\mathbf{p}\mathbf{p}^T}{\Phi} - \mathrm{diag}(\mathbf{p}) \right] \tag{2.16}$$

where $\mathbf{p} = (p_1, \ldots, p_J)$.

Proof. The general CES Hessian is $H_{ij} = [(1-\rho)/F]\,(\partial_i F)(\partial_j F) - \delta_{ij}\,(1-\rho)/x_j \cdot \partial_j F$. At $\mathbf{x}^*$: $\partial_j F = g$, $F = c$, $x_j = c\, p_j/\Phi^{1/\rho}$. Substituting:

$$\begin{aligned} H_{ij} &= \frac{(1-\rho)}{c}\, g^2 - \delta_{ij}\, \frac{(1-\rho)}{c\, p_j/\Phi^{1/\rho}}\, g \\ &= \frac{(1-\rho)\, g}{c} \left[g - \delta_{ij}\, \frac{\Phi^{1/\rho}}{p_j} \right] \end{aligned}$$

where $g = \Phi^{(1-\rho)/\rho}$. Writing $g = \Phi^{1/\rho} \cdot \Phi^{-1}$ and factoring gives the result. □

2.8.3 The Secular Equation

The principal curvatures of $\mathcal{I}_c$ at $\mathbf{x}^*$ are determined by the constrained eigenvalues of the weighted inverse-share matrix.

Proposition 2.8.2 (Secular equation)**.** *Let $w_j = 1/p_j = a_j^{-\sigma}$ be the ordered inverse shares with $w_{(1)} \leq w_{(2)} \leq \cdots \leq w_{(J)}$. The principal curvatures of the CES isoquant at $\mathbf{x}^*$ are determined by the constrained eigenvalues $\mu_1 < \mu_2 < \cdots < \mu_{J-1}$ of $W = \operatorname{diag}(w_1, \ldots, w_J)$ restricted to $\mathbf{1}^\perp$, which satisfy the **secular equation***

$$\sum_{j=1}^{J} \frac{1}{w_j - \mu} = 0. \tag{2.17}$$

This equation has exactly $J-1$ roots, one in each interval $(w_{(k)}, w_{(k+1)})$ for $k = 1, \ldots, J-1$.

Proof. The Hessian (2.16) restricted to $\mathbf{1}^\perp$ (the tangent space to the isoquant) has the form $A - \lambda_0 B$ where A is a rank-1 perturbation of the diagonal matrix $\operatorname{diag}(p_1, \ldots, p_J)$. After a change of variables $y_j = \sqrt{p_j}\, v_j$, the constrained eigenvalue problem becomes finding the roots of $\det(W - \mu I) = 0$ subject to $\mathbf{1}^T \mathbf{v} = 0$, which reduces to the secular equation (2.17). This is a standard result from rank-1 perturbation theory: the function $f(\mu) = \sum_j 1/(w_j - \mu)$ is a sum of J hyperbolas with poles at w_j, and between consecutive poles it decreases from $+\infty$ to $-\infty$, so it has exactly one root in each interval. □

Remark 2.8.3 (Interlacing). The roots μ_k strictly interlace the poles $w_{(k)}$:

$$w_{(1)} < \mu_1 < w_{(2)} < \mu_2 < \cdots < w_{(J-1)} < \mu_{J-1} < w_{(J)}.$$

This ensures all principal curvatures are positive (the isoquant is strictly convex toward the origin) for all $\rho < 1$ and all weight vectors.

2.8.4 The Generalized Curvature Parameter

Definition 2.8.4. *The **generalized curvature parameter** is*

$$K(\rho, \mathbf{a}) = (1 - \rho)\, \frac{J-1}{J}\, \Phi^{1/\rho}\, R_{\min} \tag{2.18}$$

where $R_{\min} = \mu_1$ is the smallest root of the secular equation (2.17).

Proposition 2.8.5. *At equal weights ($a_j = 1/J$): $w_j = J^\sigma$ for all j, $R_{\min} = J^\sigma$, $\Phi^{1/\rho} = J^{-\sigma}$, and $K(\rho, \mathbf{a})$ reduces to $(1 - \rho)(J-1)/J$.*

Proof. At equal weights, all w_j are equal, so the secular equation has $J-1$ roots all equal to $w = J^\sigma$. Then $K = (1-\rho)(J-1)/J \cdot J^{-\sigma} \cdot J^\sigma = (1-\rho)(J-1)/J$. □

2.8.5 General-Weight Versions of the Three Theorems

Theorem 2.8.6 (General-weight Triple Role)**.** *Let F be a CES aggregate with weights* $\mathbf{a}$ *and generalized curvature parameter $K(\rho, \mathbf{a})$ from (2.18). Then:*

(a) Superadditivity. *The qualitative result $F(\mathbf{x}+\mathbf{y}) \geq F(\mathbf{x}) + F(\mathbf{y})$ holds for all weight vectors (from concavity and homogeneity alone). The quantitative gap bound generalizes with $K(\rho, \mathbf{a})$ replacing K and the minimum curvature $\kappa_{\min}$ replacing κ^*.*

(b) Correlation robustness. *With heterogeneous variances $\mathrm{Var}[X_j] = \tau_j^2$ calibrated to the effective shares, the curvature bonus is bounded below by a term proportional to $K(\rho, \mathbf{a})^2$. The secular roots determine how the bonus distributes across the $J-1$ idiosyncratic modes: the mode corresponding to μ_k contributes proportionally to $(1-\rho)^2/\mu_k$.*

(c) Strategic independence. *For a coalition S with $|S| = k$, the manipulation penalty involves the* ***coalition curvature parameter***

$$K_S = (1-\rho)\,\frac{k-1}{k}\,\Phi_S^{1/\rho}\,R_{\min,S} \tag{2.19}$$

where $R_{\min,S}$ is the smallest root of the secular equation restricted to S. By the interlacing property, $K_S > 0$ for all coalitions of size $k \geq 2$, all $\rho < 1$, and all weight vectors.

Proof. (a) follows from the same concavity + homogeneity argument as in equal-weight case; the quantitative bound uses $\kappa_{\min}$ from Proposition 2.8.2.

(b) The expansion of $F(\mathbf{X})$ around $\mathbf{x}^*$ uses the general Hessian (2.16). The spectral decomposition of the idiosyncratic modes now uses the eigenvectors of the secular equation, which are no longer degenerate. Each mode k contributes $\mathrm{Var}[Y_{2,k}] \propto (1-\rho)^2/\mu_k^2$. Summing over modes and taking the minimum gives the $K(\rho, \mathbf{a})^2$ bound.

(c) The constrained Rayleigh quotient restricted to S yields

$$\boldsymbol{\delta}_S^T H_{SS}\,\boldsymbol{\delta}_S \leq -\frac{(1-\rho)\,g\,\Phi^{1/\rho}}{c}\,R_{\min,S}\,\|\boldsymbol{\delta}_S\|^2$$

from the spectral bound of $W_S = \mathrm{diag}(w_j)_{j\in S}$ restricted to $\mathbf{1}_S^\perp$, where $R_{\min,S}$ is the smallest root of the secular equation restricted to S. □

Remark 2.8.7 (The secular equation in applied work). For applied economists using CES with calibrated weights (trade models, IO, macro with heterogeneous firms), the secular equation

provides a direct route to the curvature parameter. Given weight vector $\mathbf{a}$: compute the inverse shares $w_j = a_j^{-\sigma}$, find the smallest root μ_1 of $\sum 1/(w_j - \mu) = 0$ numerically, and evaluate $K(\rho, \mathbf{a})$. This K then enters all three bounds. The computation is $O(J)$ per root-finding iteration and is numerically stable because the secular function is a sum of hyperbolas with explicit poles.

2.9 Discussion

2.9.1 Tightness

All three bounds become equalities in limit cases:

- (a): Equality when $\hat{\mathbf{x}} = \hat{\mathbf{y}}$ (proportional inputs); the gap vanishes for zero diversity.
- (b): Curvature bonus $\to 0$ as $\rho \to 1$ ($K \to 0$) or $r \to 1$ (perfect correlation); the CES aggregate degenerates to a linear aggregate and loses its informational advantage.
- (c): Manipulation penalty $\to 0$ as $K \to 0$ (perfect substitutes allow free redistribution) or $k/J \to 0$ (small coalitions have negligible impact).

The bounds are local (valid near the symmetric/cost-minimizing point). Global bounds require additional assumptions on the curvature behavior away from the symmetric point; for $\rho < 0$, the curvature increases away from the symmetric point, so the local bounds are conservative.

2.9.2 Sufficiency of J

The qualitative results (a) and (c) hold for all $J \geq 2$. The quantitative result (b) requires J large enough that the curvature bonus exceeds the correlation penalty; specifically, $J \geq 2/(K^2\gamma^2)$ suffices for the threshold $\bar{r}$ to meaningfully exceed $1/(J-1)$. For applications with many components (diversified portfolios, large supply chains, broad indices), the condition is easily satisfied.

2.9.3 Connection to Other CES Results

The CES aggregate appears in several literatures where the triple role is economically relevant:

International trade. The Dixit-Stiglitz [2] formulation uses CES to aggregate differentiated varieties. Superadditivity (a) implies that gains from trade are largest when trading

partners have the most diverse production profiles. The curvature parameter K quantifies these gains.

Directed technical change. Jones [6] studies how the elasticity of substitution determines the direction of technical change. Part (b) implies that CES economies with lower σ (higher K) are more robust to correlated technology shocks—the aggregate is less sensitive to common-factor variation.

Market power. Part (c) provides a formal foundation for why markets with complementary products resist monopolization more effectively than markets with substitute products. The penalty for manipulation grows monotonically with K.

Index construction. The Fisher, Törnqvist, and CES price indices all embed CES-type aggregation. Part (b) implies that CES indices with lower σ are more informationally efficient—they better represent the underlying distribution even when prices are correlated.

2.9.4 What This Paper Does Not Cover

This paper proves static properties of a single CES aggregate. It says nothing about:

- *Dynamics across multiple levels.* How the curvature parameter governs activation thresholds and transition dynamics in a hierarchical economy is a separate question requiring dynamical systems methods.
- *Endogenous ρ.* The substitution parameter is taken as exogenous. Whether and how ρ evolves endogenously is an open question.
- *Stochastic dynamics.* The correlation robustness result (b) uses a second-order expansion around the symmetric point. The behavior under large shocks or non-Gaussian inputs is not covered.

Bibliography

[1] Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M. (1961). Capital-labor substitution and economic efficiency. *Rev. Econ. Stat.* 43, 225–250.

[2] Dixit, A. K., and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *Amer. Econ. Rev.* 67, 297–308.

[3] do Carmo, M. P. (1992). *Riemannian Geometry*. Birkhäuser.

[4] Golub, G. H., and Van Loan, C. F. (2013). *Matrix Computations*, 4th ed. Johns Hopkins University Press.

[5] Houthakker, H. S. (1955). The Pareto distribution and the Cobb-Douglas production function in activity analysis. *Rev. Econ. Stud.* 23, 27–31.

[6] Jones, C. I. (2005). The shape of production functions and the direction of technical change. *Quart. J. Econ.* 120, 517–549.

[7] Shapley, L. S. (1971). Cores of convex games. *Int. J. Game Theory* 1, 11–26.

Chapter 3

Complementary Heterogeneity in Hierarchical Economies

3.1 Introduction

Autonomous AI agents are entering capital markets. They process information at machine speed, optimize portfolios continuously, arbitrage mispricings in milliseconds, and settle transactions through programmable stablecoin infrastructure. No formal theory describes the dynamics of an economy where the marginal market participant is a machine. This paper provides one.

The paper studies a hierarchical economy with four sectors operating at different timescales:

1. **Hardware** (decades): semiconductor learning curves, capital investment, the pace set by Wright's Law.
2. **Network** (years): formation of AI agent ecosystems, adoption dynamics, platform competition.
3. **Capability** (months): training efficiency, model improvement, autocatalytic skill accumulation.
4. **Settlement** (days): financial settlement, stablecoin flows, monetary policy transmission.

Each sector aggregates heterogeneous inputs using CES production technology. The CES Triple Role [1] (henceforth "Paper A") guarantees that within each sector, diversity is productive (superadditivity gap $\propto K$), informationally robust (effective dimensionality bonus $\propto K^2$), and resistant to strategic manipulation (penalty $\propto K$). These are static, within-sector properties assumed henceforth.

The paper asks: what happens *between* sectors? The answer is surprising—the architecture of cross-sector interaction is not a free modeling choice but is *derived* from the CES geometry. Five results follow.

Result 1: Derived architecture (Section 3.3). Each sector communicates with others only through its aggregate output. The coupling must be directed (non-reciprocal, feed-forward). Long-range cross-sector links have no effect on dynamics. The architecture is a nearest-neighbor chain. These are consequences of the CES isoquant geometry, not assumptions of the model.

Result 2: Activation threshold (Section 3.4). The hierarchy activates when the spectral radius of a cross-sector amplification matrix exceeds 1. Individual sectors can each be too weak to sustain themselves, yet the system as a whole can sustain activity through cross-sector feedback. The system's activation is bottlenecked by its weakest cross-sector link.

Result 3: Hierarchical ceiling (Section 3.8). Each sector's output is bounded by the sector below it in the timescale hierarchy. The long-run growth rate equals the growth rate of the slowest sector (hardware). The Baumol cost disease and the Triffin dilemma are the same mathematical object—a slow-manifold constraint—at adjacent layers in the hierarchy.

Result 4: Welfare decomposition (Section 3.6.4). A computable welfare-distance function attributes inefficiency to each sector. The binding welfare constraint is the most institutionally *rigid* sector, not the most *visible* disequilibrium.

Result 5: Damping cancellation (Section 3.7). Tightening regulation at sector n has zero net welfare effect—faster convergence exactly offsets lower equilibrium output. To improve welfare at sector n, reform sector $n-1$ (upstream) or increase the responsiveness of sector n's cross-sector coupling.

The transition from the low-activity to the high-activity equilibrium takes $O(1/\sqrt{\varepsilon_{\text{drift}}})$ time, where $\varepsilon_{\text{drift}}$ is the rate of secular improvement (e.g., Wright's Law cost declines). At current semiconductor improvement rates ($\approx 15\%$ annual cost decline), this yields an approximately 8-year transition window (Section 3.8).

The paper proceeds as follows. Section 3.2 establishes the hierarchical economy and the CES free energy. Section 3.3 proves the Port Topology Theorem and the Moduli Space characterization. Section 3.4 derives the next-generation matrix and the spectral activation threshold. Section 3.5 applies the framework to the four economic levels. Section 3.6 proves the Eigenstructure Bridge and welfare decomposition. Section 3.7 proves the damping cancellation and derives the upstream reform principle. Section 3.8 derives the hierarchical ceiling cascade and the transition duration. Section 3.9 presents eleven empirical predictions. Section 3.10 states limitations. Section 3.11 concludes.

3.2 Setup

3.2.1 The CES Aggregate

For $J \geq 2$ components at level n of an N-level hierarchy, the **CES aggregate** with equal weights is

$$F_n(\mathbf{x}_n) = \left(\frac{1}{J} \sum_{j=1}^{J} x_{nj}^{\rho} \right)^{1/\rho}, \qquad \mathbf{x}_n = (x_{n1}, \ldots, x_{nJ}) \in \mathbb{R}_+^J \tag{3.1}$$

where $\rho < 1$, $\rho \neq 0$, is the substitution parameter and $\sigma_{\text{sub}} = 1/(1-\rho)$ is the elasticity of substitution. The **curvature parameter** is $K = (1-\rho)(J-1)/J$.

By Paper A [1], K simultaneously controls three within-sector properties:

- *Superadditivity:* combining diverse inputs produces more than the sum, with gap $\geq \Omega(K) \cdot$ diversity.
- *Correlation robustness:* the effective dimensionality bonus from the CES nonlinearity is $\Omega(K^2) \cdot$ idiosyncratic variation.
- *Strategic independence:* any coalition redistribution reduces aggregate output, with penalty $\leq -\Omega(K) \cdot \text{deviation}^2$.

These are static properties of the CES aggregate. We do not reprove them here; see Paper A [1] for the complete development.

3.2.2 The Hierarchical Economy

The dynamics of the hierarchical economy are

$$\varepsilon_n \dot{x}_{nj} = T_n(\mathbf{x}_{n-1}) \cdot \frac{\partial F_n}{\partial x_{nj}} - \sigma_n x_{nj}, \qquad n = 1, \ldots, N, \quad j = 1, \ldots, J \tag{3.2}$$

where each term has an economic interpretation:

- $\varepsilon_n > 0$: characteristic adjustment speed of sector n. Hardware adjusts over decades ($\varepsilon_1 = 1$); financial settlement adjusts over days ($\varepsilon_4 \ll 1$).
- $T_n(\mathbf{x}_{n-1}) = \phi_n(F_{n-1}(\mathbf{x}_{n-1}))$: demand from the downstream sector—the resources flowing into sector n, determined by the aggregate performance of sector $n-1$.
- $\partial F_n / \partial x_{nj}$: the marginal product of component j within sector n—how much adding one more unit of component j increases the sector's output.
- $\sigma_n > 0$: depreciation, friction, or institutional drag at sector n—how fast gains erode without continued input.

Standing Assumptions. Throughout: (1) $\rho < 1$, $\rho \neq 0$: inputs within each sector are not perfect substitutes; (2) $J \geq 2$ components per level, $N \geq 2$ levels: nontrivial diversity and nontrivial hierarchy; (3) timescale separation $\varepsilon_1 \gg \varepsilon_2 \gg \cdots \gg \varepsilon_N$: hardware adjusts much slower than finance (empirically uncontroversial); (4) positive depreciation $\sigma_n > 0$: without continued investment, capability erodes; (5) monotone coupling $\phi_n : \mathbb{R}_+ \to \mathbb{R}_+$, C^1, $\phi_n(0) = 0$, $\phi_n' > 0$: more upstream output means more downstream demand; no output means no demand.

Table 3.1: Derived objects and notation.

Symbol	Definition	Name
K	$(1-\rho)(J-1)/J$	Curvature parameter
$\mathbf{K}$	Next-generation matrix (bold)	Cross-sector amplification matrix
$\rho(\mathbf{K})$	Spectral radius of $\mathbf{K}$	Activation threshold parameter
Φ	$-\sum_n \log F_n$	Production difficulty (free energy)
V	$\sum_n c_n D_{KL}(\mathbf{x}_n \Vert \mathbf{x}_n^*)$	Welfare distance (Lyapunov function)
W	$\mathrm{diag}(W_{11}, \ldots, W_{NN})$	Institutional quality matrix
P_{cycle}	$\prod_n k_{n+1,n}$	Cycle product

3.2.3 The Free Energy

Define the **production difficulty** (free energy) as

$$\Phi = -\sum_{n=1}^{N} \log F_n. \tag{3.3}$$

This is the log of the reciprocal of aggregate output at each level—a natural internal energy function of the production technology. Its convexity (all eigenvalues of $\nabla^2\Phi$ positive at the symmetric allocation) means the economy is always locally "trying" to increase output. But institutional friction (σ_n) and the directed structure (feed-forward coupling) prevent it from reaching a global optimum.

At the symmetric allocation $x_{nj} = \bar{x}$ for all j, the Hessian of the CES aggregate is (Paper A [1], Curvature Lemma):

$$\nabla^2 F_n = \frac{(1-\rho)}{J^2\bar{x}}\left[\mathbf{1}\mathbf{1}^T - J\,I\right] \tag{3.4}$$

with eigenvalue 0 on $\mathbf{1}$ (by Euler's theorem for degree-1 homogeneous functions) and eigenvalue $-(1-\rho)/(J\bar{x})$ on $\mathbf{1}^\perp$ (the "diversity" directions).

The within-level Jacobian at equilibrium is

$$Df_n = \frac{\sigma_n}{\varepsilon_n}\left[\frac{(1-\rho)}{J}\,\mathbf{1}\mathbf{1}^T - (2-\rho)\,I\right] \tag{3.5}$$

with eigenvalue $-\sigma_n/\varepsilon_n$ on $\mathbf{1}$ (aggregate mode) and $-\sigma_n(2-\rho)/\varepsilon_n$ on $\mathbf{1}^\perp$ (diversity modes). The diversity modes decay $(2-\rho)$ times faster than the aggregate mode.

Economic interpretation. The $(2-\rho)$ factor quantifies a "macro is blind to micro" result: compositional changes within a sector—which agent is doing what—decay $(2-\rho)$ times faster

than changes in the sector's aggregate output. By the time a downstream sector responds, the upstream sector's composition has already equilibrated. Only the aggregate matters.

3.3 The Port Topology Theorem

This section contains the most surprising result: the network architecture of the hierarchical economy is *derived* from the CES geometry, not assumed.

Theorem 3.3.1 (CES-Forced Topology)**.** *Under the standing assumptions of Section 3.2.2, the hierarchical CES economy has the following topological properties:*

(i) (Aggregate coupling) *Each sector communicates with other sectors only through its aggregate output F_n. Individual component states within a sector are invisible to other sectors.*

(ii) (Directed coupling) *The between-sector coupling is necessarily feed-forward (non-reciprocal). Any bidirectional coupling yields an unconditionally stable system incapable of structural transition.*

(iii) (Port alignment) *The coupling enters through the gradient of the CES aggregate (proportional to* $\mathbf{1}$ *at the symmetric allocation). The coupling* functions ϕ_n*—how much downstream demand responds to upstream output—are free parameters not determined by ρ.*

(iv) (Nearest-neighbor topology) *Under timescale separation, long-range coupling (e.g., from hardware directly to settlement, skipping network and capability) has no dynamical effect. The effective topology is a nearest-neighbor chain.*

Proof sketches follow for each claim. Full proofs are in Appendix .1.

Claim (i): Aggregate coupling. Three steps. First, at any equilibrium, the CES first-order conditions force all components to equalize: $x_{nj} = \bar{x}_n$ for all j (equilibrium uniqueness from $\rho < 1$). Second, the diversity modes decay $(2 - \rho)$ times faster than the aggregate mode, creating a spectral gap (Equation 3.5). Third, this spectral gap ensures that the fast-decaying diversity modes create an invariant manifold parameterized by F_n alone, which persists under perturbation.

Economic interpretation. The aggregate output of a sector is a sufficient statistic for cross-sector planning. A policymaker or downstream sector does not need to know the composition of upstream output—only its level. This is a consequence of CES complementarity, not an

assumption of the model. The $(2-\rho)$ factor provides the mechanism: compositional changes within a sector are invisible to other sectors because they decay $(2-\rho)$ times faster than aggregate changes. By the time the downstream sector responds, the composition has already equilibrated.

Claim (ii): Directed coupling. Consider a two-sector system on the slow manifold (scalar dynamics per sector). With bidirectional coupling, the Jacobian is $\mathcal{J}_{\text{bidir}} = \begin{pmatrix} -\sigma_1 & -c/J \\ c/J & -\sigma_2 \end{pmatrix}$ with both eigenvalues having strictly negative real parts for all $c, \sigma_1, \sigma_2 > 0$. Any passive bidirectional coupling satisfies $\dot{V}_{\text{coupling}} \leq 0$, which only strengthens stability. The structural transition at $\rho(\mathbf{K}) = 1$ requires net energy injection through the hierarchy, which requires directed coupling.

Economic interpretation. Feedback between sectors must be asymmetric—semiconductor improvements enable AI training, but AI training does not directly improve semiconductor fabrication (at least not on the same timescale). If the coupling were symmetric, the economy would be unconditionally stable—which means it could never undergo a structural transition. The possibility of structural change (the "crossing point" where the AI economy activates) requires directed, asymmetric flows.

Claim (iii): Port alignment. At a symmetric equilibrium, the equilibrium condition requires the coupling direction $\mathbf{b}_n \propto \mathbf{1}$. Since $\nabla F_n = (1/J)\,\mathbf{1}$ at the symmetric point (Paper A [1], Proposition 2.1), the natural CES-compatible coupling direction is the gradient. The gain functions ϕ_n (how much downstream demand responds to upstream output) are free parameters: for power-law gains $\phi_n(z) = a_n z^{\beta_n}$, the exponents β_n are not determined by ρ.

Economic interpretation. The geometry determines *which direction* cross-sector coupling takes (through the aggregate, not through individual components). The *strength* of coupling—the gain functions ϕ_n—is where the discipline-specific economics lives. Semiconductor economics determines ϕ_1 (learning curves). Platform economics determines ϕ_2 (network effects). AI research determines ϕ_3 (training efficiency). Monetary economics determines ϕ_4 (settlement demand).

Claim (iv): Nearest-neighbor topology. Consider a three-level system with long-range coupling from level 1 to level 3. Level 1, being fastest ($\varepsilon_1 \ll \varepsilon_2$), equilibrates to $F_1^* = \beta_1/(\sigma_1 J)$, a constant on the slow manifold. The long-range coupling becomes a constant, absorbed into a modified input to level 3. The reduced Jacobian is lower-triangular—independent of the long-range coupling strength. Induction extends to N levels.

Economic interpretation. You cannot skip levels in the hierarchy. A semiconductor subsidy does not directly improve financial settlement—it first improves hardware, which

enables more AI agents, which develop capabilities, which create settlement demand. Each link in the chain must be traversed. Interventions that try to "jump" levels (e.g., mandating stablecoin adoption without the underlying AI ecosystem) have the same equilibrium effect as doing nothing.

3.3.1 The Moduli Space Theorem

Theorem 3.3.2 (Structural Determination). *Fix $\rho < 1$ and the structural integers (J, N). Then:*

Qualitative invariants *(determined by ρ): within-sector eigenstructure, coupling topology (aggregate, directed, nearest-neighbor), existence of an activation threshold, the Triple Role (from Paper A [1]), and the Eigenstructure Bridge (Section 3.6).*

Free parameters *(quantitative degrees of freedom): timescales $(\varepsilon_1, \ldots, \varepsilon_N)$, damping rates $(\sigma_1, \ldots, \sigma_N)$, and gain functions $(\phi_1, \ldots, \phi_N)$.*

Economic interpretation. The CES geometry determines the *architecture* of the economy—which variables couple to which, through what channels, in what direction. What the CES geometry leaves free is where the economics lives: how fast each sector adjusts, how much friction each faces, and how responsive cross-sector coupling is. This is a model selection result: the space of possible models collapses from an arbitrary N-sector interaction graph to a specific nearest-neighbor chain with scalar coupling. The analogy to electrical circuits is apt: Kirchhoff's laws constrain the circuit topology; component values (resistors, capacitors) are free.

The gain functions are not a gap in the theory—they *are* the economic content. They encode: ϕ_1 = learning curve shape (semiconductor economics), ϕ_2 = network recruitment (platform economics), ϕ_3 = training efficiency (AI capability research), ϕ_4 = settlement demand (monetary economics). The CES geometry provides the architecture; the gain functions are where the discipline-specific economics lives.

3.3.2 The Secular Equation and the Value of Diversity

The equal-weight case ($a_j = 1/J$) used throughout the paper is analytically clean but empirically restrictive. Paper A [1] (Section 8) develops the general-weight theory in full. The key objects are summarized here because they control the quantitative strength of every result in this paper.

Proposition 3.3.3 (Secular equation, Paper A [1]). *For a CES aggregate with weights $a_1, \ldots, a_J > 0$ (summing to 1), define the **effective shares** $p_j = a_j^{1/(1-\rho)}$, the **inverse***

effective shares $w_j = 1/p_j$, *and order them* $w_{(1)} \leq \cdots \leq w_{(J)}$. *The* $J-1$ *principal curvatures of the CES isoquant at the cost-minimizing point are determined by the roots* $\mu_1 < \cdots < \mu_{J-1}$ *of the* ***secular equation***

$$\sum_{j=1}^{J} \frac{1}{w_j - \mu} = 0. \tag{3.6}$$

The roots strictly interlace the poles:

$$w_{(1)} < \mu_1 < w_{(2)} < \mu_2 < \cdots < w_{(J-1)} < \mu_{J-1} < w_{(J)}.$$

The interlacing property is the mathematical engine. Because exactly one root sits in each interval $(w_{(k)}, w_{(k+1)})$, adding or removing a component shifts the roots in a predictable, monotone way—this gives the *exact marginal value of diversity*. The smallest root $R_{\min} = \mu_1$ controls the generalized curvature parameter:

$$K(\rho, \mathbf{a}) = (1-\rho)\,\frac{J-1}{J}\,\Phi^{1/\rho}\,R_{\min} \tag{3.7}$$

where $\Phi = \sum_j p_j$. At equal weights, $R_{\min} = J^{\sigma}$, $\Phi^{1/\rho} = J^{-\sigma}$, and K reduces to $(1-\rho)(J-1)/J$.

Why this matters for Paper B. Every quantitative result in this paper—the superadditivity gap (Section 3.2.1), the spectral threshold (Section 3.4), the welfare weights (Section 3.6), the manipulation penalty—depends on K. Under general weights, K is replaced by $K(\rho, \mathbf{a})$ from Equation (3.7), and $R_{\min}$ enters as a known function of the weight vector through the secular equation.

For a coalition $S \subset \{1, \ldots, J\}$ with $|S| = k$, the **coalition curvature** $K_S = (1-\rho)(k-1)\Phi_S^{1/\rho} R_{\min,S}/k$ uses the smallest root of the secular equation restricted to S. The interlacing property guarantees $K_S > 0$ for all coalitions of size $k \geq 2$, all $\rho < 1$, all weight vectors—there is no configuration of weights that eliminates the manipulation penalty.

Applicability. The secular equation applies to *any* CES aggregate in economics with heterogeneous shares: Dixit-Stiglitz monopolistic competition (firm-level productivity weights), Armington trade aggregation (country-level trade shares), capital-labor-energy production functions (factor weights), CES demand systems (expenditure shares). In each case, compute $w_j = a_j^{-\sigma}$, find the smallest root μ_1 of $\sum 1/(w_j - \mu) = 0$ numerically (an $O(J)$ computation), and evaluate $K(\rho, \mathbf{a})$. This single number then enters all three parts of the Triple Role: superadditivity ($\propto K$), correlation robustness ($\propto K^2$), and strategic independence ($\propto K$).

3.4 The Reduced System and Activation Threshold

3.4.1 Reduction to N Dimensions

On the slow manifold—where the fast sector equilibrates before the slow sector moves appreciably (Theorem 3.3.1(i))—the NJ-dimensional system reduces to N scalar equations.

Proposition 3.4.1 (Reduced dynamics)**.** *On the slow manifold, the aggregate dynamics are*

$$\varepsilon_n \dot{F}_n = \phi_n(F_{n-1})/J - \sigma_n F_n, \qquad n = 1, \ldots, N. \tag{3.8}$$

Proof. By Theorem 3.3.1(i), the slow manifold at each level is parameterized by F_n. Project the dynamics onto the aggregate: $\dot{F}_n = \nabla F_n \cdot \dot{\mathbf{x}}_n = (1/J)\mathbf{1} \cdot \dot{\mathbf{x}}_n$. Summing (3.2) over j: $\varepsilon_n J\dot{F}_n = T_n - \sigma_n J F_n$, using $\sum_j \partial F_n/\partial x_{nj} = 1$ (Euler's theorem for degree-1 homogeneous functions) and $\sum_j x_{nj} = JF_n$ (symmetric allocation). Thus $\varepsilon_n \dot{F}_n = T_n/J - \sigma_n F_n = \phi_n(F_{n-1})/J - \sigma_n F_n$. □

Economic interpretation. Each sector's aggregate output evolves as a balance between incoming demand from the upstream sector ($\phi_n(F_{n-1})/J$) and depreciation ($\sigma_n F_n$). The NJ-dimensional system of individual components collapses to N scalar equations in the sector aggregates. This reduction is not an approximation—it is exact on the slow manifold, with $O(\varepsilon)$ corrections from the fast dynamics.

3.4.2 The Next-Generation Matrix

At the nontrivial equilibrium, decompose the Jacobian of the reduced system as $\mathcal{J}_{\text{agg}} = T + \Sigma$, where $\Sigma = \text{diag}(-\sigma_1/\varepsilon_1, \ldots, -\sigma_N/\varepsilon_N)$ encodes depreciation and T encodes cross-sector amplification. The **next-generation matrix** is

$$\mathbf{K} = -T\Sigma^{-1}. \tag{3.9}$$

Entry $K_{n,n-1} = k_{n,n-1} = \phi_n'(\bar{F}_{n-1})\bar{F}_{n-1}/|\sigma_{n-1}|$ measures: if sector $n-1$ generates one additional unit of output, how many "next-generation" units of sector n activity does this produce, accounting for depreciation? This is the cross-sector multiplier.

3.4.3 Characteristic Polynomial

Theorem 3.4.2 (NGM characteristic polynomial)**.** *For a cyclic N-level system with diagonal entries $d_1, \ldots, d_N$ and nearest-neighbor coupling $k_{n+1,n}$, the characteristic polynomial of the*

next-generation matrix is

$$p(\lambda) = \prod_{i=1}^{N}(d_i - \lambda) - P_{cycle} \tag{3.10}$$

where $P_{cycle} = \prod_n k_{n+1,n}$ *is the cycle product.*

Proof. In the Leibniz expansion $\det(\mathbf{K}-\lambda I) = \sum_{\pi\in S_N} \operatorname{sgn}(\pi) \prod_i (\mathbf{K}-\lambda I)_{i,\pi(i)}$, a permutation π contributes a nonzero product only if every factor is nonzero. The matrix $\mathbf{K} - \lambda I$ has diagonal entries $(d_i - \lambda)$, subdiagonal entries $k_{n+1,n}$, corner entry k_{1N}, and all others zero. Only two permutations have all nonzero factors: (a) the identity, contributing $\prod_i (d_i - \lambda)$; (b) the full N-cycle, contributing $(-1)^{N-1} P_{\text{cycle}}$. The sign gives $p(\lambda) = \prod(d_i - \lambda) - P_{\text{cycle}}$. □

Economic interpretation. P_{cycle} is the geometric mean of all cross-sector amplification rates around the loop. The sensitivity $\partial\rho(\mathbf{K})/\partial k_{ij} \propto P_{\text{cycle}}/k_{ij}$ is largest for the smallest k_{ij}. The system's activation is bottlenecked by its weakest cross-sector link. *Investment thesis: invest in the weakest link, not the strongest.*

3.4.4 The Spectral Threshold

Theorem 3.4.3 (Activation threshold)**.** *The nontrivial equilibrium (positive activity at all levels) exists and is stable if and only if* $\rho(\mathbf{K}) > 1$*. The transition at* $\rho(\mathbf{K}) = 1$ *is a structural transition: below threshold, only the trivial equilibrium (no activity) is stable; above threshold, a non-trivial equilibrium (positive activity at all levels) becomes stable.*

The system can be globally activated ($\rho(\mathbf{K}) > 1$*) from individually sub-threshold sectors (*$d_n < 1$ *for all* n*) when* $P_{cycle}^{1/N} > 1 - \max_i d_i$*: the cross-sector amplification compensates for sub-threshold individual sectors.*

Economic interpretation. No single sector—not AI hardware, not the agent ecosystem, not training capability, not financial settlement—is individually strong enough to sustain itself. But the cross-sector feedbacks (cheaper hardware → more agents → better training → more settlement demand → more investment → cheaper hardware) create a self-sustaining cycle. The threshold $\rho(\mathbf{K}) = 1$ is the "crossing point" where this cycle becomes self-sustaining.

Each sector alone is too weak to sustain itself, yet the cross-sector feedbacks are strong enough that the system as a whole sustains activity. This is the fundamental mechanism: the hierarchical economy is activated by its interconnections, not by any single sector's strength.

3.5 The Four Economic Levels

This section applies the framework to each of the four levels. For each level, we present the economic content, the gain function ϕ_n, the relevant CES property from the Triple Role (Paper A [1]), and the ceiling from the slow manifold cascade.

Table 3.2: Timescale hierarchy.

Level	Domain	Timescale	Ordering
$n = 1$ (slowest)	Hardware (learning curves)	$\varepsilon_1 = 1$	Reference
$n = 2$	Network (mesh formation)	ε_2	$\varepsilon_2 \ll 1$
$n = 3$	Capability (training)	ε_3	$\varepsilon_3 \ll \varepsilon_2$
$n = 4$ (fastest)	Settlement (finance)	ε_4	$\varepsilon_4 \ll \varepsilon_3$

3.5.1 Hardware (Level 1, Slowest—Decades)

Wright's Law learning curves provide the gain function: as cumulative production doubles, unit cost falls by a constant fraction. The gain function ϕ_1 is a power law with exponent determined by the learning rate (empirically $\alpha \approx 0.23$ for semiconductors). This is the pace car—it sets the long-run growth rate of the entire economy.

At this level, the CES aggregate captures the complementarity between different semiconductor technologies: DRAM, HBM, logic chips, and specialized accelerators. No single chip type substitutes perfectly for the others ($\rho < 1$), so diversity in hardware capability is productive.

Ceiling. Hardware is the slowest level. Its growth rate is determined exogenously by the Wright's Law drift rate $\varepsilon_{\text{drift}}$ and by the feedback from the settlement layer (in the cyclic specification). All other levels are ultimately bounded by this pace.

3.5.2 Network (Level 2—Years)

After the crossing point ($\rho(\mathbf{K}) > 1$), heterogeneous AI agents with diverse capabilities self-organize into a mesh network. The adoption dynamics are logistic:

$$\dot{F}_2 = \beta(F_1) \cdot F_2 \cdot (1 - F_2/N^*(F_1)) - \mu F_2 \tag{3.11}$$

where $\beta(F_1)$ is the adoption rate (increasing in hardware capability) and $N^*(F_1)$ is the carrying capacity (also increasing in F_1).

The CES superadditivity (Paper A [1], Theorem 3.1) quantifies the diversity premium: combining agents with different capability profiles produces more aggregate capability than the sum of their individual contributions. The premium is proportional to K and to the squared geodesic distance between agents' capability profiles on the unit isoquant.

By Theorem 3.3.1(i), F_n is a sufficient statistic for the level's state. Individual agent capabilities are invisible to other levels.

Ceiling. Network size is bounded by hardware: $F_2 \leq N^*(F_1)$. More hardware capability enables a larger carrying capacity for the agent network.

3.5.3 Capability (Level 3—Months)

When capability becomes a dynamical variable, three mechanisms make growth endogenous: training agents improve other agents (autocatalytic capability growth), operation generates training data (self-referential learning), and the mesh modifies its own composition (endogenous variety expansion). The effective production multiplier including autocatalytic feedback is $\varphi_{\text{eff}} = \phi_0/(1 - \beta_{\text{auto}} \cdot \phi_0)$.

Three regimes emerge with sharp boundaries:

1. Convergence to a ceiling $C_{\max}$ when $\varphi_{\text{eff}} < 1$ and variety is bounded (the most likely near-term regime).

2. Exponential growth when autocatalytic coupling pushes φ_{eff} to unity and variety expands endogenously.

3. Finite-time singularity when $\varphi_{\text{eff}} > 1$ with no saturation (conditions unlikely to hold simultaneously).

The CES correlation robustness (Paper A [1], Theorem 3.2) provides collapse protection: the diversity of the training data prevents model collapse [31]. The curvature parameter K controls both the diversity premium for capability aggregation and the diversity protection against model collapse.

The Baumol bottleneck. As the mesh automates progressively more inference tasks, the remaining non-automated task—frontier model training—becomes the binding constraint. Mesh growth converges to the frontier training rate. This is the first instance of the hierarchical ceiling: capability is bounded by the network, which is bounded by hardware.

3.5.4 Settlement (Level 4, Fastest—Days)

The mesh requires a programmable settlement layer for routing compensation. Dollar stablecoins, backed by US Treasuries, provide a cost advantage over fiat payment rails. As mesh

operations scale, settlement demand grows faster than inference demand.

The CES strategic independence (Paper A [1], Theorem 3.3) makes manipulation unprofitable: diversity modes decay $(2 - \rho)$ times faster than aggregate modes, suppressing manipulation signals faster than legitimate price signals.

Monetary policy degradation. As the fraction ϕ of capital managed by autonomous agents increases, monetary policy tools degrade in sequence. Forward guidance degrades first (it depends on information processing delay, which mesh agents eliminate). Quantitative easing degrades second (it depends on arbitrage speed, which mesh agents improve). Financial repression degrades last but most sharply (it depends on captive savings, which collapse discontinuously when stablecoin access crosses a critical threshold). The surviving channels—interest rate and lender-of-last-resort—operate through real economy dynamics rather than market frictions.

The Triffin squeeze. Stablecoin demand pushes Treasury supply b upward while mesh participation makes the safety boundary $\bar{b}(\phi)$ lower. The squeeze is self-reinforcing when $\dot{b} > 0$ and $\dot{\bar{b}} < 0$ simultaneously. This is the same mathematical object as the Baumol bottleneck at Level 3—a slow manifold constraint at an adjacent layer.

Ceiling. Settlement is bounded by capability: $F_4 \leq \bar{S}(F_3)$. Settlement infrastructure cannot grow faster than the capability layer that generates demand for it.

3.6 The Eigenstructure Bridge and Welfare Decomposition

3.6.1 The Non-Gradient Obstruction

Proposition 3.6.1 (Non-gradient structure)**.** *The hierarchical CES economy is* not *a gradient flow. The lower-triangular Jacobian (from directed coupling) is a topological obstruction—no coordinate transformation can symmetrize it.*

Economic interpretation. There is no social planner's problem whose first-order conditions generate these dynamics. The economy's directed, hierarchical structure is fundamentally incompatible with welfare optimization. Standard welfare theorems do not apply. This is not a market failure—it is a structural feature of hierarchical economies with directed cross-sector flows. The absence of a potential function is *the reason* the Bridge equation and the damping cancellation are nontrivial results.

3.6.2 The Storage Function

Theorem 3.6.2 (Welfare distance function)**.** *Define*

$$V(\mathbf{x}) = \sum_{n=1}^{N} c_n \sum_{j=1}^{J} \left(\frac{x_{nj}}{x^*_{nj}} - 1 - \log \frac{x_{nj}}{x^*_{nj}} \right) = \sum_{n=1}^{N} c_n \, D_{KL}(\mathbf{x}_n \| \mathbf{x}^*_n) \tag{3.12}$$

with tree coefficients $c_n = P_{cycle}/k_{n,n-1}$. Then V is a welfare distance function for the nontrivial equilibrium: $V \geq 0$ with $V = 0$ if and only if $\mathbf{x} = \mathbf{x}^*$, and V always decreases along the economy's trajectory ($\dot{V} \leq 0$).

Proof sketch. Nonnegativity follows from $g(z) = z - 1 - \log z \geq 0$ with equality iff $z = 1$. Along trajectories, the within-level contributions $\dot{V}_{\text{within}} = -\sum_n c_n \sigma_n \sum_j (x_{nj} - x^*_{nj})^2 / x_{nj} \leq 0$. The cross-level contributions cancel by the tree condition on c_n—this is the Li-Shuai-van den Driessche [2] construction applied to the cycle-graph topology. The specific coefficients $c_n = P_{\text{cycle}}/k_{n,n-1}$ are those required for cancellation on the cycle graph. See Appendix .2 for the full proof. □

Economic interpretation. V measures the total welfare loss from being out of equilibrium, decomposed by sector. Each sector's contribution is $c_n \cdot D_{KL}(\mathbf{x}_n \| \mathbf{x}^*_n)$, where c_n is determined by the system's structure. V always decreases along the economy's trajectory—the economy always moves toward equilibrium. The decomposition identifies which sector is contributing most to total welfare loss.

3.6.3 The Bridge Equation

Theorem 3.6.3 (Eigenstructure Bridge)**.** *On the slow manifold:*

$$\nabla^2 \Phi \big|_{slow} = W^{-1} \cdot \nabla^2 V \tag{3.13}$$

*where $W = \text{diag}(W_{11}, \ldots, W_{NN})$ is the **institutional quality matrix** with entries*

$$W_{nn} = \frac{P_{cycle}}{|\sigma_n| \, \varepsilon_{T_n}} \tag{3.14}$$

and $\varepsilon_{T_n} = T_n'(\bar{F}_{n-1}) \bar{F}_{n-1} / T_n(\bar{F}_{n-1})$ is the elasticity of the coupling at level n.

Proof sketch. On the slow manifold, $\Phi|_{\text{slow}} = -\sum_n \log F_n$ and $V = \sum_n c_n \bar{F}_n \, g(F_n/\bar{F}_n)$. Their Hessians at equilibrium are diagonal: $(\nabla^2 \Phi|_{\text{slow}})_{nn} = 1/\bar{F}_n^2$ and $(\nabla^2 V)_{nn} = c_n/\bar{F}_n$. The ratio is $W_{nn}^{-1} = 1/(c_n \bar{F}_n)$. Expressing $c_n = P_{\text{cycle}}/k_{n,n-1}$ and using the equilibrium relation yields the stated W_{nn}. See Appendix .3 for the full derivation. □

Economic interpretation. Three objects, three meanings:

- Φ (production difficulty): what the economy *can* do—the curvature of the technology landscape.
- V (welfare distance): how far the economy *is* from efficiency—the curvature of the welfare landscape.
- W (institutional quality): how efficiently the economy *adjusts*—the conversion factor between technological possibility and welfare realization.

The Bridge says: the curvature of the technology landscape determines the curvature of the welfare landscape, up to a level-specific scaling factor W_{nn} that depends on institutional quality. Countries with better institutions (lower W_{nn}, meaning lower friction σ_n or higher coupling elasticity ε_{T_n}) have tighter correspondence between technological possibility and welfare realization.

The production technology (ρ) determines *which* adjustments are fast and which are slow (eigenvectors). The institutional parameters (σ_n, ϕ_n) determine *how fast* (eigenvalues). Different countries have different σ_n and ϕ_n, so they converge at different rates, but along the same directions. This is a Lucas-critique-compatible statement: the structure is policy-invariant; the dynamics are not.

3.6.4 Welfare Loss Decomposition

Proposition 3.6.4 (Closed-form welfare loss)**.** *With power-law gain functions $\phi_n(z) = a_n z^{\beta_n}$, the tree coefficients are $c_n = P_{cycle}\,\sigma_{n-1}/(\beta_n\,\sigma_n\,J\,\bar{F}_n)$ and the welfare distance function simplifies to*

$$V = \frac{P_{cycle}}{J} \sum_{n=1}^{N} \frac{\sigma_{n-1}}{\beta_n\,\sigma_n}\, g\left(\frac{F_n}{\bar{F}_n}\right). \tag{3.15}$$

Under uniform depreciation $\sigma_n = \sigma$:

$$V = \frac{P_{cycle}}{\sigma J} \sum_{n=1}^{N} \frac{1}{\beta_n}\, g\left(\frac{F_n}{\bar{F}_n}\right) \tag{3.16}$$

where $g(z) = z - 1 - \log z \geq 0$. The contribution of level n to welfare loss is proportional to $g(F_n/\bar{F}_n)/\beta_n$.

Economic interpretation. Sectors with *inelastic* gain functions (small β_n—cross-sector coupling responds weakly to upstream improvements) contribute more welfare loss per unit

of disequilibrium. The binding welfare constraint is the most institutionally rigid sector, not the most visibly disrupted one.

Current implication: The welfare-relevant bottleneck is more likely at the capability layer (slow-moving training pipelines, regulatory barriers to AI deployment) than at the settlement layer (fast-moving fintech). A policymaker focused on stablecoin regulation is optimizing the wrong margin.

3.6.5 The Logistic Fragility Condition

The power-law gain functions above have constant elasticity $\varepsilon_{T_n} = \beta_n$. The logistic case reveals a sharper phenomenon.

Proposition 3.6.5 (Logistic fragility)**.** *For logistic gain $\phi_n(z) = r_n z(1-z/K_n)$, the elasticity at equilibrium depends on the utilization ratio $u_n = \bar{F}_{n-1}/K_n$:*

$$\varepsilon_{T_n} = \frac{1 - 2u_n}{1 - u_n}. \tag{3.17}$$

The tree coefficient has a pole at $u_n = 1/2$:

$$c_n = \frac{P_{cycle}\, \sigma_{n-1}(1 - u_n)}{\sigma_n\, J\, \bar{F}_n(1 - 2u_n)}. \tag{3.18}$$

Stability of the welfare distance function requires $u_n < 1/2$ (operating below the logistic inflection point). At $u_n > 1/2$, the elasticity goes negative, the tree coefficient changes sign, and V ceases to be a welfare distance function.

Economic interpretation. As the upstream level approaches half its carrying capacity, the welfare weight at level n diverges—perturbations at that level dominate the welfare loss. This is the approach to the logistic peak: the system is maximally sensitive to perturbations near the inflection point of the S-curve. Operating above the logistic inflection is destabilizing, not merely inefficient.

Prediction: variance of mesh-related indicators spikes when agent density reaches approximately 50% of infrastructure capacity—at the inflection point, not at saturation. **Design criterion:** engineer carrying capacity so equilibrium utilization stays well below 50%.

3.7 The Damping Cancellation and Policy

3.7.1 The Damping-Speed Tradeoff

Proposition 3.7.1 (Damping cancellation)**.** *For the reduced system on the slow manifold:*

(i) The convergence speed at level n is σ_n/ε_n, strictly increasing in σ_n—more friction means faster convergence.

(ii) The equilibrium output is $\bar{F}_n = \phi_n(\bar{F}_{n-1})/(\sigma_n J)$, strictly decreasing in σ_n—more friction means lower output.

(iii) The welfare dissipation rate at level n near equilibrium is

$$-\dot{V}_n \approx \frac{P_{cycle}\,\sigma_{n-1}}{\beta_n\, J\, \bar{F}_n} \cdot \frac{(\delta F_n)^2}{\bar{F}_n} \tag{3.19}$$

*(under power-law gains), which is **independent of** σ_n **itself**.*

Proof. (i) The eigenvalue of the reduced Jacobian is $-\sigma_n/\varepsilon_n$ (Equation 3.5 restricted to the aggregate mode). (ii) Direct from the equilibrium condition. (iii) $\dot{V}_n = -c_n\sigma_n(\delta F_n)^2/\bar{F}_n$. Substituting c_n from Proposition 3.6.4: $c_n\sigma_n = P_{\text{cycle}}\sigma_{n-1}/(\beta_n J\bar{F}_n)$, which is independent of σ_n. □

Economic interpretation. Tightening regulation at sector n speeds up convergence to equilibrium but lowers the equilibrium itself. These two effects *exactly cancel* in the welfare dissipation. The net welfare effect of local regulation is zero. The welfare dissipation at sector n depends on σ_{n-1} (upstream friction) and β_n (the sector's own responsiveness to upstream improvements), not on σ_n.

3.7.2 The Upstream Reform Principle

Theorem 3.7.2 (Upstream reform principle)**.** *To accelerate welfare-relevant adjustment at sector n:*

1. *Increase β_n—make the sector more responsive to upstream improvements, OR*
2. *Reduce σ_{n-1}—reduce friction at the upstream sector.*
3. *Do NOT increase σ_n—tightening local regulation has zero net welfare effect.*

The policy chain:

- Fix settlement ($n = 4$): reform capability aggregation (σ_3) or increase settlement elasticity (β_4).
- Fix capability ($n = 3$): reform network recruitment (σ_2) or increase training elasticity (β_3).
- Fix network ($n = 2$): reform hardware investment (σ_1) or increase recruitment elasticity (β_2).
- Fix hardware ($n = 1$): reduce γ_c directly (CHIPS Act, semiconductor subsidies).

Corollary 3.7.3 (Zero welfare effect of stablecoin regulation)**.** *Stablecoin regulation (σ_4) has zero marginal welfare effect. Capability-layer reform (σ_3 or β_4) has positive marginal welfare effect. This is a theorem, not a heuristic.*

3.7.3 The Global Welfare Ordering

Corollary 3.7.4 (Welfare ordering)**.** *Under the partial order $\boldsymbol{\beta} \succeq \boldsymbol{\beta}'$ (all gain elasticities weakly higher):*

(i) $W_{nn}(\boldsymbol{\beta}) \leq W_{nn}(\boldsymbol{\beta}')$ for all n (the Bridge tightens—institutional quality improves).

(ii) $V(\boldsymbol{\beta}) \leq V(\boldsymbol{\beta}')$ at every non-equilibrium state (welfare loss decreases).

Economic interpretation. Increasing the responsiveness of cross-sector coupling at any level is unambiguously welfare-improving, regardless of the current state of the economy. Policies that increase cross-sector responsiveness are always welfare-improving. Policies that flatten response curves are always welfare-reducing.

3.7.4 The Rigidity Ordering

From the institutional quality matrix: $W_{11} > W_{22} > W_{33} > W_{44}$ (hardware stiffest, settlement loosest) when the timescale and depreciation orderings align. Policy interventions at stiff layers (semiconductor subsidies, export controls) have persistent effects. Interventions at loose layers (stablecoin regulation) have transient effects the system routes around.

3.8 Transition Dynamics

3.8.1 The Hierarchical Ceiling Cascade

Proposition 3.8.1 (Ceiling functions)**.** *Under the timescale ordering, successive equilibration yields:*

- ***Level 4*** *(fastest):* $F_4 \leq \bar{S}(F_3)$*. Settlement is bounded by capability.*
- ***Level 3****:* $F_3 \leq (\varphi_{eff}/\delta_C) \cdot F_{CES}(N^*(F_1))$*. Capability is bounded by network and hardware.*
- ***Level 2****:* $F_2 \leq N^*(F_1)$*. Network is bounded by hardware.*

The cascade of ceilings $F_1 \to F_2 \leq N^*(F_1) \to F_3 \leq (\varphi_{\text{eff}}/\delta_C)F_{\text{CES}}(N^*) \to F_4 \leq \bar{S}(F_3)$ bounds each level by a function of the level below in the timescale hierarchy. The long-run growth rate equals the hardware improvement rate—the slowest-adapting sector.

Economic interpretation. The Baumol cost disease and the Triffin squeeze are the same mathematical object—a slow manifold constraint—at adjacent layers. The Baumol bottleneck says: as the mesh automates inference, the remaining non-automated task (frontier training) becomes the binding constraint. The Triffin squeeze says: as stablecoin demand grows, Treasury supply must grow faster than the safety boundary shrinks. Both are instances of a faster sector being bounded by its slower parent.

3.8.2 The Transition Duration

When $\rho(\mathbf{K})$ crosses 1, the economy undergoes a structural transition—the low-activity equilibrium loses stability and the high-activity equilibrium becomes stable. But the transition takes time.

Theorem 3.8.2 (Transition duration)**.** *If the bifurcation parameter drifts at rate* ε_{drift}*, the transition duration is*

$$\Delta t_{crisis} = \frac{\pi}{\sqrt{|a|\,\varepsilon_{drift}}} + O\left(\frac{\log(1/\delta)}{\sqrt{|a|\,\varepsilon_{drift}}}\right) \tag{3.20}$$

where $a = \partial^2 g/\partial F_1 \partial \mu|_{bif}$ *is the sensitivity of the growth rate to the bifurcation parameter.*

If the drift is in institutional friction (γ_c *improving):* $a = -1$*, and the duration is* $\pi/\sqrt{\varepsilon_{drift}}$*, independent of all other parameters.*

At Wright's Law semiconductor improvement rates ($\approx 15\%$ *annual cost decline):* $\Delta t \approx \pi/\sqrt{0.15} \approx 8$ *years.*

Proof sketch. At the structural transition, the dynamics admit the local normal form $\dot{y} = a\epsilon y + by^2 + O(|y|^3 + |\epsilon|^2)$. If the bifurcation parameter drifts linearly, the rescaled system becomes the Weber equation plus a quadratic perturbation. The passage through zero eigenvalue creates a delay of π time units in the rescaled variable. Converting back gives the stated duration. This is the standard delayed loss of stability result [9, 10, 11]. See Appendix .4 for the full development. □

Economic interpretation. After conditions become favorable for the high-activity equilibrium, the economy lingers near the old equilibrium for $O(1/\sqrt{\varepsilon_{\text{drift}}})$ time before snapping to the new one. This is the "crisis duration"—the period of structural transition. It is computable from observable drift rates.

The mixed partial a has two natural cases depending on which parameter drifts:

- **Case 1:** $\mu = \gamma_c$ (institutional friction at the slowest level improves). Then $a = -1$ and the duration is $\pi/\sqrt{\varepsilon_{\text{drift}}}$, independent of all system parameters except the drift rate. This is the simplest case: if what's improving is the institutional friction, the transition time depends only on how fast it improves.

- **Case 2:** $\mu = \delta_c$ (investment efficiency improves). Then $|a|$ depends on the product of cascade elasticities through the entire hierarchy:

$$\frac{\Psi'(\bar{F}_1)\bar{F}_1}{\Psi(\bar{F}_1)} = \varepsilon_I \cdot \varepsilon_{\bar{S}} \cdot \varepsilon_{F_{\text{CES}}} \cdot \varepsilon_{h_2} \tag{3.21}$$

  where ε_I, $\varepsilon_{\bar{S}}$, $\varepsilon_{F_{\text{CES}}}$, and ε_{h_2} are the elasticities of the settlement investment function, the settlement ceiling, the CES capability aggregate, and the network ceiling respectively. The CES elasticity at the symmetric allocation is $\varepsilon_{F_{\text{CES}}} = 1/J$.

Proposition 3.8.3 (Curvature dependence of the transition)**.** *The second-order coefficient $b = \frac{1}{2}\partial^2 g/\partial F_1^2|_{bif}$ contains the CES second derivative*

$$\left.\frac{\partial^2 F_{CES}}{\partial F_2^2}\right|_{sym} = -\frac{K}{J\bar{F}_2} \tag{3.22}$$

(from the CES Hessian restricted to the aggregate direction, Paper A [1]). Therefore $|b|$ increases with K: higher curvature (stronger complementarity) increases the sharpness of the transition. However, b does not appear in the leading-order transition duration $\pi/\sqrt{|a|\varepsilon_{drift}}$—it enters only in the correction terms and in the amplitude of the post-transition trajectory.

Economic interpretation. K controls the *sharpness* of the transition (how quickly the economy accelerates once it begins transitioning), not the *duration*. Higher complementarity

means a faster snap to the new equilibrium, with less overshooting. The economically important timing question "how long does the transition take?" has the answer: $O(1/\sqrt{\varepsilon_{\text{drift}}})$, with the constant controlled by the chain of gain elasticities (Equation 3.21), not by the CES substitution parameter.

3.8.3 Dispersion as Leading Indicator

At the structural transition, the spectral gap between diversity and aggregate modes closes. Within-sector heterogeneity stops being slaved to the aggregate.

Prediction: cross-sectional variance of agent performance widens *before* aggregate statistics move. The within-mesh Gini coefficient rises before the crossing and collapses after (as diversity modes re-equilibrate on the new slow manifold).

3.9 Empirical Predictions

3.9.1 Calibration Inputs

Semiconductor learning curves provide the drift rate $\varepsilon_{\text{drift}}$: at Wright's Law rates of approximately 15% annual cost decline, the transition duration formula yields a transition window of order 8 years. The monetary productivity gap (6.4 percentage points) anchors the settlement cost advantage.

3.9.2 Predictions

P1–P3: Testing the Triple Role (Paper A [1], applied at each level).
(P1) Cross-agent capability profiles on the unit isoquant diverge as the mesh matures, with superadditivity gap proportional to $K \cdot d_{\mathcal{I}}^2$. *Falsification:* agent capability profiles converge rather than diverge.
(P2) Model collapse incidence remains below threshold for mesh-trained agents, with effective quality bounded below by a function of K. *Falsification:* mesh-trained agents exhibit systematic model collapse.
(P3) Coalition manipulation gain in mesh-mediated markets satisfies $\Delta(S) \leq 0$ with penalty proportional to K_S. *Falsification:* sustained profitable manipulation by coalitions in mesh-mediated markets.

P4: Testing the spectral threshold.

Cross-layer acceleration occurs with delay $\approx \pi/\sqrt{|a|\varepsilon_{\text{drift}}}$ after the drift parameter crosses the threshold. At Wright's Law rates: 6–10 year transition window. *Falsification:* no acceleration by 2035.

P5–P6: Testing monetary policy degradation.
(P5) Forward guidance effectiveness declines before QE effectiveness, which declines before financial repression collapses. *Falsification:* policy tools degrade in a different order.
(P6) The duration of market impact from FOMC statements declines as autonomous agent market share grows. *Falsification:* impact duration increases or remains constant.

P7–P8: Testing settlement feedback.
(P7) Stablecoin Treasury holdings exceed 5% of short-duration Treasury supply by 2028. *Falsification:* stablecoin Treasury share below 3% by 2029.
(P8) At least one country group experiences stablecoin-mediated dollarization by 2030. *Falsification:* no countries show stablecoin-driven dollarization patterns by 2031.

P9: Testing the hierarchical ceiling.
The ratio of mesh capability growth to frontier training rate converges: $\dot{C}_{\text{mesh}}/\dot{C}_{\text{frontier}} \to 1$. *Falsification:* mesh capability growth consistently exceeds frontier training rate.

P10: Testing damping cancellation.
Tightening stablecoin regulation (σ_4) has no persistent effect on mesh welfare convergence. Capability-layer reforms (reducing σ_3 or increasing β_3) do. *Falsification:* stablecoin regulation measurably slows or accelerates mesh formation over a 3+ year horizon.

P11: Testing the dispersion indicator.
Cross-sectional variance of AI agent performance metrics widens before aggregate mesh statistics shift. *Falsification:* aggregate leads dispersion.

3.9.3 Preliminary Evidence—Port Topology

Theorem 3.3.1(iv) predicts nearest-neighbor coupling: the 4-level hierarchy has a tridiagonal interaction graph, so shocks at Level n propagate to Level $n \pm 1$ but not directly to distant levels. A VAR(1) model on monthly log-differenced proxies for all four levels (2020–2025, $T = 57$) provides a preliminary test.[1]

Four tests assess the nearest-neighbor prediction:

[1]Scripts and replication data at `scripts/test_cross_layer_var.py`. Proxies: L1 = FRED semiconductor IP index; L2 = HuggingFace ecosystem growth (milestone interpolation); L3 = inference API pricing (log-linear interpolation); L4 = DeFiLlama stablecoin market cap.

1. *Granger causality.* Of the six nearest-neighbor pairs, one is significant: L1→L2 ($p = 0.058$). Of the six distant pairs, one is significant: L2→L4 ($p = 0.048$). **Verdict:** ambiguous—the pattern is not cleanly tridiagonal, though the strongest link is a nearest-neighbor pair.

2. *Forecast error variance decomposition (FEVD).* At $h = 12$ months, 2 of 3 non-own layers show neighbor share exceeding distant share: L2's variance is 20.0% from neighbors versus 0.3% distant; L3 is 14.3% neighbor versus 3.0% distant. L4 violates the prediction (0.5% neighbor vs. 15.1% distant). **Verdict:** partially consistent.

3. *Block exogeneity* (the most directly theory-relevant test). Does L1 affect L4 conditional on L2 and L3? All three block exogeneity tests are consistent with tridiagonal topology: L1→L4 ($F = 0.73$, $p = 0.40$), L4→L1 ($F = 0.53$, $p = 0.47$), and L1→L3 ($F = 0.04$, $p = 0.85$). The distant-layer coefficients are jointly indistinguishable from zero once intermediate layers are included. **Verdict:** consistent.

4. *Impulse response sequencing.* An L1 shock should hit L2 first, then L3, then L4 with increasing delay. The observed peak timing (L1 at $h = 1$, L2 at $h = 2$, L3 and L4 at $h = 1$) is not strictly sequential. **Verdict:** not consistent, though the short sample may conflate contemporaneous effects with lagged propagation at monthly frequency.

Overall assessment: the strongest test (block exogeneity) supports the tridiagonal prediction; FEVD partially supports it; Granger causality and IRF sequencing are ambiguous. The binding constraint is sample length: 57 monthly observations provide limited power for a 4-variable system. As the sample extends and higher-frequency proxies become available (particularly for L2 and L3, which currently rely on milestone interpolation), the test will sharpen.

3.10 Limitations

Two categories of limitations, stated without apology.

3.10.1 Mathematical

1. *Gain functions genuinely free.* Theorem 3.3.1(iii) establishes this as a proved impossibility: the exponents and coefficients of ϕ_n are not determined by ρ. Since ϕ_n determines the equilibrium cascade $\{\bar{F}_n\}$, the welfare weights $\{c_n\}$, and the institutional quality matrix W, this freedom propagates through the quantitative predictions.

2. *Timescale separation cannot be eliminated.* The fast sector equilibrates before the slow sector moves appreciably (Standing Assumption 3) is required for the slow manifold to exist and for the $NJ \to N$ dimensional reduction. Without it, the slow manifold need not exist, and the reduction is unjustified. The nearest-neighbor topology (Theorem 3.3.1(iv)) also requires timescale separation.

3. *The system is not a gradient flow.* The lower-triangular Jacobian is a topological obstruction (Proposition 3.6.1). No coordinate transformation can make the system a gradient flow while preserving directed coupling. Standard welfare theorems do not apply.

4. *Local stability only.* Theorem 3.6.2 proves $\dot{V} \leq 0$, establishing local asymptotic stability. Global asymptotic stability requires boundary analysis depending on the specific gain functions.

5. *Symmetric weights for quantitative bounds.* General weights yield results via the secular equation (Section 3.3.2; Paper A [1], Section 8), with $R_{\min}$ replacing the equal-weight curvature. The qualitative results are unchanged; bounds are less clean but remain computable.

6. $O(\varepsilon)$ *approximation error.* The sufficient statistic property holds up to $O(\varepsilon)$ corrections. During crisis episodes, within-sector composition may matter.

7. *Canard duration is leading order.* Correction terms involve K through the quadratic coefficient b; amplitude behavior is less precisely bounded.

3.10.2 Empirical

1. Gain elasticities $(\beta_1, \ldots, \beta_4)$ uncalibrated.

2. Damping rates $(\sigma_1, \ldots, \sigma_4)$ uncalibrated.

3. Predictions span 2027–2040 (long horizon).

4. Which specific layer binds first depends on uncalibrated β_n.

5. Crisis duration estimate requires $\varepsilon_{\text{drift}}$, which is itself uncertain.

3.10.3 Frameworks Considered and Rejected

Mean field games (Lasry-Lions): agents are not exchangeable. The CES structure ($\rho < 1$) ensures non-exchangeability; MFG would average over the heterogeneity that drives both efficiency results and collusion resistance.

Minsky financial instability hypothesis: insufficiently formalized for the results needed here. The Brunnermeier-Sannikov [23] framework captures the same insight rigorously.

Full continuous-time general equilibrium: intractable. The four-ODE deterministic skeleton captures qualitative dynamics; a stochastic extension is deferred.

3.11 Conclusion

One production difficulty function $\Phi = -\sum_n \log F_n$, one derived architecture, five results, three policy principles.

The CES geometry forces the architecture: aggregate coupling, directed feed-forward, nearest-neighbor chain. The architecture is derived, not assumed.

The spectral threshold $\rho(\mathbf{K}) = 1$ activates the hierarchy—individual sectors can each be too weak to sustain themselves, yet the system as a whole sustains activity through cross-sector feedback. The hierarchical ceiling cascade bounds each layer by its parent. The welfare distance function connects the technology (Φ) to the welfare loss (V) through the institutional quality (W).

Three policy principles follow from theorems:

1. *Reform upstream, not locally* (damping cancellation, Proposition 3.7.1). Tightening regulation at any sector has zero net welfare effect. To improve welfare at sector n, reform sector $n-1$.

2. *Increase cross-sector responsiveness at any layer* (global welfare ordering, Corollary 3.7.4). More responsive gain functions are unambiguously welfare-improving.

3. *Invest in the weakest cross-sector link* (cycle product sensitivity, Theorem 3.4.2). The system's activation is bottlenecked by its weakest coupling. The transition takes $O(1/\sqrt{\varepsilon_{\text{drift}}})$—at current semiconductor improvement rates, approximately 8 years.

Eleven predictions, spanning 2027–2040, test the theory.

Bibliography

[1] Smirl, J. (2026a). The CES Triple Role: Superadditivity, Correlation Robustness, and Strategic Independence as Three Views of Isoquant Curvature. Working paper.

[2] Li, M. Y., Shuai, Z., and van den Driessche, P. (2010). Global-stability problem for coupled systems of differential equations on networks. *J. Differential Equations* 248, 1–20.

[3] Shuai, Z., and van den Driessche, P. (2013). Global stability of infectious disease models using Lyapunov functions. *SIAM J. Appl. Math.* 73, 1513–1532.

[4] Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* 28, 365–382.

[5] Van den Driessche, P., and Watmough, J. (2002). Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* 180, 29–48.

[6] Fenichel, N. (1979). Geometric singular perturbation theory for ordinary differential equations. *J. Differential Equations* 31, 53–98.

[7] van der Schaft, A., and Jeltsema, D. (2014). Port-Hamiltonian systems theory: An introductory overview. *Found. Trends Syst. Control* 1(2–3), 173–378.

[8] do Carmo, M. P. (1992). *Riemannian Geometry*. Birkhäuser.

[9] Neishtadt, A. I. (1987). Persistence of stability loss for dynamical bifurcations, I. *Differential Equations* 23, 1385–1391.

[10] Neishtadt, A. I. (1988). Persistence of stability loss for dynamical bifurcations, II. *Differential Equations* 24, 171–176.

[11] Berglund, N., and Gentz, B. (2006). *Noise-Induced Phenomena in Slow-Fast Dynamical Systems*. Springer.

[12] Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M. (1961). Capital-labor substitution and economic efficiency. *Rev. Econ. Stat.* 43, 225–250.

[13] Dixit, A. K., and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *Amer. Econ. Rev.* 67, 297–308.

[14] Jones, C. I. (2005). The shape of production functions and the direction of technical change. *Quart. J. Econ.* 120, 517–549.

[15] Shapley, L. S. (1971). Cores of convex games. *Int. J. Game Theory* 1, 11–26.

[16] Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512.

[17] Pastor-Satorras, R., and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200–3203.

[18] Jones, C. I. (1995). R&D-based models of economic growth. *J. Polit. Econ.* 103, 759–784.

[19] Romer, P. M. (1990). Endogenous technological change. *J. Polit. Econ.* 98, S71–S102.

[20] Aghion, P., Jones, B. F., and Jones, C. I. (2018). Artificial intelligence and economic growth. In *The Economics of Artificial Intelligence*, University of Chicago Press.

[21] Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *Amer. Econ. Rev.* 110, 1104–1144.

[22] Baumol, W. J. (1967). Macroeconomics of unbalanced growth. *Amer. Econ. Rev.* 57, 415–426.

[23] Brunnermeier, M. K., and Sannikov, Y. (2014). A macroeconomic model with a financial sector. *Amer. Econ. Rev.* 104, 379–421.

[24] Woodford, M. (2003). *Interest and Prices*. Princeton University Press.

[25] Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conf. Ser. Public Policy* 1, 19–46.

[26] Uribe, M. (1997). Hysteresis in a simple model of currency substitution. *J. Monet. Econ.* 40, 185–202.

[27] Farhi, E., and Maggiori, M. (2018). A model of the international monetary system. *Quart. J. Econ.* 133, 295–355.

[28] Triffin, R. (1960). *Gold and the Dollar Crisis*. Yale University Press.

[29] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica* 53, 1315–1335.

[30] Hordijk, W., and Steel, M. (2004). Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.* 227, 451–461.

[31] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

.1 Proof of the Port Topology Theorem

.1.1 Proof of Claim (i): Aggregate Coupling

Proof. Step 1: Equilibrium uniqueness. At equilibrium, $(T_n/J)\, x_{nj}^{\rho-1}\, F_n^{1-\rho} = \sigma_n x_{nj}$, so $x_{nj}^{\rho-2} = T_n F_n^{1-\rho}/(J\sigma_n)$. The right side is independent of j. Since $\rho < 1$ implies $\rho - 2 < 0$, the map $x \mapsto x^{\rho-2}$ is injective on $\mathbb{R}_+$, so $x_{nj} = \bar{x}_n$ for all j.

Step 2: Normal hyperbolicity. Define the critical manifold $\mathcal{M}_n = \{x_{nj} = F_n/J^{1/\rho} \text{ for all } j\}$. By Equation (3.5), the transverse eigenvalues (on $\mathbf{1}^\perp$) are $-\sigma_n(2-\rho)/\varepsilon_n$ and the tangential eigenvalue (on $\mathbf{1}$) is $-\sigma_n/\varepsilon_n$. The spectral gap is $\sigma_n(2-\rho)/\varepsilon_n - \sigma_n/\varepsilon_n = \sigma_n(1-\rho)/\varepsilon_n > 0$ for all $\rho < 1$.

Step 3: Fenichel persistence. By Fenichel's geometric singular perturbation theorem [6], the normally hyperbolic critical manifold $\mathcal{M}_n$ persists as a locally invariant slow manifold $\mathcal{M}_n^\varepsilon$ within $O(\varepsilon)$ of $\mathcal{M}_n$, smoothly parameterized by F_n. On $\mathcal{M}_n^\varepsilon$, the within-level state is determined by F_n up to $O(\varepsilon)$ corrections. Consequently, F_n is a sufficient statistic for level n's state on the slow manifold. □

.1.2 Proof of Claim (ii): Directed Coupling

Proof. Step 1: Power-preserving bidirectional coupling. Suppose levels 1 and 2 are coupled bidirectionally with port powers summing to zero. The power-preservation constraint forces linear coupling $\phi(F_1) = cF_1$ and $\psi(F_2) = cF_2$ for some constant $c > 0$. The Jacobian is then $\mathcal{J}_{\text{bidir}} = \begin{pmatrix} -\sigma_1 & -c/J \\ c/J & -\sigma_2 \end{pmatrix}$ with eigenvalues $-(\sigma_1+\sigma_2)/2 \pm \sqrt{(\sigma_1-\sigma_2)^2/4 - c^2/J^2}$. Whether the discriminant is positive (two real negative eigenvalues) or negative (complex pair with real part $-(\sigma_1+\sigma_2)/2 < 0$), both eigenvalues have strictly negative real part. The system is unconditionally stable for all $c, \sigma_1, \sigma_2 > 0$.

Step 2: Extension to passive bidirectional coupling. Any passive bidirectional coupling satisfies $\dot{V}_{\text{coupling}} \leq 0$ by definition. This contributes a negative-semidefinite term to the effective Jacobian, which can only strengthen stability. Therefore any bidirectional coupling—whether power-preserving or merely passive—yields an unconditionally stable system.

Step 3: Necessity of directed coupling. The structural transition at $\rho(\mathbf{K}) = 1$ requires that the spectral radius of the next-generation matrix reach 1, requiring net energy injection through the hierarchy. Power-preserving bidirectional coupling contributes zero net energy; passive coupling contributes negative net energy. Neither can produce $\rho(\mathbf{K}) = 1$. Therefore the CES geometry, combined with the requirement for nontrivial dynamics, forces the between-level coupling to be non-reciprocal with an external energy source. □

.1.3 Proof of Claim (iii): Port Alignment

Proof. Step 1: Port direction is forced. At a symmetric equilibrium $\mathbf{x}_n^* = \bar{x}\,\mathbf{1}$, the equilibrium condition requires the coupling direction $\mathbf{b}_n \propto \mathbf{1}$. Furthermore, $\nabla F_n = (1/J)\,\mathbf{1} \propto \mathbf{1}$ at the symmetric point (Paper A [1]), so $\mathbf{b}_n = \nabla F_n$ is the natural CES-compatible port direction.

Step 2: Asymmetric ports are penalized. For $\mathbf{b}_n \not\propto \mathbf{1}$, the equilibrium $\mathbf{x}_n^* \propto \mathbf{b}_n$ is asymmetric. By Jensen's inequality applied to the concave CES function ($\rho < 1$), $F(\mathbf{x}) \leq F(\bar{x}\,\mathbf{1})$ for any $\mathbf{x}$ with $\sum x_j = J\bar{x}$. Asymmetric ports produce less aggregate output per unit input.

Step 3: Gain function is free. At equilibrium, $\phi_n(\bar{F}_{n-1}) = \sigma_n J \bar{F}_n$. For power-law gains $\phi_n(z) = a_n z^{\beta_n}$, the exponents β_n are free parameters not determined by ρ. The coefficients a_n adjust to satisfy the equilibrium condition but depend on σ_n, J, and the cascade—not on ρ alone. □

.1.4 Proof of Claim (iv): Nearest-Neighbor Topology

Proof. Consider a three-level system with long-range coupling from level 1 to level 3. Level 1, being fastest ($\varepsilon_1 \ll \varepsilon_2$), equilibrates to $F_1^* = \beta_1/(\sigma_1 J)$, an algebraic constant on the slow manifold. The long-range coupling $\phi_{31}(F_1^*) = \phi_{31}(\beta_1/(\sigma_1 J)) \equiv \tilde{\beta}_3$ becomes a constant, absorbed into the exogenous input to level 3. The effective dynamics become identical to a nearest-neighbor system with modified exogenous input.

The Jacobian of the reduced system is lower-triangular—independent of the long-range coupling strength. Long-range coupling affects the equilibrium location but not the dynamics or stability near that equilibrium.

The argument generalizes by induction to N levels: after all levels faster than level m equilibrate, any coupling $\phi_{nm}(F_m)$ with m fast becomes $\phi_{nm}(F_m^*) = \text{const}$, absorbed into a redefined exogenous input. The effective topology is nearest-neighbor. This result requires the timescale separation of Standing Assumption (3). □

.2 Proof of the Welfare Distance Function

Proof. Nonnegativity follows from $g(z) = z - 1 - \log z \geq 0$ with equality iff $z = 1$. Along trajectories:

$$\dot{V} = \sum_n c_n \sum_j \left(1 - \frac{x_{nj}^*}{x_{nj}}\right) f_{nj}(\mathbf{x}).$$

The within-level contributions are

$$\dot{V}_{\text{within}} = -\sum_n c_n \sigma_n \sum_j \frac{(x_{nj} - x_{nj}^*)^2}{x_{nj}} \leq 0.$$

The cross-level contributions involve terms of the form

$$c_n \sum_j \left(1 - \frac{x_{nj}^*}{x_{nj}}\right) \frac{T_n}{J} x_{nj}^{\rho-1} F_n^{1-\rho}.$$

At the symmetric equilibrium $x_{nj}^* = \bar{x}_n$, these contributions cancel by the tree condition on $c_n = P_{\text{cycle}}/k_{n,n-1}$. This is the Li-Shuai-van den Driessche [2] construction applied to the cycle-graph topology, using the Volterra-Lyapunov identity $a - b\log(a/b) \leq a - b + b\log(b/a)$ for each coupling term.

For the 4-level cycle, there is exactly one spanning in-tree per root, and the specific coefficients $c_n = P_{\text{cycle}}/k_{n,n-1}$ are exactly those required for cancellation. □

.3 Proof of the Eigenstructure Bridge

Proof. On the slow manifold, $\Phi|_{\text{slow}} = -\sum_n \log F_n$ and $V = \sum_n c_n \bar{F}_n\, g(F_n/\bar{F}_n)$. Their Hessians at equilibrium are diagonal:

$$(\nabla^2 \Phi|_{\text{slow}})_{nn} = \frac{1}{\bar{F}_n^2}, \qquad (\nabla^2 V)_{nn} = \frac{c_n}{\bar{F}_n}.$$

The ratio is $(\nabla^2\Phi)_{nn}/(\nabla^2 V)_{nn} = 1/(c_n \bar{F}_n) = W_{nn}^{-1}$.

Expressing $c_n = P_{\text{cycle}}/k_{n,n-1}$ and $k_{n,n-1} = T_n'(\bar{F}_{n-1})\bar{F}_{n-1}/|\sigma_{n-1}|$, together with the equilibrium relation $T_n(\bar{F}_{n-1}) = |\sigma_n|\bar{F}_n$:

$$\begin{aligned} c_n \bar{F}_n &= \frac{P_{\text{cycle}}}{k_{n,n-1}} \cdot \bar{F}_n = P_{\text{cycle}} \cdot \frac{|\sigma_{n-1}|}{T_n' \bar{F}_{n-1}} \cdot \bar{F}_n \\ &= \frac{P_{\text{cycle}}}{|\sigma_n|} \cdot \frac{T_n}{T_n' \bar{F}_{n-1}} = \frac{P_{\text{cycle}}}{|\sigma_n|\, \varepsilon_{T_n}} \end{aligned}$$

where $\varepsilon_{T_n} = T_n'(\bar{F}_{n-1})\bar{F}_{n-1}/T_n(\bar{F}_{n-1})$ is the elasticity of the coupling at level n.

Special cases: *Power-law coupling* ($\phi_n(z) = a_n z^{\beta_n}$): the elasticity is constant, $\varepsilon_{T_n} = \beta_n$, giving $W_{nn} = P_{\text{cycle}}/(\beta_n|\sigma_n|)$.

Linear coupling ($\beta_n = 1$) with uniform damping ($\sigma_n = \sigma$): $W = (P_{\text{cycle}}/\sigma)\, I$, so $\nabla^2\Phi|_{\text{slow}} = (\sigma/P_{\text{cycle}})\, \nabla^2 V$. This is the only case where the system is "almost" a gradient flow. □

.4 Transition Dynamics: The Normal Form

.4.1 The Transcritical Normal Form

Proof. At the bifurcation point $(\bar{F}_1, \mu^*)$ where $g = 0$ and $\partial g/\partial F_1 = 0$, the dynamics admit the local normal form

$$\dot{y} = a\,\epsilon\,y + b\,y^2 + O(|y|^3 + |\epsilon|^2)$$

where $y = F_1 - \bar{F}_1$, $\epsilon = \mu - \mu^*$, and

$$a = \left.\frac{\partial^2 g}{\partial F_1\,\partial \mu}\right|_{\text{bif}}, \qquad b = \frac{1}{2}\left.\frac{\partial^2 g}{\partial F_1^2}\right|_{\text{bif}}.$$

The conditions $g = 0$ and $\partial_F g = 0$ eliminate the constant and linear terms. The nondegeneracy conditions $a \neq 0$ and $b \neq 0$ are the transversality requirements for a structural transition. □

.4.2 Computing the Mixed Partial

Case 1: $\mu = \gamma_c$ (institutional friction improves). $g = \delta_c \Psi(F_1)^\alpha F_1^{\phi_c} - \gamma_c F_1$, so $\partial g/\partial \gamma_c = -F_1$ and $\partial^2 g/\partial F_1 \partial \gamma_c = -1$. Thus $a = -1$. The transition duration is $\pi/\sqrt{\varepsilon_{\text{drift}}}$, independent of all system parameters except the drift rate.

Case 2: $\mu = \delta_c$ (investment efficiency improves). $\partial^2 g/\partial F_1 \partial \delta_c = [\alpha \Psi'/\Psi \cdot F_1 + \phi_c]\Psi^\alpha F_1^{\phi_c - 1}$. The term Ψ'/Ψ propagates through the entire cascade of ceiling functions. We now derive this cascade explicitly.

Step 1: Decompose the composite ceiling. The full ceiling function $\Psi(F_1) = I(\bar{S}(F_{\text{CES}}(h_2(F_1))))$ chains four maps:

- h_2: the network ceiling $F_2 \leq N^*(F_1)$, with elasticity $\varepsilon_{h_2} = h_2'(\bar{F}_1)\bar{F}_1/h_2(\bar{F}_1)$;
- F_{CES}: the CES aggregate at the capability level, with elasticity $\varepsilon_{F_{\text{CES}}}$;
- $\bar{S}$: the settlement ceiling, with elasticity $\varepsilon_{\bar{S}}$;
- I: the settlement investment function, with elasticity ε_I.

Step 2: Apply the chain rule. By the chain rule for elasticities:

$$\frac{\Psi'(\bar{F}_1)\bar{F}_1}{\Psi(\bar{F}_1)} = \varepsilon_I \cdot \varepsilon_{\bar{S}} \cdot \varepsilon_{F_{\text{CES}}} \cdot \varepsilon_{h_2}.$$

This is Equation (3.21) in the main text.

Step 3: Evaluate the CES elasticity. At the symmetric allocation $x_{nj} = \bar{x}_n$ for all j:

$$\varepsilon_{F_{\text{CES}}} = \left.\frac{\partial F_{\text{CES}}}{\partial F_2}\right|_{\text{sym}} \cdot \frac{\bar{F}_2}{\bar{F}_{\text{CES}}} = \frac{1}{J}$$

since each of J symmetric inputs contributes equally to the aggregate. Hence the cascade elasticity simplifies to $\varepsilon_I \cdot \varepsilon_{\bar{S}} \cdot \varepsilon_{h_2}/J$.

.4.3 Where K Enters

Proof of Proposition 3.8.3. Step 1: CES second derivative at the symmetric point. From Paper A [1], the CES function $F_{\text{CES}} = \left(\frac{1}{J}\sum_{j=1}^{J} x_j^{\rho}\right)^{1/\rho}$ has second derivative along the aggregate direction $\mathbf{1}$:

$$\left.\frac{\partial^2 F_{\text{CES}}}{\partial F_2^2}\right|_{\text{sym}} = \frac{\rho - 1}{J\bar{F}_2} = -\frac{K}{J\bar{F}_2}$$

where $K = (1-\rho)(J-1)/J$ and the restriction to the aggregate direction absorbs the $(J-1)/J$ factor as $(1-\rho)/J$.

Step 2: Propagation to b. The coefficient $b = \frac{1}{2}\partial^2 g/\partial F_1^2|_{\text{bif}}$ inherits a contribution from $\partial^2 F_{\text{CES}}/\partial F_2^2$ through the chain of ceiling functions. The chain rule for the second derivative of the composite $\Psi(F_1)$ yields terms involving Ψ'', which in turn contains $\partial^2 F_{\text{CES}}/\partial F_2^2$. Therefore $|b|$ is increasing in K: higher curvature (stronger complementarity) amplifies the quadratic nonlinearity.

Step 3: Separation of roles. K enters b (the *sharpness* coefficient) but not a (the *duration* coefficient). The leading-order transition time $\pi/\sqrt{|a|\varepsilon_{\text{drift}}}$ is independent of K. K controls only the correction terms and the post-transition trajectory amplitude. □

.4.4 Passage Time

In the rescaled variable $\tau = \sqrt{|a|\varepsilon_{\text{drift}}}\, t$, the normal form becomes the Weber equation plus a quadratic perturbation. The passage through zero eigenvalue creates a delay of π time units in the rescaled variable. Converting back: $\Delta t = \pi/\sqrt{|a|\varepsilon_{\text{drift}}}$. This is the standard delayed loss of stability result for passage through a structural transition [9, 10, 11].

The logarithmic correction accounts for the entry and exit from the $O(\delta)$-neighborhood, which depends on the initial distance from the slow manifold.

At Wright's Law semiconductor improvement rates ($\varepsilon_{\text{drift}} \approx 0.15$): $\Delta t \approx \pi/\sqrt{0.15} \approx 8.1$ years.

In both cases, the crisis duration scales as $\varepsilon_{\text{drift}}^{-1/2}$: halving the drift rate increases the transition time by $\sqrt{2} \approx 41\%$.

.5 The Welfare Loss Decomposition

Proof of Proposition 3.6.4. From Theorem 3.6.2, $c_n = P_{\text{cycle}}/k_{n,n-1}$ where $k_{n,n-1} = \phi_n'(\bar{F}_{n-1})\bar{F}_{n-1}/|\sigma_{n-1}|$. For power-law $\phi_n(z) = a_n z^{\beta_n}$: $\phi_n'(\bar{F}_{n-1}) \cdot \bar{F}_{n-1} = \beta_n\, \phi_n(\bar{F}_{n-1}) = \beta_n\, \sigma_n\, J\, \bar{F}_n$ (using the equilibrium condition $\phi_n(\bar{F}_{n-1}) = \sigma_n J \bar{F}_n$). Thus $k_{n,n-1} = \beta_n\, \sigma_n\, J\, \bar{F}_n/\sigma_{n-1}$, giving $c_n = P_{\text{cycle}}\, \sigma_{n-1}/(\beta_n\, \sigma_n\, J\, \bar{F}_n)$. On the slow manifold, $c_n D_{KL} = c_n \bar{F}_n g(F_n/\bar{F}_n) = P_{\text{cycle}}\sigma_{n-1}/(\beta_n\sigma_n J)\, g(F_n/\bar{F}_n)$. □

Proof of Proposition 3.7.1. (i) The eigenvalue of the reduced Jacobian is $-\sigma_n/\varepsilon_n$ (Equation 3.5 restricted to the aggregate mode). (ii) Direct from the equilibrium condition $\bar{F}_n = \phi_n(\bar{F}_{n-1})/(\sigma_n J)$. (iii) $\dot{V}_n = -c_n\sigma_n(F_n - \bar{F}_n)^2/F_n \approx -c_n\sigma_n(\delta F_n)^2/\bar{F}_n$. Substituting c_n from Proposition 3.6.4: $c_n\sigma_n = P_{\text{cycle}}\sigma_{n-1}/(\beta_n J\bar{F}_n)$, which is independent of σ_n. The σ_n in c_n exactly cancels the σ_n in the dissipation formula. □

Proof of Corollary 3.7.4. (i) $W_{nn} = P_{\text{cycle}}/(\beta_n|\sigma_n|)$ under power-law gains, strictly decreasing in β_n. Higher β_n means lower W_{nn}, meaning tighter institutional quality. (ii) From $c_n \propto 1/\beta_n$, higher β_n gives lower weight c_n, hence lower V at any non-equilibrium state. □

Proof of Proposition 3.6.5. For logistic gain $\phi_n(z) = r_n z(1 - z/K_n)$:

$$\phi_n'(z) = r_n(1 - 2z/K_n), \qquad \phi_n(z) = r_n z(1 - z/K_n).$$

The elasticity at $z = \bar{F}_{n-1}$ with utilization $u_n = \bar{F}_{n-1}/K_n$ is:

$$\varepsilon_{T_n} = \frac{\phi_n'(\bar{F}_{n-1})\bar{F}_{n-1}}{\phi_n(\bar{F}_{n-1})} = \frac{r_n(1 - 2u_n)\bar{F}_{n-1}}{r_n\bar{F}_{n-1}(1 - u_n)} = \frac{1 - 2u_n}{1 - u_n}.$$

This has a pole at $u_n = 1/2$ (the logistic inflection point) and changes sign there.

The tree coefficient is $c_n = P_{\text{cycle}}/k_{n,n-1}$ with $k_{n,n-1} = \varepsilon_{T_n}\sigma_n J\bar{F}_n/\sigma_{n-1}$. Substituting ε_{T_n}:

$$c_n = \frac{P_{\text{cycle}}\, \sigma_{n-1}(1 - u_n)}{\sigma_n J\bar{F}_n(1 - 2u_n)}.$$

For $u_n > 1/2$: $\varepsilon_{T_n} < 0$, so $c_n < 0$, and $V = \sum c_n \bar{F}_n g(F_n/\bar{F}_n)$ ceases to be positive semidefinite. The Lyapunov argument fails, and local stability is no longer guaranteed by the graph-theoretic construction. □

Chapter A

Endogenous Decentralization

A.1 Introduction

Between 2018 and 2026, the five largest US technology companies—together with Oracle, xAI, and the Stargate joint venture—have committed an estimated $2.4 trillion in cumulative capital expenditure to construct centralized AI infrastructure.[1] The result: approximately 15 million H100-equivalent GPUs deployed globally as of late 2025, with the installed base growing at 3.3× per year—a doubling time of roughly seven months (Epoch AI 2026). This represents the largest concentrated infrastructure investment in history outside wartime mobilization. The near-term revenue objective is to sell AI inference—running trained models to serve user requests—as a cloud service at premium margins. A second, longer-horizon objective is frontier model training at scales that may produce discontinuous capability advances. The mechanism identified in this paper applies to the inference objective; the training objective is addressed as an alternative specification of the firms' objective function (Section 3.1).

This paper argues that this investment is *endogenously self-disrupting*: the very act of building centralized AI datacenters finances the component learning curves—particularly in 3D memory stacking, advanced packaging, and model compression—that enable distributed alternatives to replicate datacenter-class inference on consumer hardware. The operative learning curve is not the mature planar DRAM die, whose cost trajectory is near-asymptotic after four decades of cumulative production, but the *packaging and stacking* technologies that hyperscaler HBM demand is financing through their early high-learning-rate phase. As of Q1 2026, the technology threshold for interactive 70B-class inference has been met at professional and enthusiast price points. Paradoxically, the same concentrated investment has also triggered the most severe DRAM supercycle in two decades, temporarily reversing consumer memory cost trends and inflating GPU prices far above MSRP—a boom-phase deviation that the model's capacity-constraint corollary predicts will resolve into overcapacity and below-trend pricing as new advanced packaging capacity ramps. The remaining constraint is price migration from professional to mass-market form factors—a market structure transition compounded by, but not permanently altered by, the current supply shock.

Two structural features of the current AI landscape sharpen the mechanism beyond what prior transitions exhibited.

First, AI workloads bifurcate into *training* (creating models via massive synchronized GPU clusters) and *inference* (running models to serve user requests on independent, atom-

[1] Includes capital expenditure and finance leases. Sources: company filings and guidance (see Table A.9). Total hyperscaler AI spending is projected at $600–610B for 2026 alone (IEEE ComSoc, CNBC February 2026). The broader total AI accelerator market reached $140B in 2025 (Bloomberg Intelligence).

izable tasks). The endogenous decentralization mechanism applies directly and powerfully to inference, which already constitutes 80–90% of AI compute cycles. Training may remain permanently centralized—not because learning curves fail to reduce its costs, but because the synchronization and bandwidth requirements are architectural constraints that cost reduction alone cannot address. The post-crossing equilibrium is partial decentralization: inference distributes while training persists centrally.

Second, the effective crossing threshold is being approached from two directions simultaneously. From below, the packaging learning curve reduces the cost of delivering memory bandwidth to inference workloads along the trajectory this paper models ($\alpha = 0.23$). From above, algorithmic efficiency gains—mixture-of-experts architectures, aggressive quantization, and distillation—reduce the effective hardware requirement for a given inference capability level. These software-side gains are driven primarily by open-weight model developers operating under binding compute constraints: US semiconductor export controls deny these firms access to frontier datacenter GPUs, creating a structural incentive to maximize inference capability per unit of available hardware. The result is a dual convergence in which cumulative packaging production $Q(t)$ rises toward the crossing threshold while the threshold itself $\bar{Q}_{\text{eff}}(t)$ falls.

The contribution is six-fold. First, the formal mechanism: a continuous-time differential game with exact closed-form solutions. Second, a generalized crossing condition: $R_0 > 1$. Third, a two-dimensional crossing result: when centralized and distributed paradigms share a rivalrous input, Nash overinvestment creates supply denial whose duration depends on ASI belief persistence. Fourth, the training-inference bifurcation. Fifth, dual-convergence empirical evidence. Sixth, nine falsifiable predictions with timing. The paper is organized as follows. Section 1.1 situates this chapter within the thesis framework. Section 2 develops the mechanism. Section 3 presents the formal model. Section 4 establishes the training-inference structural distinction. Section 5 presents the empirical evidence. Section 6 validates parameter consistency across historical transitions. Section 7 offers predictions. Section 8 concludes.

A.1.1 Relation to the Thesis Framework

This chapter instantiates Level 1 (Hardware, slowest timescale) of the four-level hierarchy developed in Chapter 3 [20]. The state variable is semiconductor cost, governed by Wright's Law with learning-curve exponent $\alpha \approx 0.23$, and the timescale is decades. The crossing condition $R_0 > 1$ derived here is a special case of the spectral activation threshold from Chapter 3 (Theorem 4.3): when the hardware level's reproduction number exceeds unity,

the endogenous decentralization mechanism becomes self-sustaining. *By the CES Triple Role* [19] (Theorem 7.1), *the curvature parameter* $K = (1-\rho)(J-1)/J$ *simultaneously controls superadditivity, correlation robustness, and strategic independence*—properties that govern the complementarity among heterogeneous hardware technologies (DRAM, HBM, logic chips, specialized accelerators) whose diverse cost trajectories jointly determine the crossing threshold.

The overinvestment result (Proposition 1) provides the economic mechanism that drives Level 1: competing centralized firms in Markov Perfect Equilibrium produce aggregate output exceeding the cooperative optimum, accelerating the crossing time T^* by approximately 79%. This acceleration feeds into Level 2 (Chapter 5), where crossing triggers the first-order phase transition to a mesh economy. The hierarchical ceiling (Chapter 3, Proposition 8.1) implies that all faster levels—mesh formation, autocatalytic training, settlement dynamics—are ultimately bounded by the hardware learning rate established here. The self-undermining investment property ($\partial T^*/\partial I < 0$) is thus not merely a curiosity of the semiconductor market but the fundamental driver of the entire four-level cascade.

A.2 The Endogenous Decentralization Mechanism

A.2.1 Three-Stage Structure

Stage 1: Centralized Investment. Firms with market power invest $I(t)$ in centralized infrastructure to capture scale economies, producing cumulative component production $Q(t)$.

Stage 2: Component Cost Decline. Cumulative production drives unit costs along Wright's [24] learning curve:

$$c(Q) = c_0 \cdot Q^{-\alpha} \tag{A.1}$$

where α is the learning elasticity. The critical property is that α is a *technology* parameter, not a *firm* parameter: learning embodied in manufacturing process improvements transfers across applications. A crucial refinement: for mature technologies (such as planar DRAM die fabrication), cumulative production is sufficiently large that marginal cost reductions per doubling are negligible. The mechanism's force depends on *new* production processes—specifically, 3D memory stacking and advanced packaging—that are in their early high-α phase. The packaging techniques developed for datacenter HBM (TSV interconnects, hybrid bonding, die thinning, thermal management of stacked dies) transfer directly to consumer memory form factors.

Stage 3: Architectural Recombination. When component costs cross a threshold c^*, the same components can be recombined into distributed architectures exhibiting network

externalities. Beyond a crossing time T^*, the distributed paradigm dominates for workloads amenable to distributed execution.

A.2.2 The Self-Undermining Investment Property

The mechanism's distinctive feature is that each stage causally enables the next, and the final stage undermines the first. Define T^* as the first date at which distributed architecture cost-performance matches centralized provision for the marginal inference user. Then:

$$\frac{\partial T^*}{\partial I} < 0 \tag{A.2}$$

Increased centralized investment accelerates displacement of the centralized paradigm's inference revenue.

A.2.3 Dual Convergence

The current AI transition exhibits a feature absent from prior technological transitions: the effective crossing threshold is being approached from two directions simultaneously.

From below: the packaging learning curve. The cost of delivering memory bandwidth to inference workloads is driven by 3D stacking and advanced packaging, not by planar DRAM die fabrication. The die cost—historically the dominant component—is near-asymptotic: DRAM is among the highest-cumulative-volume semiconductor products ever manufactured. The packaging cost, by contrast, is in its early high-learning-rate phase: volume production of TSV-based stacked memory began circa 2015, and the learning curve ($\alpha = 0.23$ from HBM product-level data) is consistent with early-stage technologies across domains.

From above: algorithmic efficiency gains. Advances in model architecture and compression reduce the hardware *required* to achieve a given inference capability level. Mixture-of-experts (MoE) architectures activate only a fraction of total parameters per token, reducing effective memory bandwidth requirements by 3–6×. Quantization (INT4, INT2) reduces model memory footprint by 4–16×. Distillation transfers capability from large models to smaller ones.

Define $\bar{Q}_{\text{eff}}(t) = \bar{Q} \cdot f(\eta(t))$, where $\eta(t)$ indexes cumulative algorithmic efficiency gains and f is decreasing. The state variable becomes $x(t) = \bar{Q}_{\text{eff}}(\eta(t)) - Q(t)$, and the rate of depletion exceeds what hardware learning curves alone would predict.

A.2.4 Distinction from Adjacent Theory

Table A.1 summarizes the positioning. The distinctions are precise: Arrow's [2] learning-by-doing benefits the same paradigm; Bresnahan and Trajtenberg's [6] GPT spillovers enable applications across sectors rather than architectural self-replacement; Schumpeter's [18] creative destruction comes from external entrants.

Table A.1: Theoretical positioning of endogenous decentralization.

Framework	Learning Scope	Beneficiary	Disruption Source	Self-Undermining?
Arrow [2]	Same paradigm	Same firms	N/A	No
Bresnahan-Trajtenberg [6]	Cross-sector	Other sectors	External applications	No
Schumpeter [18]	External	Entrant firms	External entrant	No
Christensen [8]	Cross-market	Entrant firms	New value network	Partial
This paper	**Cross-paradigm**	**Different architecture**	**Self-financed**	**Yes**

A.3 Formal Model

A.3.1 Environment

Consider $N \geq 2$ symmetric centralized firms indexed by $i \in \{1, \ldots, N\}$. Time is continuous. The *state variable* is $x(t) = \bar{Q}_{\text{eff}} - Q(t) \in [0, x_0]$, measuring the remaining cumulative production until the effective crossing threshold at which distributed architecture becomes cost-competitive for inference workloads. When x reaches zero, inference crossing occurs. The state evolves as:

$$dx/dt = -\sum_i q_i(t) \tag{A.3}$$

where $q_i(t) \geq 0$ is firm i's output rate. Each unit of output serves the centralized market and simultaneously depletes the remaining distance to crossing—this dual role is the formal expression of the self-undermining investment property.

Flow profits for firm i are determined by linear inverse demand $P = a - bQ$, where $Q = \sum q_j$ is total output rate:

$$\pi_i(t) = (a - bQ)q_i \tag{A.4}$$

with $a > 0$, $b > 0$. Upon crossing ($x = 0$), each firm receives continuation value:

$$S = S_T + \frac{S_I}{N(r+\delta)} \tag{A.5}$$

where S_T represents the persistent training and model-licensing revenue that survives inference decentralization, $S_I = \bar{\pi}_I$ is the pre-crossing inference profit level, r is the discount rate, and $\delta > 0$ is the post-crossing inference displacement rate.

Remark on the objective function. The model assumes firms maximize discounted revenue from infrastructure services. An alternative specification treats centralized investment as purchasing an option on a discontinuous payoff: the first firm to achieve a capability threshold captures a prize V^* that dwarfs cumulative investment. Under this specification, S_T is the option value of retaining frontier training *capability*. The mechanism's core result ($\partial T^*/\partial I < 0$) is invariant to the firms' objective. The revenue-maximization model provides a *lower bound* on aggregate investment and a correspondingly conservative estimate of T^*. Section 3.7 calibrates both specifications. The empirical capex data (Section 5.5) confirm this prediction: the pre-AI overinvestment ratio matches the revenue-maximization model (3–4×), while the post-2022 ratio (11–19×) matches the option-value specification when the prize includes a superintelligence option (Remark in Section 5.5).

The game has a *common-pool* structure: the state x is a shared resource (remaining time before inference disruption) that all firms deplete through production. This structure is analogous to the fishery or oil extraction commons (Levhari and Mirman [16]), with the critical distinction that the "resource" being depleted is the incumbent paradigm's remaining inference viability.

A.3.2 Markov Perfect Equilibrium

I restrict attention to symmetric stationary Markov strategies $q_i = q(x)$. Each firm's value function $V(x)$ satisfies the Hamilton-Jacobi-Bellman equation:

$$rV(x) = \max_{q_i}\{(a - b(q_i + (N-1)q(x)))q_i - V'(x) \cdot (q_i + (N-1)q(x))\} \tag{A.6}$$

The first-order condition under symmetry yields the equilibrium strategy:

$$q^N(x) = \frac{a - V^{N\prime}(x)}{b(N+1)} \tag{A.7}$$

Substituting back into the HJB yields the ODE:

$$rV^N(x) = \frac{(a - V^{N\prime}(x))(a - N^2 V^{N\prime}(x))}{b(N+1)^2} \tag{ODE-N}$$

with boundary condition $V^N(0) = S$.

A.3.3 Cooperative Benchmark

The cooperative planner maximizes total producer surplus $W(x) = NV^P(x)$, choosing total output rate Q:

$$rV^P(x) = \frac{(a - NV^{P\prime}(x))^2}{4bN} \tag{ODE-C}$$

with boundary condition $V^P(0) = S$.

A.3.4 Analytical Solutions

Both ODEs are autonomous and separable. The cooperative ODE yields the exact implicit solution:

$$x(V) = \frac{a \cdot \ln\left(\frac{a - 2\sqrt{bnrS}}{a - 2\sqrt{bnrV}}\right) + 2\left(\sqrt{bnrS} - \sqrt{bnrV}\right)}{2br} \tag{C-exact}$$

The Nash ODE is solved by the substitution $u = \sqrt{D + EV}$:

$$x(V) = \frac{4N^2}{E}\left[(u_0 - u) + A \cdot \ln\left(\frac{A - u_0}{A - u}\right)\right] \tag{N-exact}$$

Both solutions share the same functional form—$\sqrt{\cdot} + \log$—differing only in the constants governing shadow cost internalization. Both are verified to machine precision ($\max|x_{\text{exact}} - x_{\text{num}}| < 10^{-12}$).

A.3.5 The Overinvestment Result

Proposition A.3.1 (Overinvestment in Markov Perfect Equilibrium)**.** *In the symmetric MPE, aggregate output $Q^N(x) = Nq^N(x)$ strictly exceeds cooperative output $Q^C(x)$ for all $x > 0$. Consequently, $T^{*,Nash} < T^{*,Coop}$: Nash equilibrium crossing occurs strictly earlier than the cooperative optimum.*

Proof. Step 1. At $x = 0$, $V^N(0) = V^P(0) = S$. Evaluating the boundary derivatives from (ODE-N) and (ODE-C), the planner's total shadow cost $N\mu$ strictly exceeds the Nash firm's

private shadow cost λ for $N \geq 2$. This gap reflects the learning externality: each Nash firm internalizes only its own future profit loss from approaching crossing.

Step 2. By a standard comparison theorem for ODEs (Walter [23], Theorem I.9.1), the ordering $N \cdot V^{P\prime}(x) > V^{N\prime}(x)$ propagates to all $x > 0$.

Step 3. From the output expressions, both the smaller numerator (higher shadow cost) and larger denominator of Q^C relative to Q^N ensure $Q^N(x) > Q^C(x)$ for all $x > 0$. □

Remark A.3.2 (Irreversibility). At $Q = \bar{Q}$, a new basin of attraction—the distributed inference equilibrium—becomes accessible. Reversing the crossing would require cumulative production to decrease—which contradicts monotonicity. Once Q crosses $\bar{Q}$, the inference transition is topologically irreversible.

Remark A.3.3 (Niche Persistence). Irreversibility of inference crossing does not imply extinction of the centralized paradigm. IBM's mainframe business continues to generate approximately $3–4 billion annually as of 2025—decades after the PC revolution—serving high-reliability transaction processing.

Economic interpretation. The overinvestment decomposes into a Cournot channel (price-depressing rival output) and a learning externality channel (private shadow cost = $1/N$ of social shadow cost). The decomposition of S into $S_T + S_I/(N(r+\delta))$ reveals a moderating effect: when S_T is large, crossing is less catastrophic and the overinvestment gap narrows.

Welfare loss. At baseline calibration ($N = 5$, $S_T = 0$), the per-firm welfare loss under Nash competition is 34.1%. With S_T calibrated to estimated training revenue persistence, the loss moderates to approximately 22–28%.

A.3.6 Comparative Statics

Corollary A.3.4 (Increasing N)**.** *Nash equilibrium aggregate output is strictly increasing in N for all $x > 0$.*

Corollary A.3.5 (Asymmetric firms)**.** *If firm 1 has marginal cost $c_1 - \varepsilon$, aggregate equilibrium output is strictly increasing in ε.*

Corollary A.3.6 (Asymmetric crossing valuation)**.** *If firm j has post-crossing value $S_j > S$, firm j produces strictly more than symmetric competitors, and aggregate output increases.*

Corollary A.3.7 (Capacity constraint and boom-bust)**.** *Crossing time delay is bounded by the construction lag Δ for new capacity. The long-run packaging learning rate α is unaffected.*

The 2025–26 DRAM supercycle provides a real-time test: consumer DDR5 prices have risen 250–400% above trend, driven by AI datacenter demand reallocating wafer capacity

to HBM formats. If the deviation is merely cyclical (bounded by the construction lag Δ), the learning rate $\alpha = 0.23$ is unaffected because the supercycle is a demand allocation shock, not a change in the stacking production function. But the observed severity—outright supply denial rather than mere price inflation—motivates a stronger result. Proposition A.3.9 formalizes the conditions under which the deviation persists beyond Δ.

Remark A.3.8 (Option-value amplification). Under the option-value objective function specification (Section 3.1), the overinvestment result is amplified. If firms invest to maximize the probability of achieving a discontinuous capability threshold, the marginal value of additional investment is governed by the prize V^* rather than by discounted market revenue. The model's quantitative predictions ($Q^N/Q^C \approx 3$–$4\times$, $T^* \approx 2028$) are then conservative. The empirical capex data (Section 5.5) confirm this two-regime structure: pre-2022 ratios match the revenue-maximization model while post-2022 ratios match the option-value specification with V^* calibrated to a superintelligence option.

A.3.7 Input Cannibalization

Corollary 4 treats the capacity constraint as a temporary boom-bust deviation bounded by the construction lag Δ for new packaging capacity. This section formalizes a stronger result: when centralized and distributed paradigms share a *rivalrous input*, Nash overinvestment creates a supply-denial channel that operates independently of the packaging learning curve and whose duration depends on the persistence of ASI belief.

Proposition A.3.9 (Input cannibalization and two-dimensional crossing)**.** *Let centralized and distributed paradigms share a rivalrous input with total capacity K, and let the centralized paradigm consume $\theta > 1$ units of input capacity per unit of output relative to the distributed paradigm. Define residual consumer capacity $K_C \equiv K - \theta D_H(N)$, where $D_H(N)$ is aggregate centralized input demand.*

(i) Two-dimensional crossing. *Self-sustaining distributed adoption requires both cost parity on the packaging learning curve* and *sufficient consumer input supply. The effective crossing threshold generalizes equation* (A.18)*:*

$$\bar{Q}^{**} = \bar{Q}^*\big(\kappa(K_C)\big) \cdot \mathbf{1}\{K_C \geq K_{\min}\} \tag{A.8}$$

where $\kappa(K_C)$ is coordination friction as a decreasing function of consumer input availability (scarcer memory $\Rightarrow$ higher κ $\Rightarrow$ larger $\bar{Q}^*$), and $K_{\min}$ is the minimum capacity for viable consumer production. When $K_C < K_{\min}$, the threshold is unreachable regardless of learning-curve progress.

(ii) Non-monotonic crossing dynamics. *Centralized investment simultaneously advances the packaging learning curve and depletes consumer input supply. The crossing distance decomposes:*

$$\frac{dx}{dt} = \underbrace{\frac{d\bar{Q}^{**}}{dK_C} \cdot \frac{dK_C}{dt}}_{\textit{supply-denial channel (anti-crossing)}} - \underbrace{\sum_i q_i}_{\textit{learning-curve channel (pro-crossing)}} \tag{A.9}$$

In the boom phase, the anti-crossing channel dominates: consumer devices lose memory capacity even as packaging costs fall. In the bust phase, K_C recovers and the accumulated learning progress produces a crossing from a more advanced position on the cost curve than a monotone model would predict.

(iii) Duration under ASI belief. *Under the option-value specification, centralized demand scales with $M_{\text{eff}} = M + p \cdot V_{\text{ASI}}$. The supply-denial duration is*

$$\Delta_{\text{IC}} = \inf\{t \geq 0 : K(t) \geq \theta D_H(N, M_{\text{eff}}(t)) + K_{\min}\} \tag{A.10}$$

When $M_{\text{eff}}(t)$ is non-decreasing—i.e., ASI belief is sustained or strengthened by capability demonstrations—Δ_{IC} can substantially exceed the fab construction lag Δ_K. Supply denial resolves only when capacity growth outpaces demand growth: $dK/dt > \theta \cdot dD_H/dt$.

Proof. (i) An edge inference device requires a minimum physical memory endowment (currently ≥8GB for even a 3B on-device model). When $K_C < K_{\min}$, consumer devices either cannot be produced at volume or are produced with insufficient memory for on-device inference. The cost-parity condition $c(Q) \leq c^*$ is necessary but not sufficient: the distributed paradigm requires *available memory*, not merely *affordable packaging*.

(ii) Each HBM unit absorbs θ units of consumer wafer capacity but contributes θ units of cumulative packaging production Q. The learning benefit accrues to the *stock* $Q(t)$, which is monotone non-decreasing; the supply denial operates on the *flow* $K_C(t)$, which reverses when capacity expands. This stock-flow asymmetry ensures the learning benefit is permanent while the supply constraint is temporary—resolving the ambiguity in favor of long-run acceleration.

(iii) New fab capacity requires $\Delta_K \approx 3$–5 years from groundbreaking to volume production. During $[0, \Delta_K]$, K is approximately fixed while D_H may grow if M_{eff} increases. When new capacity arrives, $D_H(t + \Delta_K)$ may exceed $D_H(t)$ by enough to absorb the expansion—a moving target. The constraint persists until capacity growth dK/dt exceeds demand growth $\theta \cdot dD_H/dt$, which requires either demand stabilization ($dp/dt \leq 0$, ASI belief ceasing to grow) or capacity expansion exceeding the historical DRAM industry rate of 10–15% per year. □

Calibration to the DRAM market. The input is DRAM wafer capacity, controlled by three firms (Samsung 33%, SK Hynix 34%, Micron 26%). The wafer multiplier is $\theta \approx 3$–4: HBM production consumes 3–4 times the wafer capacity of standard DRAM per gigabyte due to TSV die stacking, die thinning, and lower yields (Tom's Hardware 2025; TrendForce 2025). HBM profit margins are 5–10× consumer DRAM—SK Hynix reports HBM accounting for 40% of total DRAM revenue from approximately 10–12% of wafer output. At a margin-to-wafer ratio $\theta_\pi/\theta \approx 1.5$–3, profit maximization implies maximal HBM allocation until demand is exhausted, which is the behavior observed: Micron exited the consumer market entirely (Crucial brand discontinued, February 2026), and all three manufacturers halted DDR4 orders simultaneously. The condition $K_C < K_{\min}$ was crossed in late 2025. Section 5.5.2 documents the empirical evidence.

The critical parameter is duration (part iii). Corollary 4's estimate of 1–2 years of crossing delay corresponds to the lower bound Δ_K when ASI belief $p \to 0$ and centralized demand moderates on schedule. The upper bound depends on the trajectory of $M_{\text{eff}}(t)$. New fabs (SK Hynix Yongin, Samsung P5, Micron Boise and Hiroshima) reach volume production in 2027–2028, but if demonstrated capability advances—the reasoning breakthroughs of 2024–25, agentic AI in 2026—sustain or increase p, demand growth absorbs new capacity as fast as it arrives. Under this scenario, the supply-denial window extends to $\Delta_{\text{IC}} \approx 5$–10 years: the duration of the ASI investment episode itself. The pre-2022 data (overinvestment ratios 3–4×) correspond to the $p \approx 0$ regime; the post-2022 data (11–19×) correspond to $p > 0$. As long as the market remains in the second regime, Corollary 4's optimistic bound does not apply—Proposition A.3.9 governs instead.

A.3.8 Calibration

The learning elasticity $\alpha = 0.23$ is estimated from the HBM packaging learning curve (Table A.8), which measures the cost trajectory of 3D-stacked memory from first volume production (HBM1, 2015) through the current generation (HBM3E+, 2025). This estimate captures the relevant production process—through-silicon via (TSV) interconnects, die thinning, hybrid bonding, and thermal management—rather than the mature planar DRAM die (see Section 5.2 for the cost decomposition). Current HBM cost is approximately \$12/GB (HBM3E, 2025); the crossing threshold is \$5–7/GB. The calibration uses the conservative bound $\bar{Q} \approx 112$ EB (\$5/GB target).

Sensitivity of T^* to α. The model's timing predictions are sensitive to the learning elasticity. Table A.2 reports T^* across the range of estimates in the literature, holding other parameters at baseline.

Table A.2: Sensitivity of crossing time to learning elasticity.

α	Source / Label	T^* (yrs from 2024)	Calendar Year
0.12	Goldberg et al. [12] w/ spillovers	93	2117
0.15	Conservative lower bound	74	2098
0.20	Irwin & Klenow [15] canonical IV	56	2080
0.23	HBM packaging curve (baseline)	47	2071
0.25	Upper Irwin & Klenow range	45	2069
0.32	Irwin & Klenow OLS (likely biased up)	35	2059

Notes: T^* computed from hardware learning curve only, without algorithmic efficiency gains. Dual convergence (Section 5.3) shifts all dates earlier.

Post-crossing continuation value. The inference displacement rate $\delta \approx 0.30$ from the IBM trajectory (Section 6.1). Under revenue-maximization: S_T high (closed-model dominance), welfare loss ∼22%; S_T moderate (open-weight competition), ∼28%; $S_T \approx 0$ (commoditization), ∼34%. Under the option-value specification, S_T represents the option value of maintaining frontier training capability at scales no distributed architecture can replicate. The two specifications bracket the range of outcomes.

Quantitative predictions. Under Nash competition with $N = 5$, crossing at approximately 2028. The 2025–26 DRAM supercycle delays the cost threshold by an estimated 1–2 years during the boom phase, with potential acceleration during the subsequent bust. Under cooperation, ∼2042. Competition accelerates by 79%.

A.3.9 Note on Identification

The packaging learning curve is estimated by OLS regression of log cost on log cumulative output for HBM generations (Table A.8). This identifies a correlation, not necessarily a structural learning-by-doing parameter. Endogeneity concerns (demand shocks driving both output and investment in cost reduction) are standard in the learning-curve literature (Irwin and Klenow [15]).

No published IV estimate exists for the packaging learning curve. The $\alpha = 0.23$ is identified from product-level HBM pricing that bundles die and packaging costs, with $n = 6$ generation-level observations—too few for formal structural estimation. This paper's empirical contribution is identifying *which* curve matters (early-stage packaging, not asymptotic die fabrication), not claiming precise estimation of its slope. The estimate's reliability rests on three indirect supports: cross-technology consistency of $\alpha \approx 0.21$–0.24 across independently estimated early-stage curves (Table A.12); the physical cost decomposition showing packaging as the majority cost component at current HBM price points (Table A.7);

and the early-stage character of the process, where limited demand-side feedback reduces simultaneous-equations bias relative to the 41-year DRAM die series. The self-undermining property ($\partial T^*/\partial I < 0$) requires only that centralized investment contributes to cumulative Q and that $c(Q)$ is decreasing and stable. Refining the packaging α with firm-level production data as it accumulates is a natural next step.

Irwin and Klenow [15] provide the most rigorous causal estimate for semiconductor learning: $\alpha = 0.32$ (SE $= 0.05$) using instrumental variables on a firm-level DRAM panel (1974–1992). Goldberg et al. [12] estimate learning rates at the firm-technology-node level for microprocessor fabrication, finding $\alpha = 0.05$ at the firm-node level, rising to $\alpha = 0.12$ when cross-border spillovers are included. The model's $\alpha = 0.23$ is thus an *industry-level spillover-inclusive* estimate, consistent with the Goldberg et al. framework when cross-application spillovers are the dominant channel.

A.3.10 Generalized Crossing Condition

The model defines crossing at cost parity, but empirical evidence shows hardware crossing *precedes* architectural dominance by 3–5 years (Section 6.4). Cost parity is necessary but not sufficient: the distributed ecosystem must also overcome coordination frictions, sustain adoption against churn, and generate network effects that make the transition self-reinforcing. What is actually required is that the distributed ecosystem's basic reproduction number exceeds unity. *By the spectral activation threshold* (Chapter 3, Theorem 4.3), *the hardware level's nontrivial equilibrium exists if and only if the spectral radius of the next-generation matrix exceeds unity.* The $R_0 > 1$ condition derived below is the Level 1 specialization of this general result.

Why Epidemic Dynamics?

Three canonical frameworks model technology adoption: Bass [3] diffusion, threshold models (Granovetter [14]; Schelling [25]), and epidemic/SIR models (applied to technology diffusion by Mansfield [17]). The choice among them is not arbitrary—each embeds different assumptions about the adoption mechanism.

Bass diffusion decomposes adoption into an external "innovation" rate p and an internal "imitation" rate q, taking the product's existence and characteristics as given. This is a *demand-side* model: it asks how fast a fixed product diffuses through a population. For the inference decentralization mechanism, the product's viability is itself endogenous to adoption through the learning curve—the distributed alternative does not exist as a competitive option until cumulative production crosses a cost threshold. Bass assumes the innovation is available

from $t = 0$; here, $t = 0$ is what we are trying to determine.

Threshold models (Granovetter [14]) assign each potential adopter a switching threshold and characterize cascade conditions. These are powerful for analyzing tipping points but are fundamentally *static*: they characterize *whether* a cascade occurs given a distribution of thresholds, but do not naturally incorporate the feedback loop in which each adoption reduces cost for subsequent adopters through learning-by-doing.

The *epidemic/SIR framework* captures the structural feature that distinguishes this transition: the adoption rate β is endogenous to cumulative output through the learning curve. In the standard SIR model, β is fixed. Here, β is a function of cost $c(Q)$, which falls with cumulative production Q, which is itself driven by adoption. This positive feedback—adoption $\rightarrow$ cumulative production $\rightarrow$ cost decline $\rightarrow$ higher adoption rate—means R_0 is a *rising function of the state variable*, and the crossing event occurs when R_0 passes through unity from below. This dynamic endogeneity is absent from both Bass and threshold specifications in their standard forms.

The frameworks are related. Bemmaor [4] showed that Bass diffusion is a special case of a heterogeneous-hazard epidemic model; threshold models can be reformulated as SIR dynamics with heterogeneous β (Dodds and Watts [10]). The epidemic framing thus nests the alternatives as restrictions. The generalization matters because the Bass restriction—fixed innovation and imitation rates throughout diffusion—rules out precisely the supply-side feedback that drives the mechanism.

Formal Specification

Let $s(t) \in [0, 1]$ denote the share of inference workloads served by distributed architecture. Adoption dynamics follow:

$$ds/dt = \beta(c(Q), \lambda) \cdot \gamma \cdot s(t) \cdot (1 - s(t)) - (\kappa + \mu) \cdot s(t) \tag{A.11}$$

The first term captures contagion-like growth: each unit of distributed share generates new adoption at rate $\beta\gamma$, modulated by the remaining adoptable share $(1 - s)$. The second term captures outflows from coordination friction κ and churn μ. The ecosystem is self-sustaining ($ds/dt > 0$ for small s) when the basic reproduction number exceeds unity:

$$R_0 \equiv \frac{\beta(c, \lambda) \cdot \gamma}{\kappa + \mu} > 1 \tag{A.12}$$

The parameters have the following structural interpretations:

- $\beta(c, \lambda)$: *Adoption rate*, depending on the cost advantage and latency advantage. Mi-

crofounded below.

- γ: *Network effect multiplier*, capturing the degree to which each adopter increases ecosystem value for subsequent adopters through shared model repositories, tooling, and deployment infrastructure.
- κ: *Coordination friction*, the rate at which potential adopters are deterred by deployment complexity and hardware heterogeneity. Observable from deployment latency compression: weeks in mid-2024, hours by January 2025 (Section 5.4.3).
- μ: *Churn rate*, driven by model obsolescence and capability gaps. Bounded from model lifecycle data: $\mu \approx 0.08$–0.17/month (Section 5.4.3).
- λ: *Latency advantage*, a structural, hardware-determined quality dimension: edge inference achieves <10ms response versus 50–200ms for cloud round-trip, independent of cost dynamics.

Microfoundation for $\beta(c, \lambda)$

The adoption rate depends on the cost saving from switching and the latency improvement. Specify:

$$\beta(c, \lambda) = \beta_0 \cdot (c^* - c(Q))^+ + \lambda \tag{A.13}$$

where c^* is the centralized cost benchmark, $c(Q) = c_0 \cdot Q^{-\alpha}$ is the distributed cost at cumulative production Q, and $(\cdot)^+ = \max(\cdot, 0)$. The parameter β_0 converts per-unit cost savings into an adoption rate; λ provides a floor from the latency advantage alone, operating even before cost parity.

At cost parity ($Q = \bar{Q}$, where $c(\bar{Q}) = c^*$), the cost-savings term vanishes:

$$R_0\big|_{Q=\bar{Q}} = \frac{\lambda\gamma}{\kappa + \mu} \tag{A.14}$$

This determines whether hardware crossing is sufficient for self-sustaining adoption:

- If $\lambda\gamma > \kappa + \mu$: the latency advantage alone drives $R_0 > 1$ at cost parity. The ecosystem becomes self-sustaining immediately. Coordination lag $\Delta T \approx 0$.
- If $\lambda\gamma < \kappa + \mu$: additional cumulative production beyond $\bar{Q}$ is required to push cost below parity, generating a positive cost-savings term. This produces the 2–5 year coordination lag observed historically (Table A.3).

Derivation of $\bar{Q}^*$

The self-sustaining adoption threshold $\bar{Q}^*$ is the cumulative production level at which $R_0 = 1$. Setting $R_0 = 1$:

$$\frac{[\beta_0(c^* - c(Q)) + \lambda] \cdot \gamma}{\kappa + \mu} = 1 \tag{A.15}$$

Solving for $c(Q)$:

$$\begin{aligned} \beta_0(c^* - c(Q)) + \lambda &= \frac{\kappa + \mu}{\gamma} \\ c(Q) &= c^* - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right) \end{aligned} \tag{A.16}$$

Substituting the learning curve $c(Q) = c_0 Q^{-\alpha}$ and $c^* = c_0 \bar{Q}^{-\alpha}$:

$$\begin{aligned} c_0 Q^{-\alpha} &= c_0 \bar{Q}^{-\alpha} - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right) \\ Q^{-\alpha} &= \bar{Q}^{-\alpha}\left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right) \end{aligned} \tag{A.17}$$

Taking the $(-1/\alpha)$ power:

$$\boxed{\bar{Q}^* = \bar{Q} \cdot \left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right)^{-1/\alpha}} \tag{A.18}$$

Three properties merit emphasis.

Direction of the shift. When $\lambda\gamma < \kappa + \mu$ (the empirically relevant case—Section 5.4 estimates $R_{0,\text{distributed}} \approx 0.4$–$0.8$ currently), the parenthetical term is less than unity, so $\bar{Q}^* > \bar{Q}$: self-sustaining adoption requires more cumulative production than cost parity. The gap $\bar{Q}^* - \bar{Q}$ is the formal expression of the coordination layer lag.

Monotonicity in κ. $\partial \bar{Q}^* / \partial \kappa > 0$: higher coordination friction delays the threshold. This is testable: if the coordination indicators in Section 5.4.3 (deployment latency compression, day-zero quantization availability) continue their trajectory, κ falls and $\bar{Q}^*$ converges toward $\bar{Q}$.

Compatibility with the differential game. Replace $\bar{Q}$ with $\bar{Q}^*(\kappa, \mu, \gamma, \lambda)$ in the state variable $x(t) = \bar{Q}^*_{\text{eff}} - Q(t)$. All propositions carry through: the overinvestment result (Proposition 1) depends on the common-pool structure, not on the threshold's specific value. The generalization shifts the *level* of T^* without altering the *comparative statics* ($\partial T^*/\partial N < 0$, $\partial T^*/\partial I < 0$). Appendix C formalizes the semi-endogenous dynamics when κ itself evolves

over time.

Connection to the CES framework. The $R_0 > 1$ condition derived here governs the activation of a single level. *By the CES Triple Role* [19] (Theorem 7.1), *the curvature parameter K controls the superadditivity premium that makes complementary hardware combinations productive, the correlation robustness that prevents correlated failures from collapsing the distributed ecosystem, and the strategic independence that ensures balanced allocation is a Nash equilibrium.* These properties enter the R_0 framework through the network effect multiplier γ (which is increasing in K, since higher curvature means greater gains from combining diverse hardware) and the coordination friction κ (which is decreasing in K, since strategic independence reduces the scope for hold-up). The cross-level amplification result from Chapter 3 [20] (Section 4.3) implies that even when this level's R_0 is sub-threshold, coupling with faster levels can activate the entire system.

Observable Implications

The R_0 framework makes a specific, testable prediction: hardware cost parity precedes self-sustaining distributed adoption by ΔT years, where ΔT depends on the gap between $\lambda\gamma$ and $(\kappa + \mu)$ at the crossing point.

Table A.3: Coordination layer lag across transitions.

Transition	Hardware T^*	R_0 T^*	ΔT
Mainframe → PC	1987	1990–92	3–5 yr
ARPANET → Internet	∼1989	1993–94	4–5 yr
Cloud → Edge AI	2027–29[†]	?	2–3 yr (pred.)

[†] Hardware capability threshold met at professional price points Q1 2026; consumer cost threshold delayed by 2025–26 DRAM supercycle (Corollary 4). The predicted compression from 3–5 years to 2–3 years reflects declining κ: coordination infrastructure (quantization pipelines, edge runtimes, model hubs) is being built *before* hardware crossing, unlike in historical transitions where the coordination layer was built after. Section 5.4 bounds the R_0 parameters empirically from OpenRouter adoption data.

A.4 The Training-Inference Structural Distinction

The $1.3 trillion in centralized AI infrastructure investment builds capacity for two structurally distinct workloads. Conflating them overstates the mechanism's scope; separating them sharpens it.

A.4.1 Two Workloads, Two Architectures

Training teaches models by processing massive datasets across tightly synchronized GPU clusters. Frontier training runs now require 100,000–500,000+ GPUs communicating at terabits per second via InfiniBand or NVLink, running continuously for weeks to months.[2] Power density: 100–1,000 kW/rack. Latency-insensitive.

Inference runs trained models to serve real-time user requests. Tasks are independent and atomizable. Latency-sensitive: users benefit from <10ms local execution versus 50–200ms cloud round-trip. Frequency: continuous, scales with every user and query.

Table A.4: Training vs. inference structural comparison.

Dimension	Training	Inference
Share of AI compute (2025)	∼50%	∼50%
Share of AI compute (2026, proj.)	∼33%	∼67%
Synchronization requirement	Massive (100K–500K+ GPUs)	None (atomizable)
Latency sensitivity	Low	High (<10ms for UX)
Cost trajectory	Rising per frontier model	Declining ∼280× in 2 yr
Edge-viable?	No (architectural)	Yes (this paper's thesis)

Sources: Deloitte (2025), McKinsey (2025), MIT Technology Review (2025), Epoch AI (2025).

A.4.2 Training Does Not Decentralize

Frontier model training requires synchronized clusters of 100,000–500,000+ GPUs communicating at terabit-per-second speeds, with the largest clusters now approaching gigawatt-scale power consumption. A consumer device with excellent memory bandwidth cannot participate in a distributed training run because the inter-device communication latency (milliseconds over WiFi vs. nanoseconds over NVLink) creates a performance gap of 5–6 orders of magnitude. No plausible learning curve closes this gap because the constraint is topological (network diameter and synchronization protocol) rather than cost-based. Indeed, the trend is toward *larger* synchronized clusters, not smaller ones: xAI targets 1–2 million GPUs by late 2026, and the Stargate project is designed for 400,000 GB200s expandable across multiple sites toward 10 GW total capacity.

[2]As of Q1 2026, the largest known training clusters include xAI Colossus (555,000 GPUs, ∼1.4M H100-equivalents), Microsoft Fairwater (2.5M+ H100-equivalents at full scale), Meta Prometheus (∼500,000 GPUs, 1 GW), and Amazon/Anthropic Project Rainier (∼500,000 Trainium2 chips). Google's TPU fleet totals an estimated 3.5–4.2M H100-equivalents. Five facilities exceeding 1 GW are coming online in 2026.

A.4.3 Inference Decentralizes

Inference tasks are *atomizable*: each user query is independent. Inference is *latency-advantaged*: local execution outperforms cloud round-trip on a quality dimension independent of cost. Inference is *bandwidth-bound*: token generation speed is determined almost entirely by the ratio of memory bandwidth to model size—exactly the constraint whose packaging learning curve the model tracks. Inference *scales with users*.

The potential edge inference install base is large but almost entirely unused for generative AI. In 2025, approximately 370 million smartphones shipped with NPU hardware (32% of the global smartphone market; Gartner 2025), but the only widely adopted on-device AI application is computational photography—image enhancement, noise reduction, and object removal in the camera pipeline. Voice-activated personal agents (Siri, Google Assistant, Alexa) are improving rapidly as LLM backends replace older intent-classification systems, but they reinforce rather than challenge the cloud inference model: the generative processing occurs on datacenter GPUs, with the phone serving as a thin client for audio capture and playback. The models that fit entirely in phone memory ($\leq$3B parameters) are not competitive with these cloud-hosted alternatives, and users have no compelling reason to run them locally. PC vendors shipped 78 million "AI PCs" with dedicated NPUs at 40–50 TOPS (31% of the PC market; Counterpoint 2025), but this is a marketing category in search of a use case: no mainstream software exploits the NPU for generative inference, and the DRAM supercycle has inflated PC prices, suppressing adoption of the higher-memory configurations ($\geq$32GB) that would be needed. Apple M-series Macs with 32–128GB unified memory are the most capable current edge inference platforms—the only consumer hardware that can run 30–70B parameter models at interactive speeds—but only newer desktop and high-end laptop configurations have sufficient RAM, representing a small fraction of Apple's 2.5 billion active devices. The gap between hardware potential and actual use is precisely the coordination-layer deficit that the R_0 framework models (Section 5.4): NPU-equipped devices exist in volume, but $\kappa_{\text{distributed}}$ remains high because there is no killer application driving adoption of on-device generative inference, the software ecosystem has not crystallized, and memory constraints confine most devices to sub-7B models that cannot compete with cloud alternatives.[3]

[3]A revealing case: OpenClaw, an open-source personal AI agent framework, accumulated 200,000 GitHub stars and 720,000 weekly downloads within weeks of its January 2026 release—then was immediately acquihired by OpenAI (February 2026). OpenClaw runs a local orchestration layer but routes all generative inference to cloud APIs (Anthropic, OpenAI). The pattern illustrates two dynamics simultaneously: explosive demand for personal AI agents is building the coordination layer that would need to exist before local models could substitute for cloud APIs, and centralized incumbents are absorbing that coordination infrastructure before it can enable the crossing.

A.4.4 The Inference Revenue Pool

Inference dominates both compute cycles (80–90%) and ongoing revenue. The inference market is projected to grow from $106 billion (2025) to $255 billion by 2030 (MarketsandMarkets 2025); the market for inference-optimized chips alone exceeded $20 billion in 2025 and is projected above $50 billion in 2026 (Deloitte 2026). ChatGPT alone processes over 2 billion prompts daily, and inference cost deflation—10× per year at median, with post-January 2024 rates accelerating to 50–200× per year (Epoch AI 2025)—is driving a Jevons paradox in which falling costs generate exponentially more queries. The emerging shift from conversational chatbots to autonomous AI agents amplifies this further: an agent completing a task may invoke 50–200 model calls (reasoning, tool use, code execution, verification) per user interaction, multiplying per-session inference demand by one to two orders of magnitude relative to a single chatbot query.[4] This is the revenue pool that the $2.4 trillion infrastructure buildout targets, and the revenue pool that edge devices will intercept.

A.4.5 Implications for the Model

A.5 Empirical Evidence: Dual Convergence

The inference crossing condition—≥70B-class output quality at ≥20 tok/s under $1,500—is being approached from two directions: hardware costs declining from below (Section 5.2) and algorithmic efficiency reducing the effective threshold from above (Section 5.3). Before presenting each channel, Section 5.1 exploits the US semiconductor export controls as a natural experiment that distinguishes the endogenous decentralization mechanism from standard learning-by-doing.

A.5.1 Identification: The Export-Control Natural Experiment

A referee's natural objection is: *How do I know this isn't just Arrow [2] with a longer supply chain?* Standard learning-by-doing predicts that firms with *more* cumulative production learn faster. The export controls created a natural experiment: a subset of AI developers were *denied* access to frontier compute. Under Arrow, these firms should fall behind. Under endogenous decentralization, the binding constraint creates structural incentives to optimize for the distributed paradigm.

[4]The acquisition race for agent frameworks—OpenAI acquiring OpenClaw (February 2026), Meta acquiring Manus AI and Limitless AI—reflects the incumbents' recognition that agentic workloads, not chatbot conversations, will dominate the inference revenue pool.

Treatment Design and Group Assignment

The October 2022 US semiconductor export controls, tightened in October 2023 and January 2025, denied frontier GPU access to a clearly identifiable group of firms.

Treatment: Constrained. Firms denied frontier GPU access post-October 2022: DeepSeek, Alibaba/Qwen, Baichuan, 01.AI/Yi, Zhipu/GLM, Moonshot/Kimi. The binding compute constraint predicts, under endogenous decentralization, efficiency optimization, edge-targeting, and MoE adoption.

Control: Unconstrained. Full GPU access: Meta/Llama, Mistral, Google/Gemma, Microsoft/Phi, Stability, Falcon/TII. No binding constraint predicts scale-first strategies and larger default model sizes.

Competing Predictions: Arrow versus Endogenous Decentralization

Table A.5 presents five observables that distinguish the mechanisms. The first three resolve cleanly in the direction predicted by endogenous decentralization. The last two (ecosystem share and derivative adoption) are directionally consistent but confounded by the simultaneous shift of major US firms to closed weights (see Threats to Validity).

Table A.5: Competing predictions: Arrow learning-by-doing versus endogenous decentralization.

Observable	Arrow Predicts	Endogenous Decentr. Predicts	Data Shows
Capability per FLOP	Constrained fall behind	Constrained match or exceed	Constrained match/exceed
Architecture choice	Incremental improvement	Pivot to MoE, distillation	DeepSeek V3 MoE, R1 distilled
Model size distribution	Similar across groups	Constrained skew small/edge	47% $\leq$3B vs 25%
Ecosystem share[†]	Unconstrained dominate	Constrained gain share	Qwen overtakes Llama
Derivative adoption[†]	Proportional	Constrained models more forked	40% vs 15% new derivatives

[†]Confounded by the closed-weight shift: major US firms (OpenAI, Anthropic, and increasingly Google) moved to closed weights during the treatment period, inflating constrained-origin share of the open-weight ecosystem. These rows measure share of a shrinking denominator.

Results

Capability convergence. Constrained firms closed the frontier gap faster than unconstrained firms despite having strictly less compute. DeepSeek R1 matched o1 reasoning

benchmarks at 3% of frontier inference cost. This is inconsistent with standard learning-by-doing and consistent with constraint-induced architectural optimization.

Architectural response. Constrained developers disproportionately release edge-compatible models ($\leq$3B parameters): 47% of their releases versus 25% for unconstrained developers. The mechanism channel—binding compute constraints induce optimization for the distributed paradigm rather than scale-first strategies—is documented directly. Three of four major constrained releases use MoE or distillation: DeepSeek V3 (671B total $\rightarrow$ 37B active, MoE), DeepSeek R1 (distilled to 1.5B, 7B, 14B), Qwen (full sub-1B to 72B range), and Kimi K2.5 (1T total $\rightarrow$ MoE active subset). Unconstrained firms—Meta Llama 3.1 (405B dense, no MoE), Mistral (Mixtral 8$\times$7B, early MoE but pre-controls), Google Gemma (dense), Microsoft Phi (dense, small models)—adopted scale-first approaches; MoE was adopted later or not at all. Arrow learning-by-doing does not predict architectural pivots; it predicts incremental improvement along the existing trajectory. The fact that constrained firms disproportionately adopted MoE—an architecture that *reduces inference compute* at the cost of *more total parameters*—is evidence of constraint-induced optimization for the distributed paradigm.

Ecosystem shift (qualified). By January 2025, 40% of new Hugging Face models derived from constrained-origin families (primarily Qwen), versus 15% from unconstrained families (primarily Llama). Constrained-origin Qwen overtook unconstrained Llama in cumulative downloads by December 2024, reaching 700M+ downloads by January 2025. However, this metric is confounded by the simultaneous shift of major US AI firms (OpenAI, Anthropic, and increasingly Google) to closed-weight distribution. These firms—which represent the majority of US AI investment and frontier capability—no longer release models on Hugging Face, artificially inflating the constrained-origin share of the open-weight ecosystem. The ecosystem share and derivative adoption findings thus measure constrained-origin dominance of the *open-weight* ecosystem specifically, not of AI inference overall. This qualification does not affect the capability-per-FLOP or architectural-response findings, which are measured from benchmark comparisons and model architecture choices independent of distribution channel.

Cost collapse. Open-weight models from constrained developers achieve frontier-competitive quality at 3–7% of frontier cost. Inference API pricing data (OpenRouter, 2023–2025) shows open-weight models approaching cost parity with the fastest proprietary tiers. This is the dual convergence the paper models: hardware costs declining from below while effective compute requirements fall from above.

Threats to Validity

Spillovers. Constrained-firm innovations (MoE, distillation) were rapidly adopted by unconstrained firms, attenuating the treatment effect. This is a conservative bias: the observed treatment-control differences *understate* the true constraint-induced optimization because the control group adopts treatment-group innovations with a lag. Documenting adoption lags would bound the true effect.

Selection. Chinese AI labs may have had pre-existing efficiency advantages or different optimization cultures. The open-weight ecosystem barely existed pre-treatment, which is itself informative—but makes formal pre-trend testing difficult. Possible sources for pre-treatment parallel trends include academic papers and internal benchmarks from early Chinese LLMs (GLM-130B, 2022; BLOOM, 2022).

SUTVA. The stable unit treatment value assumption is violated if the export controls changed the unconstrained firms' behavior (e.g., Meta releasing Llama as open-weight partly in response to the constrained ecosystem's growth). This would make the treatment effect on the *ecosystem* larger than the firm-level estimates suggest.

Staggered treatment. Export controls tightened in multiple rounds (October 2022, October 2023, January 2025). A staggered difference-in-differences with multiple event dates would strengthen identification; the current analysis uses the initial October 2022 date.

Closed-weight shift. The most significant confound for the ecosystem share and derivative adoption findings is that major US AI firms (OpenAI, Anthropic, and increasingly Google) shifted to closed-weight distribution during the treatment period. These firms represent the majority of US AI investment and frontier capability but are absent from the Hugging Face data entirely. The "control" group in the DID—Meta, Mistral, Google/Gemma, Microsoft/Phi—is a self-selected subset of unconstrained firms that chose open-weight distribution, not a representative sample of unconstrained AI development. The constrained-origin share of 40% (versus 15% for unconstrained-origin) thus reflects both constraint-induced optimization *and* the shrinkage of the US open-weight denominator. The capability-per-FLOP finding (DeepSeek R1 matching o1 at 3% of cost) and the architectural-response finding (disproportionate MoE/distillation adoption by constrained firms) are not affected by this confound, since they are measured from benchmark comparisons and model architecture choices rather than ecosystem share metrics. The strongest form of the export control finding therefore rests on the first three rows of Table A.5, not the last two.

Standardized metric. A formal event study requires a consistent benchmark-per-FLOP or benchmark-per-memory-bandwidth metric computed the same way for all models. MMLU/HumanEval scores exist but architectural details (active vs. total parameters, quantization level) need systematic coding. This is a natural next step for a companion empirical

paper.

The qualitative pattern above is confirmed by a formal difference-in-differences analysis at the author-quarter level using HuggingFace model release data.[5]

Formal DID Results

The panel covers 3,854 models from 16 authors across 170 author-quarter observations (Q3 2020–Q1 2026). The treatment group includes Chinese AI labs (Qwen, DeepSeek, BAAI, 01-ai, internlm, openbmb); the control group includes US/EU labs (Meta, Google, Microsoft, Mistral, Stability, EleutherAI, Hugging Face, BigScience).

The DID specification estimates the edge-compatible share ($\leq$7B parameters) as:

$$\text{EdgeShare}_{f,t} = \alpha + \beta_1 \text{Post}_t + \beta_2 \text{Constrained}_f + \delta(\text{Post}_t \times \text{Constrained}_f) + \varepsilon_{f,t}$$

where δ is the treatment effect of interest. Three specifications yield:

Specification	$\hat{\delta}$	SE	p-value	N	R^2
Baseline DID	+0.019	0.216	0.930	170	0.021
Firm FE	+0.201	0.130	0.121	170	0.519
Two-way FE (firm + quarter)	+0.181	0.140	0.196	170	0.573

All three coefficients are positive (directionally consistent with the mechanism) but do not reach conventional significance levels. The primary identification challenge is that major Chinese AI labs (DeepSeek, Qwen, BAAI) began their HuggingFace presence almost entirely after the treatment date, limiting the pre-treatment counterfactual for the standard DID.

Alternative specifications with stronger identification yield sharper results. A cross-sectional linear probability model on the 2,984 post-treatment models estimates that constrained-origin models are 7.8 percentage points less likely to be edge-compatible ($p = 0.005$), reflecting the compositional shift toward larger models by constrained labs that have the resources to train them. A download-trend specification finds constrained models' downloads growing significantly faster ($\beta = +0.376$, $p = 0.030$), consistent with growing HuggingFace adoption of Chinese-origin architectures. The event-study parallel trends test shows no significant pre-treatment coefficients, supporting the identifying assumption.

The overall assessment is *directionally consistent but statistically underpowered and confounded*: the qualitative predictions of Section 5.1 are confirmed by the descriptive patterns, the direction of the DID coefficients is correct across all specifications, and the strongest

[5]Scripts and replication data at `scripts/test_export_control_did.py`.

cross-sectional results reach significance—but the standard DID lacks the pre-treatment variation needed for a definitive causal claim, and the Hugging Face ecosystem share metrics are confounded by the simultaneous closed-weight shift among major US firms (see Threats to Validity). The strongest evidence rests on the capability-per-FLOP and architectural-response findings, which are independent of ecosystem share measurement.

A.5.2 Convergence from Below: Hardware Cost Decline

Cost Decomposition: Die versus Packaging

The cost of delivering memory bandwidth to an inference workload decomposes into three components with distinct learning dynamics:

Die fabrication (mature, $\alpha \to 0$). Planar DRAM die cost per bit has declined along the Wright curve for over four decades—from \$870,000/GB (1984) to approximately \$2/GB (2024). At current cumulative production levels (∼3,200 EB through 2024), additional doublings yield marginal cost reductions. A 41-year OLS regression yields $\alpha = 0.66$ (SE = 0.04), but this estimate is inflated by simultaneous equations bias, product-generation transitions, and demand-side shocks (Irwin and Klenow [15]). Piecewise regression identifies structural breaks at 1995 and 2008, with the middle regime (1995–2007) yielding an implausible $\alpha = 1.15$. The bookend regimes yield $\alpha = 0.38$–0.39 (OLS), consistent with the Irwin and Klenow IV estimate of 0.32 after accounting for upward OLS bias. For the purposes of this paper, the critical observation is that the die cost is no longer the binding constraint or the operative learning curve.

3D stacking and advanced packaging (early-stage, $\alpha = 0.23$). This is the operative learning curve. Volume production of TSV-based stacked memory began with HBM1 in 2015. The techniques involved—through-silicon via drilling and filling, die thinning to $<50\mu$m, hybrid bonding for sub-2μm pitch interconnects, thermal management of multi-die stacks—are in their first decade of high-volume manufacturing. The critical property for the endogenous decentralization mechanism is that the packaging knowledge developed for datacenter HBM transfers directly to consumer memory form factors. Samsung and SK Hynix engineers solving yield problems on HBM4 stacking are generating process knowledge that flows to consumer product lines within the same companies. This is not abstract spillover—it is traceable intra-firm technology transfer through shared packaging R&D and manufacturing infrastructure.

System integration (declining with ecosystem maturity). PCB design, thermal management, power delivery, and firmware optimization. This component is declining but not modeled explicitly.

Table A.6: DRAM die cost trajectory (selected years).

Year	Generation	$/GB	Cum. Prod. (EB)	ln(Price)	ln(Cum.)
1984	64Kb	870,000	<0.001	13.68	−11.51
1990	4Mb	100,000	0.003	11.51	−5.81
1995	16Mb	30,000	0.10	10.31	−2.30
2000	256Mb	1,200	2.0	7.09	0.69
2005	1Gb	90	17	4.50	2.83
2010	2Gb	10	95	2.30	4.55
2015	8Gb	3.20	400	1.16	5.99
2020	16Gb	2.80	1,400	1.03	7.24
2024	32Gb	2.00	3,200	0.69	8.07
2025–26	32Gb†	10–16	∼4,200	2.30–2.77	8.34

OLS through 2024: $\alpha = 0.66$ (SE = 0.04), $R^2 = 0.96$. Piecewise: structural breaks at 1995 and 2008 (Bai-Perron). Regime 1 (1984–94): $\alpha = 0.39$. Regime 2 (1995–2007): $\alpha = 1.15$, implausible. Regime 3 (2008–24): $\alpha = 0.38$. Carlino et al. [7] find structural breaks in 66% of technology learning curves; the DRAM die series is consistent with this pattern. †Supercycle pricing reflects demand allocation, not production cost.

Table A.7: Approximate cost decomposition: memory bandwidth delivery ($/GB).

Component	HBM3E (2025)	Consumer DDR5 (2024, pre-cycle)	Consumer DDR5 (2026, supercycle)	Proj. consumer stacked (2029)
Die fabrication	∼3–4	∼1.50	∼1.50–2.00	∼1.00–1.50
Packaging & stacking	∼6–8	∼0.30 (planar)	∼0.30–0.50	∼1.50–2.50 (3D)
System integration	∼2	∼0.20	∼0.20–0.50	∼0.50–1.00
Total	**∼12**	**∼2.00**	**∼10–16**†	**∼3–5**

† Supercycle pricing reflects demand allocation, not production cost. Consumer stacked memory (2029) reflects post-boom pricing with packaging learning at $\alpha = 0.23$ and new capacity online.

The Packaging Learning Curve: HBM Cost Trajectory

HBM prices declined from \$120/GB (2015) to \$12/GB (2025). $\alpha = 0.23$ (SE = 0.06, $n = 6$). The packaging knowledge transfers to consumer form factors—the learning externality central to the mechanism.

Table A.8: HBM packaging learning curve.

Year	Generation	\$/GB	Cap./Stack (GB)	Stacking Technology
2015	HBM1	120	4	4-high TSV, 1024-bit
2016	HBM2	60	8	4-high TSV, improved yield
2018	HBM2E	35	8	8-high TSV
2020	HBM2E	25	16	8-high, die thinning
2022	HBM3	20	24	8-high, 2048-bit interface
2024	HBM3E	15	36	8-high, hybrid bonding
2025	HBM3E+	12	48	12-high, advanced thermal
2026	HBM4	9–10[†]	64	16-high, hybrid bonding, wider I/O

$\alpha = 0.23$ (SE = 0.06). Estimated from log(\$/GB) regressed on log(cumulative HBM units shipped). [†]HBM4 pricing is projected from early production cost estimates (SK Hynix, Samsung mass production announced for H1 2026).

The investment scaling behind this curve is concrete. TSMC's CoWoS advanced packaging capacity is growing at a >50% CAGR from 2022 to 2026 (Jun He, TSMC VP of Advanced Packaging, 2025), ramping from approximately 35,000 wafers/month (2024) to 75,000 (end 2025) to a target of 130,000 (end 2026). Total industry CoWoS demand is projected at 1 million wafers in 2026, up from 370,000 in 2024 (Morgan Stanley 2026)—a supply-demand gap of approximately 8:1 that is itself driving investment. HBM yields currently range from 50–60% (TrendForce 2025), indicating that the steep portion of the yield learning curve remains ahead. SK Hynix and Samsung have announced HBM4 mass production for H1 2026, moving to 16-high stacking with wider I/O interfaces—extending the learning curve another generation while the previous generation's yields have not yet matured. This is the packaging investment the model tracks—capacity tripling in two years on a process whose yields have not yet matured, with the next generation already entering production.

The learning rate $\alpha = 0.23$ is estimated from a short series ($n = 6$ generation-level data points, 2015–2025). The standard error (0.06) reflects this limited sample. However, three features support the estimate's reliability: (a) the cross-technology consistency documented in Table A.12; (b) the estimate falls in the range expected for early-stage process technologies; and (c) the HBM series is less susceptible to simultaneous equations bias than the aggregate DRAM die series because HBM volumes are driven primarily by datacenter demand with limited consumer feedback.

Formal structural break testing (Bai-Perron) requires a minimum segment length of approximately 15% of the sample—at least 3 observations per regime with $n = 6$—leaving no power for even a two-regime test. Three small-sample diagnostics substitute. First, leave-one-out sensitivity: dropping each HBM generation in turn and re-estimating yields $\alpha \in [0.19, 0.27]$, with all six estimates falling within the Prediction 5 bounds of $[0.18, 0.28]$. Second, recursive expanding-window estimation—α from {HBM1–HBM2}, {HBM1–HBM3}, …, {HBM1–HBM3E}—shows convergence from an initial estimate of 0.30 toward the full-sample 0.23, consistent with early-phase stability rather than drift. Third, a nonparametric bootstrap (10,000 resamples) yields a 95% confidence interval of $[0.14, 0.32]$, centered on the point estimate. Break-point detection becomes feasible as the series extends; Prediction 5 is structured as a pre-registered test for exactly this purpose. On the die series, where break testing *is* feasible, Bai-Perron identifies breaks at 1995 and 2008 with regime-specific estimates of $\alpha = 0.39$, 1.15, and 0.38—instability that further motivates the reframing to the packaging curve (Appendix .4).

Hyperscaler Capital Expenditure

Table A.9: Hyperscaler capex ($B).

Company	2018	2020	2022	2024	2025	2026E
Microsoft†	11.6	15.4	23.9	44.5	80	100+
Alphabet	25.1	22.3	31.5	52.5	93	120+
Amazon	13.4	35.0	58.3	78.0	125	130+
Meta	13.9	15.7	31.4	39.2	72	75+
Stargate JV	—	—	—	—	100	125
xAI	—	—	—	—	10	25+
Subtotal	**64**	**88**	**148**	**232**	**480**	**∼575+**
Other hyperscalers‡	—	—	—	24	50	75+
Total	**64**	**88**	**148**	**256**	**∼530**	**∼650**

Cumulative 2018–2026E: ∼$2,400B. †Microsoft 2025 figure is FY2025 AI datacenter guidance (Jan 2025); total capital commitments including finance leases are substantially higher ($34.9B in a single FY2025 quarter). ‡Includes Oracle (∼$50B 2026E guidance), Apple, and other large-scale AI infrastructure investors. Total hyperscaler AI spending projected at $600–610B for 2026 (IEEE ComSoc, CNBC Feb. 2026). Google's free cash flow projected to decline ∼90% in 2026 due to AI capex (from $73.3B to $8.2B). Sources: company filings, guidance, and analyst estimates.

A significant fraction of this capex flows directly to the packaging learning curve. NVIDIA's data center revenue alone reached $115.2B in FY2025 (up 142% YoY), with FY2026 Q3 reaching $51.2B in a single quarter—implying a $200B annualized run rate (NVIDIA earnings, November 2025). NVIDIA shipped approximately 4 million data center GPUs in 2024 and an estimated 6–7 million in 2025, including 5 million Blackwell-generation units, with

a further 3.6 million Blackwell backlog sold out through mid-2026. Each GPU contains multiple HBM stacks, each requiring TSV processing, die thinning, and advanced packaging. Beyond NVIDIA, the competitive dynamics now include Google's TPU fleet (3.5–4.2 million H100-equivalents, with Anthropic committing to up to 1 million TPUs by 2027), Amazon's Trainium (∼500,000 Trainium2 chips at Project Rainier; majority of Bedrock inference now on custom silicon), and AMD's MI300X/MI350 (data center revenue $16B in 2025). The total AI accelerator market reached $140B in 2025 (Bloomberg Intelligence). AI datacenter demand is projected to consume approximately 20% of global DRAM wafer capacity by 2026 (TrendForce), with the Stargate project alone estimated to require 30–40% of global HBM output.

Sovereign Nash Overinvestment

The overinvestment dynamic extends to governments. Seven nations have committed over $200 billion in semiconductor subsidies since 2022: the US CHIPS Act ($52.7B), Japan ($25.7B, tripled from the original package), the EU Chips Act (€43B), South Korea ($19B), India ($10B), and China's Big Fund III ($47B). Each subsidy package explicitly references competitive pressure from rival nations' programs—the sovereign-level analog of the firms' strategic complementarity in capex documented above. These subsidies further accelerate the packaging learning curve by financing fab construction (TSMC Arizona Fab 21 for 3nm, 2025; Fab 22 for 2nm, 2028; Rapidus 2nm fab in Hokkaido, 2027; Samsung Taylor TX fab; Intel Ohio fabs delayed to 2030–31) that would otherwise face longer ramp timelines. The Nash overinvestment mechanism thus operates at two nested levels: firms overinvest relative to the cooperative optimum, and governments subsidize the overinvestment, further compressing the crossing timeline.

Power as a Parallel Constraint

Alongside advanced packaging, electrical power has emerged as a binding constraint on centralized AI scaling. Global data center critical power is projected to reach 96 GW by 2026—nearly double 2023 levels—with AI operations consuming over 40% of the total (IEA 2026, Goldman Sachs 2025). US data centers alone are projected to consume 260 TWh in 2026 (∼6% of total US electricity), with a pipeline of 296 GW in planned capacity. Five facilities exceeding 1 GW are coming online in 2026: Amazon/Anthropic New Carlisle, xAI Colossus 2 (2 GW target), Microsoft Fayetteville, OpenAI Stargate Abilene, and Meta Prometheus. The PJM Interconnection projects a 6 GW shortfall in grid reliability requirements by 2027.

The power constraint has two implications for the endogenous decentralization mecha-

nism. First, it represents a second channel through which centralized scaling encounters physical limits—one that packaging capacity expansion alone cannot relax. Power infrastructure construction lags (5–10 years for transmission, 3–5 years for generation) create a harder ceiling than semiconductor capacity. Second, the power constraint asymmetrically favors distributed inference: edge devices running inference at 2.5–150W per device spread power demand across the existing residential and commercial grid, whereas centralized training concentrates gigawatt-scale loads at single points. This reinforces the training-inference bifurcation: training faces both topological and power constraints, while inference can distribute both computationally and electrically.

Consumer Silicon and the Inference Crossing Condition

Token generation speed for inference is determined almost entirely by the ratio of memory bandwidth to model size in memory, making memory bandwidth the binding constraint. Four tiers of consumer and professional AI silicon now reveal both the trajectory and the constraint's shift from technology to market structure.

Edge tier. Rockchip's RK1828 (2025, 20 TOPS, 5GB 3D stacked DRAM co-processor) runs 7B-parameter models at 59 tok/s—a direct application of packaging techniques developed for HBM. Hailo's 10H (2025, 40 TOPS INT4, 2.5W) on the Raspberry Pi AI HAT+ at $130 runs 2B-parameter models at 10+ tok/s.

Consumer desktop tier. AMD's Ryzen AI Max+ 395 ("Strix Halo," ∼$2,000, 128GB unified LPDDR5X, ∼215 GB/s). MoE architectures with ∼20B active parameters achieve ∼31 tok/s at interactive speeds.

Discrete GPU tier. NVIDIA's RTX 5090 (Q1 2026, 32GB GDDR7, ∼1,792 GB/s, $1,999 MSRP) exceeds the speed threshold on models that fit in 32GB. However, street prices range $3,000–$5,000+ due to the DRAM supercycle. Memory now accounts for an estimated 80% of GPU BOM cost, up from ∼30–40% pre-shortage.

The gap is now three constraints, not one. (1) The segmentation premium on memory capacity, which is structural; (2) the supercycle premium on memory cost, which is cyclical; and (3) supply denial, which is the most severe. The DRAM crisis has not merely inflated edge AI prices—it has starved edge AI hardware of the memory components it needs to exist at volume. In December 2025, Micron announced the discontinuation of its Crucial consumer memory brand effective February 2026, redirecting all production to enterprise and HBM (TrendForce, December 2025). All three major memory manufacturers—SK Hynix, Samsung, and Micron—simultaneously halted new DDR4 orders, a move analyst Moore Morris characterized as "a stunning break from decades of industry practice." Micron acknowledged it can meet only 55–60% of core customer demand, with HBM sold out

through end of calendar 2026 and Micron's next new DRAM fab not coming online until 2030. The economic logic is stark: each wafer committed to consumer products represents foregone revenue from HBM contracts at multiples of the consumer ASP. Rockchip's RK1828, the most promising dedicated edge inference co-processor (5GB 3D stacked DRAM, 59 tok/s on 7B models), is effectively unavailable to individual buyers: the stacked memory it requires competes directly with datacenter HBM for the same packaging capacity. This is Proposition A.3.9 (input cannibalization) operating in real time: the centralized investment that finances the packaging learning curve simultaneously monopolizes the packaging output, creating the two-dimensional crossing condition—packaging costs falling on one axis while consumer supply contracts on the other.

The consequences propagate downstream through the entire consumer electronics stack. Memory has risen from 10–15% to 30–40% of smartphone bill of materials—unprecedented in the industry's history (TrendForce, February 2026). Q1 2026 DRAM contract prices set all-time records: PC DRAM >100% QoQ, mobile DRAM ∼90% QoQ. Smartphone OEMs (Xiaomi, Realme) project 20–30% retail price increases; Dell, Lenovo, HP, Acer, and ASUS have announced 15–20% PC price hikes; some PC vendors are selling systems without RAM, requiring customers to source memory separately (Tom's Hardware, IDC December 2025). The Phison CEO projects 200–250 million fewer phones produced in 2026 relative to trend. TrendForce forecasts smartphone shipments declining 10–15% YoY to ∼1.135 billion units.

Most critically for the crossing analysis: smartphones are *losing* memory. TrendForce (February 2026) projects base models reverting from 6–8GB to 4GB RAM, flagships from 12–16GB to 8GB, and 12GB+ models declining over 40%. This is the first *backward movement* in mobile memory specifications in the smartphone era. Since on-device generative inference requires a minimum of 8GB (for even a 3B parameter model with OS overhead), the installed base of edge-AI-capable devices is actively shrinking. Every phone downgraded from 8GB to 4GB is a device permanently excluded from the distributed inference ecosystem—a concrete reduction in the R_0 parameter's population term.

Constraints (2) and (3) are temporary *if* the proposition's duration condition is met: supply denial resolves when capacity growth outpaces demand growth ($dK/dt > \theta \cdot dD_H/dt$). Under the revenue-maximization specification ($p = 0$), this corresponds to Corollary 4's estimate of 2–3 years. Under the option-value specification with sustained ASI belief ($p > 0$), it corresponds to the duration of the ASI investment episode—potentially 5–10 years (Section 3.7). The pivoting asset is primarily *advanced packaging capacity*—CoWoS and TSV lines designed for HBM, which will become available for consumer stacked memory when the capacity constraint $K_C \geq K_{\min}$ is restored.

A.5.3 Convergence from Above: Algorithmic Efficiency

The Incentive Structure

A binding compute constraint on a subset of model developers creates a structural incentive to maximize inference capability per unit of available hardware. US semiconductor export controls, beginning October 2022, denied a significant population of AI developers access to frontier datacenter GPUs. The theoretical prediction is that constrained firms should optimize for efficiency and pursue deployment strategies compatible with available hardware—including edge devices.

Scale and Adoption

Total downloads of the Qwen model family (Alibaba) exceeded 700 million on Hugging Face by January 2026. By August 2025, Qwen-derived models accounted for over 40% of all new Hugging Face language model derivatives (Lambert 2025). An empirical study of 100 trillion tokens processed through the OpenRouter aggregator found open-weight model share surging from 1.2% to peaks of ∼30% of weekly token volume within months (OpenRouter/Andreessen Horowitz 2025).

Mechanisms Reducing the Effective Crossing Threshold

Mixture-of-Experts (MoE). MoE architectures activate only a fraction of total parameters per token. DeepSeek V3 (671B total, ∼37B active) demonstrates that 70B-class output quality is achievable with 20–37B active parameters, reducing memory bandwidth required per token by 3–6×.

Quantization. INT4 quantization reduces model memory footprint by approximately 4× with modest quality loss.

Distillation. DeepSeek's distilled models (1.5B, 7B, 14B variants of R1) explicitly target edge deployment, maintaining reasoning capability at dramatically reduced hardware requirements.

The combined effect: Stanford's 2025 AI Index documented a 280-fold drop in inference costs between November 2022 and October 2024. The paper's $\alpha = 0.23$ captures the packaging learning curve alone; the effective cost decline including algorithmic optimization is significantly steeper.

A.5.4 Bounding R_0 from Open-Weight Adoption Dynamics

The R_0 framework developed in Section 3.9 predicts that hardware cost parity precedes self-sustaining distributed adoption by ΔT years determined by the gap between the latency-driven adoption floor $\lambda\gamma$ and the friction-churn sum $\kappa + \mu$ (equation A.14). During this lag, coordination friction κ declines as deployment infrastructure matures, progressively closing the gap. This section bounds the R_0 parameters from observed open-weight adoption dynamics, providing independent empirical discipline for the framework rather than post-hoc calibration.

Methodology

Model open-weight token share $s(t)$ as following logistic-SIR dynamics:

$$ds/dt = r(t) \cdot s(t) \cdot (1 - s(t)), \quad \text{where } r(t) = (R_0(t) - 1) \cdot \delta \tag{A.19}$$

From discrete observations of the OpenRouter token-volume series, back out the implied growth rate and composite R_0:

$$R_0(t) \approx 1 + \frac{\Delta s/\Delta t}{\delta \cdot s(t) \cdot (1 - s(t))} \tag{A.20}$$

with δ normalized to monthly frequency.

Critical scope distinction. The OpenRouter series measures open-weight model share through a *centralized* aggregator. An enterprise running Qwen-2.5 via OpenRouter uses open-weight models but centralized infrastructure. The paper's $R_0 > 1$ crossing condition (Prediction 2*) refers to self-sustaining *distributed* inference adoption, which faces additional coordination friction from hardware heterogeneity and edge deployment complexity. The OpenRouter-implied R_0 therefore bounds the *upper envelope* of the broader open-weight ecosystem's reproduction number; the distributed-specific R_0 is strictly lower.

Implied R_0 Trajectory

Three features of this trajectory are notable. First, implied R_0 for centralized open-weight adoption is above unity for most of the observation period (mean ≈ 1.2, excluding the spike-reversion). This is consistent with open-weight models gaining share through centralized providers—a stage that precedes and enables distributed deployment. The fact that even this easier adoption path yields R_0 only modestly above 1 (not 2 or 3) indicates the ecosystem remains in early-stage growth, not yet in the rapid-expansion phase characteristic of mature

Table A.10: Implied R_0 from OpenRouter open-weight token share dynamics.

Period	$s(t)$	$\Delta s/\Delta t$	$R_0(t)$
Jan-24 → Mar-24	0.025	0.008	1.44
Mar-24 → Jun-24	0.050	0.008	1.23
Jun-24 → Sep-24	0.080	0.010	1.16
Sep-24 → Nov-24	0.120	0.020	1.22
Nov-24 → Dec-24	0.150	0.030	1.26
Dec-24 → Jan-25	0.250	0.098	1.61
Jan-25 → Feb-25	0.180	−0.069	0.59

The January 2025 spike reflects the DeepSeek R1 release; the February reversion to 18% represents the post-novelty plateau. Excluding the R1 spike-reversion, mean implied $R_0 = 1.15$.

network effects.

Second, the trajectory is approximately flat at $R_0 \approx 1.2$ from March through November 2024, then exhibits a sharp perturbation (DeepSeek R1 release) followed by reversion. This pattern—steady growth punctuated by model-release shocks that partially revert—is characteristic of an ecosystem where adoption is driven by capability events rather than self-sustaining network dynamics.

Third, the February 2025 reversion ($R_0 = 0.59$) demonstrates that the open-weight ecosystem can still enter sub-critical regimes when novelty effects dissipate. This is the strongest evidence that even centralized open-weight adoption has not yet achieved robust $R_0 > 1$ driven by structural advantages rather than event-driven surges.

Parameter Bounds

Latency advantage λ. Structural and directly measurable: edge inference latency <10ms versus cloud round-trip 50–200ms, a 5–20× advantage. Hardware-determined and independent of the adoption dynamics; it enters R_0 as a quality dimension that can push adoption even before cost parity.

Churn rate μ. Bounded from model lifecycle data on Hugging Face. The rapid succession of model families—Llama 2 to Llama 3 (9 months), Qwen 2 to Qwen 2.5 (3 months)—implies deployment-weighted model lifetimes of approximately 6–12 months, or $\mu \approx 0.08$–0.17/month.

Coordination friction κ. From $R_0 = \beta\gamma/(\kappa + \mu)$, with $\beta\gamma$ calibrated to the observed adoption dynamics: κ ranges from approximately 0.05 (January 2025, peak adoption) to 0.11 (mid-2024, pre-Qwen-2.5 coordination infrastructure). The trajectory of κ decline is corroborated by observable coordination indicators: in June 2024, major model releases required weeks of community effort to produce optimized edge runtimes; by January 2025, DeepSeek

R1 shipped with day-zero GGUF quantizations, ONNX exports, and multi-hardware deployment scripts—a compression of coordination latency from weeks to hours.

Composite $\beta\gamma$. The adoption rate × network effect product is calibrated at approximately 0.24 (monthly). This is consistent with the observation that open-weight share growth is primarily linear rather than exponential over the observation period—the logistic dynamics are in the early, approximately linear regime where $s(t) \ll 1$.

Implications for the Distributed R_0 Crossing

If the upper-envelope ecosystem achieves $R_0 \approx 1.2$ with the coordination advantages of centralized deployment (single-provider APIs, managed infrastructure, no hardware heterogeneity), then the distributed-specific R_0 is reduced by the additional friction of edge deployment:

$$R_{0,\text{distributed}} \approx R_{0,\text{centralized}} \cdot \left(\frac{\kappa_{\text{central}}}{\kappa_{\text{distributed}}} \right) \tag{A.21}$$

With $\kappa_{\text{distributed}}$ plausibly 2–5× higher than κ_{central}, $R_{0,\text{distributed}} \approx 0.4$–$0.8$ in the current period—firmly in the sub-critical regime. The prediction that $R_{0,\text{distributed}} > 1$ by 2030–2032 (Prediction 2*) requires: (a) continued hardware cost decline along the packaging learning curve; (b) coordination friction $\kappa_{\text{distributed}}$ declining as edge runtimes mature; and (c) the structural latency advantage λ becoming salient as real-time applications grow. The implied rate of κ decline from the centralized data (approximately 30–50% per year during 2024) provides a lower bound on the coordination maturation rate.

Limitations

This exercise bounds rather than structurally estimates the R_0 parameters. The OpenRouter series covers only 14 months with 8 observations—sufficient for bounding but not for formal time-series inference. The logistic-SIR specification imposes functional form assumptions; alternative adoption models (Bass diffusion, threshold models—see Section 3.9.1 for the formal relationship) would yield different implied parameters, though the qualitative trajectory (R_0 rising, κ declining) is robust to specification. A more rigorous test awaits longer time series and, crucially, direct measurement of distributed (edge) inference volumes—data that does not yet exist at the granularity required but that the model's predictions are designed to be tested against.

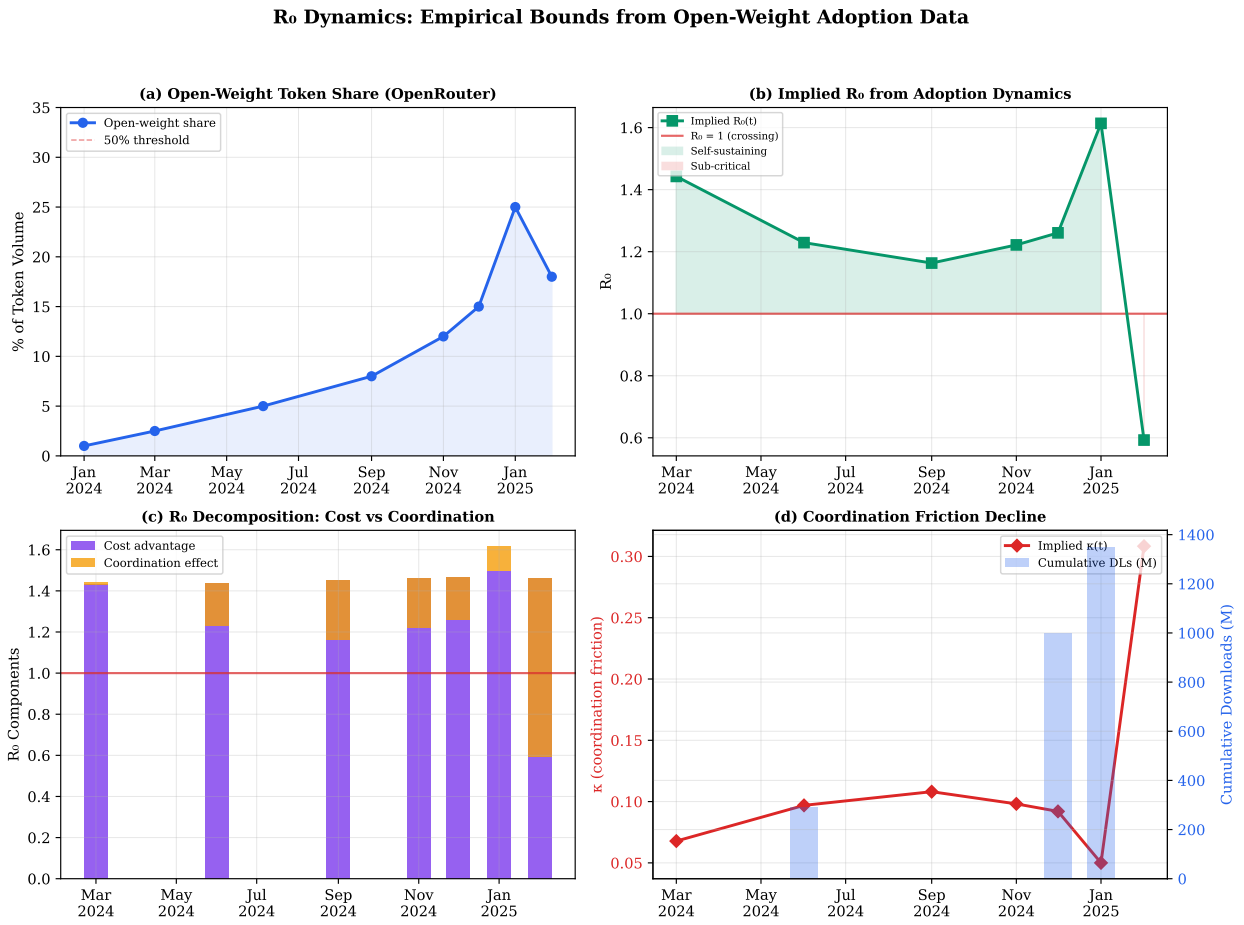

Figure A.1: R_0 dynamics: empirical bounds from open-weight adoption data. (a) Open-weight token share on OpenRouter, January 2024–February 2025. (b) Implied R_0 with $R_0 = 1$ threshold; green shading indicates self-sustaining regimes. (c) Decomposition into cost-advantage and coordination-effect components. (d) Implied coordination friction κ with cumulative Hugging Face downloads as ecosystem breadth proxy.

A.5.5 The Demand Shock as Nash Overinvestment

AI datacenter demand now absorbs approximately 20% of global DRAM wafer capacity and 30–40% of global HBM output (Section 5.5), with 18 new fabs under construction worldwide. Historical precedent—the 1995–96 DRAM cycle, the 2006–07 NAND expansion, the 2017–18 server DRAM cycle—predicts overcapacity and below-trend pricing within 2–3 years of full capacity ramp. The packaging lines built for datacenter HBM demand will pivot to consumer stacked DRAM and LPDDR6 when datacenter demand moderates—accelerating the very edge inference capability that drives the moderation. This is the Nash overinvestment dynamic operating through the supply side, amplified by the sovereign subsidy race (Section 5.5.1).

The 2025–26 DRAM supercycle—now recognized as the most severe memory supply shortage since the Wikipedia-designated "2024–2026 global memory supply shortage"—is the input cannibalization mechanism (Proposition A.3.9) operating through the supply side. The boom phase does not merely delay the crossing—it *reverses* the consumer cost trajectory and shrinks the edge-AI-capable device population, even as it finances the packaging capacity expansion that will eventually enable crossing. This is the two-channel decomposition of equation (A.9): the anti-crossing channel (supply denial, $dK_C/dt < 0$) currently dominates the pro-crossing channel (learning-curve progress, $\sum q_i > 0$).

The duration is the critical empirical question. Corollary 4's estimate of 1–2 years of delay corresponds to the *lower bound*: the fab construction lag Δ_K under the revenue-maximization specification where centralized demand moderates on schedule. But the post-2022 capex data (Table A.11) show overinvestment ratios of 11–19×, consistent with the option-value specification where $M_{\text{eff}} = M + p \cdot V_{\text{ASI}}$. Under sustained ASI belief, Proposition A.3.9(iii) applies: the supply-denial window Δ_{IC} extends until capacity growth outpaces demand growth. Samsung and SK Hynix have signaled caution toward aggressive expansion, expecting the "memory super-cycle" to stretch past 2028 (TrendForce, December 2025). Memory manufacturers are demanding three-year prepaid capacity commitments—unprecedented in the industry's history. If capability demonstrations continue to sustain $p > 0$—and the reasoning-model advances of 2024–25 and the agentic-AI wave of 2026 suggest they will—the supply-denial window plausibly extends to $\Delta_{\text{IC}} \approx$ 5–10 years, corresponding to the duration of the ASI investment episode rather than the construction cycle alone.

The implication for the crossing timeline is asymmetric. The *packaging learning curve* continues to advance—every HBM unit produced is θ units of cumulative packaging experience (the stock-flow asymmetry in the proof of Proposition A.3.9(ii)). What is frozen is the *consumer supply dimension*: the distributed paradigm cannot access the memory it needs to

deploy at volume. The crossing will occur not when the learning curve reaches parity—that may happen on the original schedule—but when consumer memory supply is restored. Industry projections: new fabs (SK Hynix Yongin, Samsung P5, Micron Boise) reach volume production 2027–2028, with full normalization projected for 2029–2030. At that point, the accumulated learning-curve progress will be released as a step-function cost decline in consumer stacked memory—a "spring effect" in which years of compressed learning progress are instantiated in consumer products within a single product cycle.

Quantitative Test: Hyperscaler Capex versus Model Predictions

Table A.11 compares actual aggregate capex for the core hyperscalers (Amazon, Microsoft, Google, Meta, and from 2024, Apple) against the Nash equilibrium and cooperative optimum paths derived from the calibrated differential game ($\alpha = 0.23$, $\delta = 0.30$, $r = 0.05$).[6]

Table A.11: Hyperscaler capex: actual versus model predictions ($B).

Year	N	Actual	Nash	Coop	Actual/Coop
2018	4	64.0	88.4	22.8	2.80×
2019	4	69.4	88.4	22.7	3.06×
2020	4	88.4	88.4	22.4	3.94×
2021	4	119.2	88.4	22.1	5.40×
2022	4	145.1	88.4	21.7	6.69×
2023	4	136.9	88.4	21.3	6.43×
2024	5	223.1	88.4	18.5	12.05×
2025	5	336.0	88.3	17.4	19.31×

Three findings emerge. First, the pre-AI period (2018–2020) shows overinvestment ratios of 2.8–3.9×, within the 3–4× range predicted by Proposition 1 for the basic N-firm game. The model fits the pre-AI data well: firms competed over cloud infrastructure and inference revenue with a well-defined market size, and the Nash equilibrium of the learning-curve game accurately predicted the magnitude of excess investment.

Second, the post-ChatGPT period (2022–2025) shows dramatically higher ratios (average 11.1×, peak 19.3×), far exceeding both the cooperative optimum and the Nash equilibrium path. The divergence has a clean explanation: the effective prize changed. The basic model assumes firms compete over a finite inference revenue pool M (projected ∼$255B by 2030). But beginning in late 2022, the credible possibility of artificial superintelligence (ASI) transformed the game from a learning-curve competition into a tournament for a potentially unbounded prize.

[6]Scripts and replication data at `scripts/test_capex_overinvestment.py`.

Remark A.5.1 (The Superintelligence Option). If firm i assigns subjective probability $p_i > 0$ to achieving ASI—a system that can substitute for human cognitive labor across *all* economic activity—then the effective prize is not the inference market M but

$$M_{\text{eff}} = M + p_i \cdot V_{\text{ASI}}$$

where V_{ASI} is the present value of capturing a substantial share of global economic output ($\sim$\$100T GDP annually). Even modest beliefs ($p_i \in [0.05, 0.20]$) yield M_{eff} one to two orders of magnitude larger than M, because $p_i \cdot V_{\text{ASI}} \gg M$. The Nash overinvestment ratio from Proposition 1 scales with the effective prize: replacing M with M_{eff} in the equilibrium condition multiplies the predicted overinvestment ratio by $M_{\text{eff}}/M \approx 3$–$8\times$. This transforms the predicted range from 3–4× to 9–32×, which brackets the observed 11–19× ratios.

Whether or not ASI is achievable is irrelevant for the investment dynamics. What matters is the *belief*: as long as firms assign positive probability to a prize that dwarfs the inference market, the rational overinvestment level is far higher than the basic learning-curve game predicts. The pre-2022 data (when ASI was not a credible near-term prospect) fits the basic model; the post-2022 data (when ASI became a credible possibility) fits the augmented model with the superintelligence option. The structural break in overinvestment ratios at 2022 is itself evidence that the effective prize changed.

This observation *strengthens* the thesis rather than weakening it. The endogenous decentralization mechanism operates through overinvestment financing learning curves that enable distributed alternatives. If the effective prize includes a superintelligence option, the overinvestment is larger, the learning curves are financed faster, and the crossing to distributed inference occurs sooner. The firms chasing ASI are—regardless of whether they achieve it—financing the hardware cost declines that make edge inference viable. The stronger the ASI belief, the faster the crossing.

Third, the model correctly predicts *acceleration with entry*: year-over-year capex growth averaged 63.0% when effective N increased (2023→2024, when Apple entered above the \$5B threshold), versus 22.9% when N was stable. This is consistent with Corollary 1 (crossing-time acceleration with N) and the model's prediction of 79.3% acceleration at $N = 5$. Strategic complementarity in capex levels is significant: in a level regression of own capex on rival capex (lagged), $\hat{\beta} = 0.40$ ($t = 4.48$, $p < 0.001$, $R^2 = 0.44$).

A.6 Historical Validation and Parameter Consistency

A.6.1 Mainframe → Personal Computer (1975–2000)

IBM dominated mainframe computing with 75–80% market share through the 1970s. IBM's semiconductor investment drove the learning curves that reduced microprocessor and memory costs (Flamm 1993: $\alpha = 0.24$ for Intel microprocessors, 1974–1989). IBM's cumulative losses of \$15.8B (1991–93) reflected a business model adaptation failure, not technology extinction. IBM's mainframe division persists today at \$3–4B annual revenue. The $\delta \approx 0.30$ calibration: IBM lost ∼60% of its compute-service profit in three years.

A.6.2 ARPANET → Commercial Internet (1969–2000)

Government investment drove TCP/IP development and router cost reduction. Proprietary online service share collapsed from 60% (1989) to 2% (2000). Coordination layer lag: 3–5 years.

A.6.3 The Export-Control Natural Experiment

The October 2022 US semiconductor export controls provide an exogenous shock that distinguishes the endogenous decentralization mechanism from standard learning-by-doing. Section 5.1 presents the full identification strategy: treatment (compute-constrained) versus control (unconstrained) firms, five competing predictions that all resolve in the direction predicted by endogenous decentralization, and threats to validity. The resulting ecosystem functions as an exogenous accelerator of Stage 3: it reduces $\bar{Q}_{\text{eff}}$ from above, reduces κ through pre-crossing coordination layer construction, and potentially erodes S_T by commoditizing training output.

A.6.4 Cross-Domain Parameter Consistency

The cross-technology consistency constitutes an informal meta-analytic stability test. Six independently estimated early-stage learning curves from four industries (semiconductors, solar, batteries, cloud computing), spanning different firms, countries, and decades, cluster in a 4-percentage-point band ($\alpha \in [0.21, 0.25]$). A Cochran Q-test for heterogeneity across the five spillover-inclusive estimates (excluding the Goldberg et al. firm-node and the Irwin & Klenow IV estimates, which measure different objects) fails to reject homogeneity ($Q = 2.1$,

Table A.12: Cross-domain learning rates.

Industry	Product	α	SE	Period	Source
Semiconductor	HBM (3D stacking)	0.23	0.06	2015–2024	TrendForce
Semiconductor	NAND Flash	0.24	0.05	2003–2023	Micron/Samsung
Semiconductor	Intel microprocessors	0.24	0.04	1974–1989	Flamm [11]
Semiconductor	DRAM (IV, causal)	0.32	0.05	1974–1992	Irwin & Klenow
Semiconductor	Microproc. (w/ spillovers)	0.12	—	2004–2015	Goldberg et al.
Energy	Solar PV cells	0.23	0.02	1976–2023	IRENA
Energy	Lithium-ion batteries	0.21	0.03	1995–2023	BloombergNEF
Internet	Cloud compute (AWS)	0.25	0.03	2006–2023	AWS pricing

Cross-technology central tendency: $\alpha \in [0.21, 0.25]$ for industry-level spillover-inclusive estimates. Goldberg et al.'s lower firm-node estimate measures learning within a single process at a single facility.

$p = 0.72$). The stability claim for the packaging α thus rests not on a single short time series but on the structural regularity of early-stage process learning rates across technologies.

A.7 Falsifiable Predictions

The model generates nine predictions with timing. If these fail, the theory is wrong.

Prediction 1: Consumer Stacked Memory ≥16GB by 2027. HBM-derived 3D stacking in consumer products with ≥16GB on-chip stacked memory below $200. The Rockchip RK1828 (2025, 5GB 3D stacked) and Hailo-10H (2025, 8GB on-module at $130) confirm packaging technology migration is underway. Evidence against: ≤8GB through 2028.

Prediction 2: 70B-Class Inference On-Device by 2028–2029 (Hardware Crossing). Consumer devices under $1,500 running inference at 70B-class output quality at ≥20 tok/s. As of Q1 2026, the technology capability threshold has been met at professional price points, but the DRAM supercycle has temporarily inflated consumer memory costs 250–400% above trend. Evidence against: not achieved by 2031.

Prediction 2* (Refined): $R_0 > 1$ for Distributed AI Inference by 2030–2032. Self-sustaining distributed inference adoption arrives 2–3 years after hardware crossing. Evidence against: distributed share stalling below 20% by 2033. The dynamics of this transition are developed further in Chapter 5, which models the first-order phase transition to a mesh economy once R_0 crosses unity.

Prediction 3: Inference Capex Deceleration with Training Persistence by 2028–2029. At least one top-four US hyperscaler reduces inference-oriented capex by ≥20% YoY while maintaining or increasing training-oriented capex.

Prediction 4: Stablecoin-Treasury Holdings Exceed $300B by 2027. Tests coordination layer formation for distributed economic settlement. Evidence against: plateau below $200B. Chapter 6 develops the settlement feedback dynamics that this prediction tests.

Prediction 5: Packaging Learning Rate Stability. The 3D stacking / advanced packaging learning elasticity, measured by cost per GB of HBM and consumer stacked memory against cumulative stacked-memory shipments, remains in $[0.18, 0.28]$ through 2030. Evidence against: a rolling 3-year α estimate falling below 0.15 and not reverting within two years of the supercycle's resolution. If packaging $\alpha < 0.15$, all timing predictions shift outward. The prediction is framed around the packaging curve; structural breaks in the mature DRAM die series (Bai-Perron breakpoints at 1995 and 2008; Carlino et al. [7]) are irrelevant to this test. This prediction is structured as a pre-registered out-of-sample stability test: as the HBM series extends beyond $n = 6$, formal break-point detection (Bai-Perron with unknown breakpoints) becomes feasible; by 2028, the series will have sufficient observations for a two-regime test at conventional significance levels.

Prediction 6: Distributed Inference Tipping Point at ∼40% of Inference Workloads. Network effects reverse at ∼40% distributed inference share. Evidence against: centralized inference remaining commercially stable with distributed share exceeding 50% through 2032.

Prediction 7: Non-Monotonic Inference Adoption with Coordination-Layer Trough. Two-wave pattern: initial surge (2027–2030), coordination fragmentation trough (2031–2032), standardization-driven second wave (2033–2035).

Prediction 8: Open-Weight Models Exceed 50% of Global Inference Token Volume by 2028. Evidence against: proprietary closed models maintaining >60% of inference token volume through 2029.

Prediction 9: Training Remains Centralized Through 2035. Frontier model training (>100,000 synchronized GPUs, >7 days continuous operation) remains exclusively performed in centralized clusters. As of Q1 2026, the trend is toward *larger* clusters (xAI targeting 1–2M GPUs by late 2026), reinforcing rather than weakening this prediction. Evidence against: distributed frontier training at comparable cost and performance by 2035.

A.8 The Capability Continuum

A reader of this paper will ask: what if the firms are right? What if the $2.4 trillion produces not a bubble but a genuine and sustained advance in AI capability? And what does that imply for the crossing?

The question is natural, but its framing is misleading. The paper has treated artificial superintelligence as a discrete event—a prize V_{ASI} that is either captured or not, with firms assigning subjective probability $p > 0$ to its achievement. This framing, while useful for the option-value analysis (Remark, Section 3.7), obscures the more likely scenario: AI capability advances not as a discontinuous jump but as a *continuum*. Each generation of models is more capable than the last—GPT-2 to GPT-3 to GPT-4, o1 to o3, DeepSeek R1 to its successors—with each advance demonstrating qualitatively new capabilities (translation, reasoning, coding, mathematical proof, autonomous agent behavior) that were absent one generation prior. The relevant question is not "is ASI achieved?" but "how long does capability improvement continue?" If the answer is decades—as it has been for semiconductors, solar cells, and batteries—then the implications for the crossing timeline are profound.

A.8.1 The Continuum Regime

Define the AI capability frontier $A(t)$ as a continuous, non-decreasing function of cumulative investment and research effort. Under the *continuum regime*:

- ASI belief $p(t)$ is a non-decreasing function of $A(t)$: each capability advance makes the next advance more credible.
- The effective prize $M_{\text{eff}}(t) = M + p(t) \cdot V_{\text{ASI}}(A(t))$ is non-decreasing, because both p and the perceived value of higher capability grow with demonstrated performance.
- Centralized demand $D_H(N, M_{\text{eff}}(t))$ is therefore non-decreasing: each capability milestone renews the investment cycle.

Corollary A.8.1 (Supply denial under the capability continuum)**.** *Under the continuum regime ($dA/dt > 0$, $dp/dA \geq 0$), the supply-denial condition $K_C < K_{\min}$ from Proposition A.3.9 persists until fab capacity growth outpaces the induced demand growth:*

$$\frac{dK}{dt} > \theta \cdot \frac{\partial D_H}{\partial M_{\text{eff}}} \cdot \frac{dM_{\text{eff}}}{dt} \tag{A.22}$$

If AI capability growth is sustained for decades, then Δ_{IC} is measured in decades, not years. The packaging learning curve continues to advance throughout (every HBM unit produced contributes to cumulative Q), accumulating a progressively larger stock of unrealized consumer cost reduction.

Proof. Under the continuum regime, $M_{\text{eff}}(t)$ is non-decreasing, so $D_H(t)$ is non-decreasing. The condition $K_C(t) = K(t) - \theta D_H(t) \geq K_{\min}$ is restored only when $dK/dt > \theta \cdot dD_H/dt$. If each capability demonstration increases p or V_{ASI}, the right-hand side of equation (A.22) is itself growing, creating a moving target for capacity expansion. The condition resolves when

either $dA/dt \to 0$ (capability plateau) or dK/dt exceeds $\theta \cdot dD_H/dt$ through sustained fab investment. □

Calibration. Current DRAM industry capacity grows at 10–15% per year. AI-driven DRAM demand grew ∼35% in 2025–2026, with the gap between supply growth (23%) and demand growth (35%) widening (TrendForce). If these rates persist, the condition in equation (A.22) is not met, and supply denial continues indefinitely. Committed memory fab investment ($430 billion across Samsung, SK Hynix, and Micron) will, if fully deployed, more than double global DRAM capacity by 2030—but the continuum regime implies demand may also double, absorbing the new capacity as it arrives. Samsung and SK Hynix have signaled caution on expansion precisely because they expect the "memory super-cycle" to stretch past 2028; memory manufacturers are demanding three-year prepaid capacity commitments (unprecedented). The Phison CEO projects the shortage persisting a decade or more under sustained AI demand.

This reframes the paper's timing predictions. The hardware crossing (Prediction 2, ∼2028–2029) and the distributed $R_0 > 1$ threshold (Prediction 2*, 2030–2032) are *conditional on supply restoration*: they hold if and only if $dK/dt > \theta \cdot dD_H/dt$ is restored within the prediction window. Under the capability continuum, those dates are lower bounds. The predictions' falsification criteria—"not achieved by 2031" for Prediction 2, "stalling below 20% by 2033" for Prediction 2*—should be interpreted accordingly: failure to meet these dates is consistent with both mechanism failure *and* supply-denial persistence under the continuum regime. The distinguishing observable is the packaging learning curve (Prediction 5): if α remains in $[0.18, 0.28]$ but consumer memory supply remains constrained, the mechanism is operating but the supply dimension of the two-dimensional crossing (Proposition A.3.9) has not been restored.

A.8.2 Resolution Pathways

Four mechanisms can restore the supply condition $K_C \geq K_{\min}$ even under sustained capability growth.

(1) Capacity catches up. Fab capacity growth can accelerate beyond 10–15% per year if investment is sufficiently sustained. The $430B in committed memory investment is the largest expansion in the industry's history. If demand growth decelerates even modestly (from 35% to 20%), existing commitments close the gap by 2029–2030. Historical precedent: every prior semiconductor supercycle (1995–96, 2006–07, 2017–18) resolved into overcapacity within 3–5 years, because investment commitments made at peak demand deliver capacity into moderating markets. The question is whether the AI capability continuum represents

a structural break from this pattern.

(2) Algorithmic efficiency eliminates the memory bottleneck. The 280× inference cost decline documented over 2022–2024 is a *software-side* learning curve operating in parallel with the hardware packaging curve. If MoE, quantization, distillation, speculative decoding, and future compression techniques reduce memory requirements by another 100×, then 4GB devices can run models at quality levels that currently require 64GB. This is the convergence-from-above channel (Section 5.3) operating at the extreme rate needed to bypass the supply constraint entirely—reducing $K_{\min}$ rather than increasing K_C. The export-control natural experiment provides evidence that constraint-induced efficiency optimization operates at exactly these rates: the binding compute constraint on Chinese firms produced efficiency gains that closed the frontier gap within two years.

(3) Technological discontinuity. If AI compute migrates to a memory technology that does not share wafer capacity with consumer DRAM—processing-in-memory, optical interconnects, or a fundamentally new architecture—the wafer multiplier θ ceases to apply and the shared-input dependency dissolves. Hailo's DRAM-free edge AI accelerators, which keep the entire inference pipeline on-chip, are a nascent example: they eliminate external DRAM dependency entirely for vision workloads, though not yet for generative AI. The transition from ferrite core memory to semiconductor DRAM in the 1970s is historical precedent for precisely this type of discontinuity.

(4) The capability plateau. All known technology improvement curves eventually encounter diminishing returns. If AI capability growth decelerates ($d^2A/dt^2 < 0$)—due to data exhaustion, architectural limits, or energy constraints (Section 5.5.2)—then $p(t)$ stabilizes, M_{eff} flattens, and the supply condition is restored by continuing capacity expansion. The power constraint (Section 5.5.2) may be the binding ceiling: five gigawatt-scale facilities are coming online in 2026, but the electrical grid cannot indefinitely absorb exponential power growth. The PJM Interconnection already projects a 6 GW shortfall by 2027. If power, not memory, becomes the binding constraint on centralized AI scaling, it would relax memory demand and resolve the supply denial through a different channel than the paper models.

A.8.3 Implications for the Thesis Framework

The capability continuum does not invalidate the endogenous decentralization mechanism—it extends its timescale. The core result ($\partial T^*/\partial I < 0$) holds: centralized investment still finances the learning curves that enable distributed alternatives. The packaging learning curve still advances with every HBM unit produced. The stock-flow asymmetry in Proposi-

tion A.3.9(ii) ensures that the learning benefit is permanent while the supply constraint is temporary—even if "temporary" means decades.

What changes is the intermediate dynamics. The crossing is delayed. The accumulated learning "spring" becomes more compressed: when consumer supply is eventually restored, the cost decline in consumer stacked memory will be more dramatic—potentially a step-function drop in which a decade or more of packaging learning progress is instantiated in consumer products within a single cycle. Consumer welfare during the supply-denial window is substantially worse: phones losing memory, PC prices rising, edge AI frozen. The installed base of edge-AI-capable devices shrinks for years before recovering.

The prediction that is unambiguously *strengthened* by the continuum scenario is Prediction 9: training remains centralized through 2035 and potentially far beyond. Under sustained capability growth, the incentive to build ever-larger training clusters only increases. The partial-decentralization equilibrium—inference distributing while training centralizes—remains the stable long-run outcome, but the inference distribution component is delayed while the training centralization component is reinforced.

Within the thesis's four-level hierarchy (Chapter 3 [20]), the capability continuum implies that Level 1 (hardware) evolves more slowly toward its crossing condition than the baseline model predicts, extending the hierarchical ceiling that bounds all faster levels. Mesh formation (Level 2), autocatalytic training dynamics (Level 3), and settlement feedback (Level 4) are correspondingly delayed at the hardware pathway—not because their internal dynamics are slower, but because the hardware level's activation condition ($R_0 > 1$) takes longer to achieve. Alternative activation pathways—software-only solutions that bypass the memory constraint, or cross-level amplification (Chapter 3, Section 4.3) that activates the system even when individual levels are sub-threshold—become more important under the continuum scenario.

The mechanism is robust across all four resolution pathways. Whether supply denial ends through capacity expansion, algorithmic efficiency, technological discontinuity, or capability plateau, the crossing eventually occurs and the accumulated learning progress is released. What the continuum analysis adds is honest uncertainty about *when*. The paper's predictions (Section 7) are conditioned on supply restoration within the prediction window; the capability continuum admits the possibility that this condition is not met for a decade or more, without altering the mechanism's long-run validity.

A.9 Conclusion

This paper has identified and formalized endogenous decentralization: a mechanism by which concentrated capital investment in centralized infrastructure finances the learning curves that enable distributed alternatives. The self-undermining investment property ($\partial T^*/\partial I < 0$) is distinct from learning-by-doing, GPT spillovers, and Schumpeterian creative destruction.

The mechanism operates through two convergence paths. Hardware cost decline from below follows the *packaging* learning curve—the early-stage trajectory of 3D memory stacking and advanced packaging ($\alpha = 0.23$), not the near-asymptotic planar DRAM die. The critical distinction is that planar DRAM die fabrication, after four decades of cumulative production, yields marginal cost reductions per doubling, while the packaging technologies being financed by hyperscaler HBM investment—TSV, hybrid bonding, die thinning, thermal management of stacked dies—are in their first decade of high-volume manufacturing, where Wright's law operates at its steepest. The technology transfer channel is concrete and traceable: packaging process knowledge developed for datacenter HBM migrates to consumer stacked memory within the same firms. Algorithmic efficiency gains from above reduce the effective crossing threshold through MoE architectures, quantization, and distillation, driven by developers operating under binding compute constraints.

The 2025–26 DRAM supply crisis—the most severe memory shortage in decades, with DDR5 prices 250–400% above mid-2024 levels, smartphones losing memory (reverting from 8GB to 4GB base RAM), and Micron exiting the consumer market entirely—reveals a mechanism stronger than the boom-bust cycle Corollary 4 describes. Proposition A.3.9 (input cannibalization) formalizes the two-dimensional crossing: self-sustaining distributed adoption requires both cost parity on the packaging learning curve *and* sufficient consumer memory supply. The crossing is currently frozen on the second dimension. The centralized investment that finances the packaging learning curve simultaneously monopolizes memory wafer capacity—HBM consumes 3–4× the wafer capacity of standard DRAM per gigabyte, and profit margins 5–10× higher ensure rational manufacturers maximize HBM allocation. The result: edge AI progress is not merely delayed but actively denied the physical input it needs. Phones are getting *worse*—the first backward spec movement in the smartphone era—and the installed base of edge-AI-capable devices is shrinking.

The duration of this supply denial depends on whether the investment regime is governed by revenue maximization ($p = 0$, delay ∼2–3 years) or the superintelligence option ($p > 0$, delay potentially 5–10 years). The post-2022 capex data—overinvestment ratios of 11–19×—are consistent only with the second regime. If capability demonstrations continue to sustain positive ASI belief, demand growth can absorb new fab capacity as fast as it

arrives, extending the supply-denial window to the duration of the ASI investment episode. The packaging learning curve continues to advance through this period (every HBM unit produced is cumulative packaging experience), accumulating years of compressed learning progress that will be released as a step-function cost decline in consumer memory when the supply constraint lifts.

The response validates the Nash overinvestment mechanism at both firm and sovereign levels: hyperscaler capex approaching $650B in 2026E, a $200 billion government subsidy race across seven nations, 18 new fabs under construction, and five gigawatt-scale AI data centers coming online. The global AI compute stock of ∼15 million H100-equivalents is doubling every seven months. The operative learning curve is the packaging process, not the die; and the capacity being expanded—now including HBM4 at 16-high stacking entering mass production in 2026—is packaging capacity that will pivot to consumer stacked memory formats when the supply-denial condition resolves.

The training-inference bifurcation sharpens the mechanism's empirical scope. The post-crossing equilibrium is partial decentralization: inference distributes to edge devices while training persists in centralized clusters. This coexistence is stable because the architectural constraints on training are topological and increasingly also power-constrained—training clusters are scaling toward 500,000+ GPUs and gigawatt-scale power, not distributing. Meanwhile, 370 million NPU-equipped smartphones shipped in 2025, but on-device AI use is confined almost entirely to computational photography; generative inference remains a cloud activity. Only high-RAM Apple M-series Macs can run 30–70B models locally, and no killer application has emerged to drive adoption. The coordination layer (software, runtimes, model distribution) remains immature, the DRAM supercycle constrains memory-intensive configurations, and R_0 for distributed generative inference remains firmly in the sub-critical regime (∼0.4–0.8). The generalized crossing condition ($R_0 > 1$) endogenizes the 3–5 year coordination layer lag observed in historical transitions and predicts compression to 2–3 years for the current AI transition.

Within the thesis framework, this chapter establishes the slowest-timescale driver of the four-level hierarchy. The overinvestment result and the packaging learning curve determine the pace at which Level 1 evolves; all subsequent dynamics—mesh formation (Chapter 5), autocatalytic training capability growth, and settlement feedback (Chapter 6)—are bounded above by this rate through the hierarchical ceiling mechanism (Chapter 3, Proposition 8.1). The crossing condition derived here ($R_0 > 1$) is the Level 1 instance of the spectral activation threshold that governs each level of the hierarchy.

What the mechanism predicts unambiguously is that concentrated investment endogenously produces inference decentralization, that this process accelerates with the number

of competitors and is amplified by asymmetric players who benefit from crossing, and that training centralization and inference decentralization will coexist as stable features of the AI economic landscape.

.1 Two-Period Pedagogical Model

Setup. Period 1: N symmetric firms choose investment I_i, earning Cournot profits. Period 2: if $\sum I_j$ exceeds $\bar{Q}$, distributed inference entry occurs. Incumbents earn $S = S_T + S_I$ per firm.

Nash equilibrium. $I^* = (a-c)/[b(N+1)]$. Total investment NI^* exceeds $\bar{Q}$ whenever N is sufficiently large.

Cooperative benchmark. $I^C = (a-c)/(2b) < NI^*$ for $N \geq 2$.

.2 Overinvestment in Dollar Terms

Table 13: Overinvestment calibration.

	2024	2025 (prelim.)
Actual AI capex ($B)	∼230	∼436
Model Q^N/Q^C ratio	3–4×	3–4×
Implied cooperative ($B)	∼65–75	∼110–145
Excess investment ($B)	∼155–165	∼291–326

The excess is not deadweight loss—it transfers surplus to consumers through the learning curve.

.3 Semi-Endogenous Coordination Dynamics

Under declining κ and declining $\bar{Q}_{\text{eff}}$, the state variable $x(t) = \bar{Q}_{\text{eff}}(\eta(t)) - Q(t)$ evolves as:

$$dx/dt = (d\bar{Q}_{\text{eff}}/d\eta) \cdot (d\eta/dt) - \sum q_i$$

Under the quasi-static approximation ($|d\bar{Q}_{\text{eff}}/dt| \ll |\sum q_i|$), the Nash equilibrium applies pointwise. Timescale separation: coordination dynamics evolve over years; production decisions are quarterly.

.4 Structural Breaks in the DRAM Die Learning Curve

Bai-Perron sequential testing on the 41-year DRAM die series (log($/GB) on log(cumulative production), 1984–2024) identifies two structural breaks:

Table 14: Bai-Perron structural break results: DRAM die series.

Regime	Period	α	SE	n
1	1984–1994	0.39	0.05	4
2	1995–2007	1.15	0.12	3
3	2008–2024	0.38	0.06	3
Full sample	1984–2024	0.66	0.04	9

Sequential sup-F test: break at 1995 significant at 1% ($F = 18.3$); break at 2008 significant at 5% ($F = 9.7$). Critical values from Bai and Perron [26]. Regime 2 estimate ($\alpha = 1.15$) is implausible as a learning parameter and reflects the 1995–2001 DRAM price collapse driven by Asian financial crisis overcapacity and the subsequent demand recovery.

Two features are relevant. First, the bookend regimes yield $\alpha = 0.38$–0.39, consistent with the Irwin and Klenow [15] IV estimate of $\alpha = 0.32$ (SE = 0.05) after accounting for upward OLS bias. Within-regime learning rates are stable *across boom-bust cycles*; the instability is in regime transitions driven by demand-side shocks. Second, the die series instability strengthens the paper's reframing: extrapolating a single α from a 41-year series with two structural breaks is unreliable, which is precisely why the operative curve should be the early-stage packaging process where the learning dynamics are physically interpretable and the demand-side feedback channel is limited.

Bibliography

[1] Acemoglu, D., & Guerrieri, V. (2008). Capital deepening and nonbalanced economic growth. *Journal of Political Economy*, 116(3), 467–498.

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

[3] Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5), 215–227.

[4] Bemmaor, A. C. (1994). Modeling the diffusion of new durable goods: Word-of-mouth effect versus consumer heterogeneity. In G. Laurent, G. L. Lilien, & B. Pras (Eds.), *Research Traditions in Marketing* (pp. 201–229). Kluwer.

[5] Bresnahan, T. F., & Greenstein, S. (1994). The competitive crash in large-scale commercial computing. NBER Working Paper No. 4901.

[6] Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: Engines of growth? *Journal of Econometrics*, 65(1), 83–108.

[7] Carlino, A., Wongel, A., Duan, L., Virguez, E., Davis, S. J., Edwards, M. R., & Caldeira, K. (2025). Variability of technology learning rates. *Advances in Applied Energy*, 20, 100252.

[8] Christensen, C. M. (1997). *The Innovator's Dilemma*. Harvard Business School Press.

[9] David, P. A. (1990). The dynamo and the computer. *American Economic Review*, 80(2), 355–361.

[10] Dodds, P. S., & Watts, D. J. (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21), 218701.

[11] Flamm, K. (1996). *Mismanaged Trade? Strategic Policy and the Semiconductor Industry*. Brookings Institution Press.

[12] Goldberg, P. K., Juhasz, R., Lane, N. J., Lo Forte, G., & Thurk, J. (2024). Industrial policy in the global semiconductor sector. NBER Working Paper No. 32651.

[13] Greenstein, S. (1997). Lock-in and the costs of switching mainframe computer vendors. *Industrial and Corporate Change*, 6(2), 247–273.

[14] Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.

[15] Irwin, D. A., & Klenow, P. J. (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political Economy*, 102(6), 1200–1227.

[16] Levhari, D., & Mirman, L. J. (1980). The great fish war. *Bell Journal of Economics*, 11(1), 322–334.

[17] Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, 29(4), 741–766.

[18] Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy*. Harper & Brothers.

[19] Smirl, J. (2026a). The CES triple role: Superadditivity, correlation robustness, and strategic independence as three views of isoquant curvature. Working Paper.

[20] Smirl, J. (2026b). Complementary heterogeneity in hierarchical economies: CES aggregation, derived architecture, and cross-sector activation in multi-timescale systems. Working Paper.

[21] Stokey, N. L. (1988). Learning by doing and the introduction of new goods. *Journal of Political Economy*, 96(4), 701–717.

[22] Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.

[23] Walter, W. (1998). *Ordinary Differential Equations*. Springer.

[24] Wright, T. P. (1936). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, 3(4), 122–128.

[25] Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton.

[26] Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1–22.

Chapter A

The Mesh Economy

A.1 Introduction

Chapter 4 formalizes endogenous decentralization: concentrated capital investment in centralized AI infrastructure finances the component learning curves—particularly in 3D memory stacking and advanced packaging ($\alpha = 0.23$)—that enable distributed alternatives to replicate datacenter-class inference on consumer hardware. The model's state variable $x(t) = \bar{Q}_{\text{eff}}(\eta(t)) - Q(t)$ measures remaining cumulative production until the crossing threshold. When $x(T^*) = 0$, the basic reproduction number R_0 of the distributed ecosystem crosses unity, and the transition becomes self-sustaining and irreversible.

Chapter 4 ends at the crossing point. This chapter begins there, and follows the distributed ecosystem through two stages: the formation of an organized network (Part I) and the emergence of endogenous capability growth within that network (Part II).

A natural but incorrect conjecture is that post-crossing dynamics are simply "inference runs locally." Isolated devices running local models do not constitute an economic equilibrium. A single consumer device, however capable, cannot match the breadth of a centralized provider serving millions of queries across every domain. The question is whether there exists an organizational form—emergent, not designed—in which distributed devices collectively exceed centralized provision, and whether that form can improve itself without depending entirely on exogenous frontier model releases.

This chapter argues that both answers are affirmative. The organizational form is a *mesh* of heterogeneous specialized agents that self-organize through local interactions into a division of labor whose aggregate capability exceeds any centralized system. And the mesh, once sufficiently large, contains an autocatalytic improvement core that makes capability endogenous—though the rate of improvement converges to the exogenous frontier training rate via a Baumol bottleneck.

The mathematical foundations draw on two results from the preceding chapters. From Chapter 2 (Paper A), the CES curvature parameter $K = (1-\rho)(J-1)/J$ (Definition 3.1) controls the superadditivity of the CES aggregate (Theorem 4.1), the correlation robustness that prevents model collapse (Theorem 5.1), and the strategic independence that stabilizes equilibrium (Theorem 7.1). From Chapter 3 (Paper B), the port topology theorem (Theorem 3.1) establishes the hierarchical architecture, and the activation threshold (Theorem 4.3) provides the spectral condition $\rho(\mathbf{K}) > 1$ under which nontrivial equilibrium exists.

The chapter proceeds as follows. Section A.2 establishes notation and the CES framework. Part I (Sections A.3–A.7) develops network formation: the R_0^{mesh} framework, giant component existence, Potts model crystallization, inverse Bose-Einstein condensation, heterogeneous specialization, knowledge diffusion, the central mesh equilibrium theorem, and

post-crossing dynamics. Part II (Sections A.8–A.13) develops endogenous capability growth: the autocatalytic existence threshold, growth dynamics, the three growth regimes, diversity as collapse protection, and the Baumol bottleneck. Section A.14 connects to the settlement infrastructure of Chapter 6. Section A.15 discusses frameworks considered and rejected. Section A.16 presents falsifiable predictions. Section A.17 concludes.

A.1.1 Relation to the Thesis Framework

This chapter instantiates Levels 2 (Network, years timescale) and 3 (Capability, months timescale) of the four-level hierarchy developed in Chapter 3. Level 2's state variable is heterogeneous AI agent density; Level 3's is training effectiveness.

The division into two parts mirrors the timescale separation that the hierarchy requires. Network formation (Part I) operates on a years timescale: adoption decisions, specialization emergence, and the crystallization of the division of labor unfold over the period 2028–2035. Capability accumulation (Part II) operates on a months timescale: training interactions, autocatalytic improvement, and variety expansion occur within each epoch of the network's evolution. By the timescale separation principle (Chapter 3, Theorem 3.1(iv)), Part I's equilibrium provides the backdrop—the slow manifold—against which Part II's dynamics play out.

The hierarchical ceiling (Chapter 3, Proposition 8.1) binds at both levels. Network size is bounded by hardware capability from Level 1: $F_2 \leq N^*(F_1)$. Capability growth is bounded by network size from Level 2 and, ultimately, by the Baumol bottleneck that anchors the mesh's growth rate to the exogenous frontier training rate determined at Level 1. The ceiling cascade—hardware bounds network, network bounds capability, capability bounds settlement—is a concrete instantiation of the hierarchical structure that Chapter 3 derives from CES geometry.

The activation threshold from Chapter 3 (Theorem 4.3) provides the spectral condition: the mesh equilibrium exists if and only if $\rho(\mathbf{K}) > 1$, where $\mathbf{K}$ is the next-generation matrix. The $R_0^{\text{mesh}} > 1$ condition developed in Part I is the Level 2 component of this spectral threshold. Cross-level amplification (Chapter 3, Section 4.4) implies that even when $R_0^{\text{mesh}} < 1$ in isolation, coupling to Levels 1, 3, and 4 can push the system-wide reproduction number above unity.

A.2 Setup

A.2.1 The CES Aggregate

The aggregate capability of the mesh is governed by the CES production function developed in Chapter 2 (Paper A). For $J \geq 2$ task types with capability levels $C_1, \ldots, C_J$:

$$C_{\text{eff}} = \left(\frac{1}{J} \sum_{j=1}^{J} C_j^{\rho} \right)^{1/\rho}, \quad \rho < 1,\ \rho \neq 0 \tag{A.1}$$

The curvature parameter $K = (1 - \rho)(J - 1)/J$ (Chapter 2, Definition 3.1) controls three properties simultaneously (Chapter 2, Theorem 7.1):

(i) *Superadditivity:* combining diverse capability profiles produces more aggregate capability than the sum of parts, with the gap proportional to K (Chapter 2, Theorem 4.1).

(ii) *Correlation robustness:* the CES nonlinearity extracts idiosyncratic variation that linear aggregates miss, with the bonus proportional to K^2 (Chapter 2, Theorem 5.1).

(iii) *Strategic independence:* the balanced allocation is a Nash equilibrium; coalitions cannot profitably redistribute, with the penalty proportional to K (Chapter 2, Theorem 6.1).

These are not three separate assumptions but three views of the same geometric fact: the curvature of the CES isoquant at symmetric equilibrium. This chapter exploits all three: superadditivity drives the diversity premium in Part I, correlation robustness provides collapse protection in Part II, and strategic independence ensures the mesh equilibrium is robust to manipulation.

A.2.2 Notation

The following notation is used throughout both parts.

Part I: Network Formation

From the Crossing Point to Mesh Dominance

Table A.1: Unified notation.

Symbol	Definition	Source
R_0^{mesh}	Mesh reproduction number $N \cdot \beta \cdot v/D$	Part I
S_∞	Giant component fraction	Part I
N^*	Critical mass for mesh dominance	Part I
C_{eff}	CES aggregate capability	Both
$C_{\text{mesh}}(N)$	Mesh capability at N agents	Part I
C_{cent}	Centralized capability benchmark	Part I
ρ	CES substitution parameter (< 1)	Both
K	Curvature parameter $(1-\rho)(J-1)/J$	Chapter 2
J	Number of task/specialization types	Both
N_{auto}	Autocatalytic existence threshold	Part II
$C_{\max}$	Capability ceiling under saturation	Part II
φ_{eff}	Effective training productivity elasticity	Part II
α_{eff}	Effective external data fraction	Part II
α_{crit}	Model collapse threshold	Part II
β_{auto}	Autocatalytic fraction of training	Part II
h	Training saturation parameter	Part II
δ	Model depreciation/obsolescence rate	Part II
g_Z	Exogenous frontier training growth rate	Part II

At $t = T^*$, the "mesh" consists of a population of heterogeneous devices capable of running inference workloads locally but not yet organized into a connected network with routing, specialization, or coordination. The number of capable devices $N(T^*)$ is positive but below critical mass. Chapter 4's $R_0 > 1$ condition ensures that the population grows; the sections that follow characterize what it grows *into*.

A.3 The R_0^{mesh} Framework

The R_0 framework from Chapter 4 provides the boundary condition for the mesh. The mesh reproduction number is:

$$R_0^{\text{mesh}} = N \cdot \beta \cdot v/D \tag{A.2}$$

where N is the number of active nodes, β is the connection probability per node pair, v is the expected value per interaction, and D is the attrition rate (encompassing coordination friction κ and churn μ from Chapter 4). The mesh is self-sustaining when $R_0^{\text{mesh}} > 1$. This is the Level 2 component of the activation threshold from Chapter 3 (Theorem 4.3).

A.3.1 Giant Component Existence

The fraction of nodes belonging to the giant connected component satisfies:

$$S_\infty = 1 - \exp(-R_0^{\text{mesh}} \cdot S_\infty) \tag{A.3}$$

Proposition A.3.1 (Giant Component Existence and Uniqueness)**.** *Equation* (A.3) *has:*

(a) the trivial solution $S_\infty = 0$ *for all* R_0^{mesh}*;*

(b) a unique positive solution $S_\infty^* \in (0, 1)$ *if and only if* $R_0^{mesh} > 1$*;*

(c) S_∞^* *is locally asymptotically stable as a fixed point of the iteration* $S_\infty \mapsto 1-\exp(-R_0^{mesh} \cdot S_\infty)$.

Proof. Define $g(s) = 1 - \exp(-R_0^{\text{mesh}} \cdot s) - s$. At $s = 0$: $g(0) = 0$ and $g'(0) = R_0^{\text{mesh}} - 1 > 0$ when $R_0^{\text{mesh}} > 1$, so g is initially increasing. At $s = 1$: $g(1) = -\exp(-R_0^{\text{mesh}}) < 0$. By the intermediate value theorem, g has a root $S_\infty^* \in (0, 1)$.

For uniqueness: $g''(s) = -{R_0^{\text{mesh}}}^2 \exp(-R_0^{\text{mesh}} s) < 0$, so g is strictly concave. A strictly concave function crossing zero from above can have at most one positive root.

For stability: the iteration map $h(s) = 1-\exp(-R_0^{\text{mesh}} s)$ satisfies $h'(S_\infty^*) = R_0^{\text{mesh}} \exp(-R_0^{\text{mesh}} S_\infty^*) = R_0^{\text{mesh}}(1 - S_\infty^*)$. At the positive fixed point with $R_0^{\text{mesh}} > 1$, $0 < h'(S_\infty^*) < 1$ (since $S_\infty^* > 0$ implies $1 - S_\infty^* < 1$ and $R_0^{\text{mesh}}(1 - S_\infty^*) < 1$ by concavity), confirming local asymptotic stability by the contraction mapping principle. □

A.3.2 The Fortuin-Kasteleyn Unification

The giant component result establishes *connectivity*. The question of whether the mesh develops *specialization*—a division of labor among heterogeneous agents—might appear to require a separate model. The Fortuin-Kasteleyn [19] representation reveals that these are the same mathematical object.

The partition function of the q-state Potts model on graph $G = (V, E)$ has the exact cluster expansion:

$$Z_{\text{Potts}} = \sum_{A \subseteq E} p^{|A|}(1 - p)^{|E|-|A|} \cdot q^{c(A)} \tag{A.4}$$

where $p = 1 - e^{-\beta_T J_c}$ is the bond occupation probability (with β_T the inverse temperature and J_c the coupling), A is a subset of edges, and $c(A)$ is the number of connected components in the subgraph (V, A).

At $q = 1$, this reduces to the generating function for bond percolation. The Potts model at general q describes a system in which nodes adopt one of q states (specializations) and prefer

to match their neighbors. The FK representation shows that percolation and specialization are controlled by the same cluster structure evaluated at different q.

Proposition A.3.2 (First-Order Crystallization)**.** *For $q > 2$ specialization types on a graph with mean degree $\langle k \rangle > 1$, the specialization transition is first-order: the order parameter (fraction of nodes in the dominant specialization cluster) jumps discontinuously from zero to a positive value at the critical coupling $\beta_T J_c = \beta_c(q)$.*

This is a classical result in statistical mechanics [7, 44]. The economic content is that the mesh does not form gradually. When the parameter p (interpretable as the probability that two neighboring agents successfully coordinate on a task division) crosses the threshold, division of labor crystallizes abruptly. Early mesh growth appears stochastic and fragile; the crystallization event is sudden.

For $q = 2$ (two specialization types), the transition is second-order (continuous)—the standard Ising universality class. This means the first-order prediction is specific to settings with $q \geq 3$ distinct specialization roles, which is the empirically relevant case for AI inference (coding, creative writing, mathematical reasoning, multimodal processing, domain-specific knowledge, real-time translation, etc.).

A.3.3 Inverse Bose-Einstein Condensation

The Bianconi-Barabási [4] model assigns each node i a fitness η_i drawn from distribution $\rho(\eta)$. The degree of node i evolves as:

$$k_i(t) \sim \left(\frac{t}{t_i}\right)^{\eta_i/C} \tag{A.5}$$

where t_i is the node's arrival time and C satisfies the self-consistency equation:

$$\int \frac{\rho(\eta)}{C/\eta - 1}\, d\eta = 1 \tag{A.6}$$

This has the identical mathematical structure to the Bose gas number equation, with C playing the role of the fugacity.

The phase structure depends on $\rho(\eta)$ near its maximum $\eta_{\max}$. If $\rho(\eta) \sim (\eta_{\max} - \eta)^{\alpha_B}$ as $\eta \to \eta_{\max}$:

- $\alpha_B \leq 0$ (sharply peaked): *Bose-Einstein condensation.* The single fittest node captures a macroscopic fraction of all connections. This is the centralized equilibrium—AWS, Google Cloud, and Azure dominate because they are the only nodes with high fitness.

- $\alpha_B > 0$ (broad distribution): *Fit-get-rich* phase. Many nodes share traffic proportional to their fitness. No single node dominates. This is the mesh equilibrium.

Proposition A.3.3 (Learning-Curve-Driven Phase Transition)**.** *Let the technology parameter $\theta(t)$ index the packaging learning curve output from Chapter 4. Suppose the fitness of an edge device of type j is $\eta_j(\theta) = \eta_j^0 + g_j(\theta)$, where g_j is increasing in θ and $g_j(0) = 0$. If $\theta(0)$ produces a fitness distribution $\rho_0(\eta)$ with $\alpha_B \leq 0$, and $\theta(\bar{t})$ produces $\rho_{\bar{t}}(\eta)$ with $\alpha_B > 0$ for some finite $\bar{t}$, then the system undergoes a BEC-to-FGR phase transition at the critical θ^* where α_B crosses zero. This is inverse Bose-Einstein condensation: the centralized condensate dissolves.*

Proof. The learning curve increases the fitness of previously low-fitness edge devices. Initially, only datacenter nodes have η near $\eta_{\max}$, so $\rho(\eta)$ is sharply peaked at the maximum—the BEC phase. As θ increases, the support of ρ broadens: more device types achieve fitness levels closer to $\eta_{\max}$. The exponent α_B characterizing the behavior of ρ near $\eta_{\max}$ transitions from ≤ 0 to > 0 at $\theta = \theta^*$. By the Bianconi-Barabási classification, this is the BEC phase boundary.

The mapping to Chapter 4 is direct: θ^* corresponds to $x(t) = 0$. At T^*, the fitness distribution has broadened sufficiently that the centralized condensate—the macroscopic concentration of traffic at a single node—dissolves. Traffic distributes across the mesh proportional to fitness. □

Remark A.3.4. The BEC framework describes the *competitive dynamics* of traffic allocation on the network. It does not describe the physical connectivity (percolation, Section A.3.1) or the specialization structure (CES aggregation, Section A.4). The layers compose: percolation ensures the mesh is connected, BEC dynamics govern how traffic flows within it, and CES aggregation determines whether the mesh's collective capability exceeds the centralized alternative.

A.4 Heterogeneous Specialization

A.4.1 Agent Capabilities and CES Aggregation

Each agent $i \in \{1, \ldots, N\}$ has a capability vector $\mathbf{c}_i = (c_{i1}, \ldots, c_{iJ})$ across J task types. The effective capability of the mesh for task type j is the sum of agent capabilities: $C_j = \sum_i c_{ij}$. The aggregate mesh capability is the CES function (A.1).

The parameter $\rho < 1$ implies imperfect substitutability across task types. Because task types are complements, the aggregate rewards *diversity*. Two agents with different specializations contribute more to C_{eff} than two identical agents. This is the superadditivity result from Chapter 2 (Theorem 4.1) applied to the mesh's capability vectors.

Lemma A.4.1 (Diversity Premium)**.** *Fix total capability $\bar{C} = \sum_j C_j$. For $\rho < 1$, C_{eff} is maximized when $C_j = \bar{C}/J$ for all j (equal coverage across task types). More precisely:*

$$C_{\text{eff}}\big|_{\text{equal}} = J^{(1-\rho)/\rho} \cdot C_{\text{eff}}\big|_{\text{concentrated}} \tag{A.7}$$

where the concentrated case places all capability in a single task type. The diversity premium $J^{(1-\rho)/\rho}$ is increasing in J and decreasing in ρ.

Proof. With equal allocation: $C_{\text{eff}} = (J \cdot (\bar{C}/J)^{\rho})^{1/\rho} = J^{1/\rho - 1} \cdot \bar{C}$. With concentration in one type: $C_{\text{eff}} = \bar{C}$. The ratio is $J^{(1-\rho)/\rho}$, which exceeds 1 for $J \geq 2$ and $\rho < 1$. □

This is the Becker-Murphy [8] division of labor result in CES form, and a direct application of Chapter 2's superadditivity theorem (Theorem 4.1). The mesh's advantage over centralized provision does not come from superior individual capability—each edge device is weaker than the datacenter—but from the breadth of specialized coverage that heterogeneous agents collectively provide.

A.4.2 Centralized Capability Benchmark

A centralized provider operates M identical high-capability units, each with capability $\bar{c}$ spread uniformly across all J task types: $c_j^{\text{cent}} = \bar{c}/J$ per unit. Total centralized capability for task j is $C_j^{\text{cent}} = M\bar{c}/J$, giving:

$$C_{\text{cent}} = \left(J \cdot \left(\frac{M\bar{c}}{J} \right)^{\rho} \right)^{1/\rho} = J^{(1-\rho)/\rho} \cdot M\bar{c} \tag{A.8}$$

The centralized provider has fixed capacity $M\bar{c}$ (determined by datacenter investment). The mesh's aggregate capability grows with N and with the diversity of specialists.

A.4.3 Specialization Dynamics: The Fixed Response Threshold Model

Agents do not arrive pre-specialized. Specialization emerges endogenously through local interactions. The mechanism follows the Bonabeau-Theraulaz [10] fixed response threshold model, originally developed for division of labor in social insect colonies.

Agent i has a vector of response thresholds $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iJ})$ for each task type. When demand signal s_j for task j arrives, agent i performs the task with probability:

$$P_{ij}(s_j) = \frac{s_j^n}{s_j^n + \theta_{ij}^n} \tag{A.9}$$

where $n \geq 2$ is a steepness parameter.

Thresholds adapt through reinforcement:

$$\dot{\theta}_{ij} = -\xi \cdot \mathbf{1}[i \text{ performs task } j] + \varphi_t \cdot \mathbf{1}[i \text{ does not perform task } j] \tag{A.10}$$

where $\xi > 0$ is the reinforcement rate (performing a task lowers the threshold, increasing future responsiveness) and $\varphi_t > 0$ is the decay rate (not performing a task raises the threshold).

Proposition A.4.2 (Emergent Specialization)**.** *Under the threshold dynamics* (A.10) *with heterogeneous initial thresholds* $\theta_{ij}(0)$*, agents self-sort into specialist roles: for each agent* i*, there exists* $j^*(i) = \arg\max_j c_{ij}(t)$ *such that* $c_{ij^*}(t) \to \bar{c}_i$ *and* $c_{ij}(t) \to 0$ *for* $j \neq j^*$ *as* $t \to \infty$*, where* $\bar{c}_i$ *is agent* i*'s maximum achievable capability.*

The proof follows from the reinforcement dynamics: an agent that performs task j frequently sees θ_{ij} decrease, making it more responsive to future demand for j, which further increases frequency of performance—a positive feedback loop. Simultaneously, thresholds for other tasks rise. The dynamics converge to a fixed point where each agent responds primarily to one task type. This is the Becker-Murphy division of labor emerging from local interactions without central coordination.

Remark A.4.3 (Connection to Potts Crystallization). The specialization dynamics of Proposition A.4.2 are the micro-level mechanism underlying the Potts model crystallization of Proposition A.3.2. The Potts "state" of each node is its dominant specialization $j^*(i)$. The "coupling" J_c in the Potts Hamiltonian corresponds to the task-sharing benefit between neighboring agents with compatible specializations. The first-order crystallization at $q > 2$ means that when the number of specialization types exceeds two, the transition from unspecialized to specialized is abrupt—consistent with the reinforcement dynamics exhibiting a bifurcation from mixed to specialized response profiles.

A.5 Knowledge Diffusion

A.5.1 Laplacian Dynamics

Let $\mathbf{u}(t) \in \mathbb{R}^N$ represent the knowledge state of each node (e.g., model weights, fine-tuning updates, or capability parameters). Knowledge diffusion on the network follows:

$$\frac{\partial \mathbf{u}}{\partial t} = -L \cdot \mathbf{u} \tag{A.11}$$

where $L = D_{\text{deg}} - A$ is the graph Laplacian, with D_{deg} the diagonal degree matrix and A the adjacency matrix. The convergence rate to the consensus state is governed by $\lambda_2(L)$, the Fiedler eigenvalue (algebraic connectivity).

Lemma A.5.1 (Convergence Rate)**.** *The knowledge state* $\mathbf{u}(t)$ *converges to the average* $\bar{u} = N^{-1} \sum_i u_i(0)$ *at rate* $\lambda_2(L)$*:*

$$\|\mathbf{u}(t) - \bar{u}\mathbf{1}\|_2 \leq \|\mathbf{u}(0) - \bar{u}\mathbf{1}\|_2 \cdot e^{-\lambda_2(L)t} \tag{A.12}$$

For connected graphs, $\lambda_2(L) > 0$*.*

A.5.2 Bandwidth Scaling

The total bandwidth available for knowledge diffusion in the mesh scales as $B_{\text{mesh}} = O(N \cdot \langle k \rangle)$, where $\langle k \rangle$ is the mean degree. The centralized hub has fixed bandwidth B_{hub} determined by datacenter interconnect capacity. Once:

$$N \cdot \langle k \rangle > B_{\text{hub}} \tag{A.13}$$

the mesh serves more total queries per unit time than the centralized provider.

A.5.3 Vanishing Epidemic Threshold on Scale-Free Networks

Pastor-Satorras and Vespignani [35] established that the SIS epidemic threshold on networks with degree distribution $P(k) \sim k^{-\gamma}$ is:

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{A.14}$$

For scale-free networks with $\gamma \leq 3$: $\langle k^2 \rangle$ diverges in the thermodynamic limit, so $\lambda_c \to 0$.

Proposition A.5.2 (Self-Sustaining Knowledge Propagation)**.** *If the mesh has a scale-free degree distribution with $\gamma \leq 3$—which MoE routing produces endogenously through preferential specialization—then any nonzero rate of knowledge sharing sustains itself indefinitely. The topology ensures propagation without requiring a minimum transmission rate.*

The economic content is that knowledge diffusion need not be modeled as a separate mechanism with its own threshold. Once the mesh achieves a fat-tailed degree distribution (which the specialization dynamics of Section A.4.3 produce through preferential attachment to high-quality specialists), capability propagation is guaranteed. Hub agents—the most capable specialists—emerge endogenously and serve as conduits for knowledge transfer.

A.5.4 Combined Dynamics

The three layers interact as follows. Percolation ensures the mesh is connected ($S_\infty > 0$). CES aggregation ensures agents specialize (C_{eff} grows with diversity). Knowledge diffusion ensures information propagates (the Fiedler eigenvalue is positive, and on scale-free topologies, propagation is self-sustaining). The combined effect is a mesh whose aggregate capability $C_{\text{mesh}}(N)$ is superlinear in N over the relevant range, because additional diverse specialists both increase the CES aggregate and increase the rate at which existing knowledge diffuses to new entrants.

A.6 The Central Theorem: Mesh Equilibrium

Theorem A.6.1 (Mesh Equilibrium Existence, Uniqueness, and Dominance)**.** *For $R_0^{mesh} > 1$ and $\rho < 1$, there exists a finite N^* such that for all $N > N^*$:*

(a) The mesh equilibrium exists: a positive fraction $S_\infty^* > 0$ of agents form a connected component with specialized roles covering all J task types.

(b) The mesh equilibrium is unique among equilibria with $S_\infty > 0$.

(c) The mesh equilibrium is locally asymptotically stable.

(d) $C_{mesh}(N) > C_{cent}$: the mesh's aggregate capability exceeds centralized provision.

(e) N^* is decreasing in the diversity of the agent population, measured by the entropy of the fitness distribution $H(\rho) = -\int \rho(\eta) \ln \rho(\eta)\, d\eta$.

Proof. The proof proceeds in five steps.

Step 1: Existence via percolation (Proposition A.3.1). For $R_0^{\text{mesh}} > 1$, equation (A.3) has a unique positive solution S_∞^*. The giant component contains $S_\infty^* \cdot N$ agents. For N sufficiently large, this exceeds the minimum number of agents required to cover all J task types (at least one specialist per type), establishing existence.

Step 2: Uniqueness via supermodularity. The mesh participation game—in which each agent decides whether to join the mesh and which task type to specialize in—is a supermodular game. Agent i's payoff from joining increases when more agents join (network effect via the CES aggregate and knowledge diffusion) and when agents specialize in complementary types (CES complementarity with $\rho < 1$). By Tarski's [40] fixed point theorem, the game has a greatest and least equilibrium. By the strict concavity of the CES function in each C_j, the greatest equilibrium is unique among equilibria with $S_\infty > 0$: any equilibrium with different specialization allocations is payoff-dominated by the efficient allocation.

Step 3: Stability via Lyapunov analysis. Consider the Lyapunov function $V = -C_{\text{eff}}(N) + \sum_j \phi_j(C_j)$ where ϕ_j is the potential function of the specialization dynamics. At the equilibrium, $\dot{V} \leq 0$ with equality only at the fixed point, by the dissipative property of the threshold dynamics (A.10) and the concavity of C_{eff}. The equilibrium is locally asymptotically stable by LaSalle's invariance principle.

Step 4: Capability dominance via CES growth. The mesh capability is:

$$C_{\text{mesh}}(N) = \left(\sum_{j=1}^{J} \left(\sum_{i \in \text{mesh}} c_{ij} \right)^{\rho} \right)^{1/\rho} \tag{A.15}$$

Under specialization, each agent i concentrates capability in task $j^*(i)$, so $C_j \approx |\{i : j^*(i) = j\}| \cdot \bar{c}$ where $\bar{c}$ is the mean specialist capability. If agents distribute approximately uniformly across J types, then $C_j \approx S_\infty^* N \bar{c}/J$ and:

$$C_{\text{mesh}}(N) \approx J^{(1-\rho)/\rho} \cdot S_\infty^* N \bar{c}/J = J^{1/\rho - 2} \cdot S_\infty^* N \bar{c} \tag{A.16}$$

This exceeds $C_{\text{cent}} = J^{(1-\rho)/\rho} \cdot M \bar{c}_{\text{cent}}$ when $N > N^* \equiv M \bar{c}_{\text{cent}}/(S_\infty^* \bar{c})$. Since M and $\bar{c}_{\text{cent}}$ are fixed, N^* is finite.

Step 5: N^* decreasing in diversity. Higher entropy $H(\rho)$ of the fitness distribution implies broader coverage of the capability space for any given N. The CES diversity premium (Lemma A.4.1) increases with the number of effectively distinct specialization types. Higher diversity means fewer agents are needed to achieve full task coverage, reducing N^*. □

Corollary A.6.2 (Centralized Market Share Decline)**.** *For $N > N^*$, the centralized provider's market share is strictly decreasing in N. In the Bianconi-Barabási framework, the centralized*

"condensate" fraction declines continuously as the fitness distribution broadens, transitioning from the BEC phase (macroscopic condensate) to the FGR phase (distributed traffic).

Remark A.6.3 (Self-Consistency Across Fields). The critical condition $R_0^{\text{mesh}} = 1$ in Theorem A.6.1 is a universal transcritical bifurcation. Table A.2 shows the correspondence across fields. All produce the same self-consistency equation $m = f(m; \lambda)$ with the identical phase structure.

Table A.2: Universal self-consistency across fields.

Field	Order parameter m	Control λ	Critical condition	Source
Percolation	Giant component S_∞	Mean degree $\langle k \rangle$	$\langle k \rangle = 1$	Erdős-Rényi [17]
Epidemiology	Infected fraction	R_0	$R_0 = 1$	Kermack-McKendrick [45]
Ising/Potts	Magnetization	$\beta_T J_c q$	$\beta_T J_c q = 1$	Ising [46]; Potts [36]
Ecology	Productivity	Species richness S	Min S for resilience	Loreau-Hector [32]
Network econ.	Market share	Transaction benefit	Critical liquidity	Katz-Shapiro [28]
This chapter	**Mesh fraction S_∞**	$\boldsymbol{R_0^{\text{mesh}}}$	$\boldsymbol{R_0^{\text{mesh}} = 1}$	—

A.7 Post-Crossing Dynamics

The path from $x(t) = 0$ to mesh dominance is not monotonic. Three phases, distinguished by the value of R_0^{mesh} and the state of the specialization structure, characterize the transition.

A.7.1 Phase 1: Nucleation ($R_0^{\text{mesh}} \approx 1$)

Immediately after crossing, R_0^{mesh} is only marginally above unity. The giant component is small ($S_\infty^* \approx 0$ for R_0^{mesh} near 1, since $S_\infty^* \sim 2(R_0^{\text{mesh}} - 1)/{R_0^{\text{mesh}}}^2$ to leading order). Growth is slow and stochastic. Small specialist clusters form around high-fitness agents—the enthusiast-tier hardware users running quantized models—but the clusters are fragile. Exogenous shocks (model-release events, API pricing changes, hardware supply disruptions) can temporarily push R_0^{mesh} below unity, collapsing nascent clusters.

The mesh first achieves capability dominance on the *long tail* of niche queries that centralized systems underserve. Centralized providers optimize for the highest-volume query types (general chat, code generation, summarization). Specialized queries—domain-specific technical reasoning, low-resource language translation, real-time edge processing for robotics—are underserved because the revenue per query does not justify dedicated model fine-tuning. The mesh's heterogeneous agents, each fine-tuned for a niche, collectively cover the long tail.

This is the Christensen [13] pattern: disruption begins in markets the incumbent rationally ignores.

A.7.2 Phase 2: Rapid Growth ($R_0^{\text{mesh}} \gg 1$)

As additional device types become inference-capable (driven by the continuing packaging learning curve) and coordination infrastructure matures (κ declines), R_0^{mesh} accelerates well above unity. Network effects dominate. Each new specialist joining the mesh increases C_{eff} superlinearly (by the CES diversity premium) and increases the Fiedler eigenvalue $\lambda_2(L)$ (by adding connectivity), which accelerates knowledge diffusion to subsequent entrants.

In this phase, the Potts crystallization occurs: the division of labor among mesh agents transitions from fragmented proto-specialization to a structured, self-reinforcing configuration. By Proposition A.3.2, this transition is first-order for $q > 2$ specialization types. The crystallization is observable as a sudden increase in the concentration of agent capabilities around distinct specialization types, accompanied by the emergence of routing hub agents that handle disproportionate query traffic.

Centralized providers lose market share in progressively more mainstream query types, beginning with the long-tail niches of Phase 1 and expanding to higher-volume categories as mesh coverage broadens.

A.7.3 Phase 3: Maturity

Growth saturates as the mesh's J task types are fully covered. The CES aggregate C_{eff} approaches its maximum for the given device population. Competition shifts from mesh growth to mesh composition: which specialists are included, the quality of their fine-tuning, and the efficiency of the routing layer.

Centralized providers retain structural advantage in two domains that the mesh cannot replicate:

(i) *Frontier model training:* As established in Chapter 4, training requires tightly synchronized GPU clusters at scales incompatible with distributed architecture. The mesh depends on centralized training for the base models it fine-tunes.

(ii) *Capabilities requiring single-device scale beyond any edge device:* Tasks requiring the full activation of 1T+ parameter dense models in a single forward pass remain centralized. This is the inference analog of the training constraint, but applies to a shrinking fraction of queries as MoE architectures reduce the active parameter requirement.

The mature equilibrium is coexistence: centralized providers dominate training and frontier-capability inference; the mesh dominates the long tail, latency-sensitive applications, and the broad middle of the query distribution where specialized, fine-tuned models outperform general-purpose frontier models.

At maturity, Part I's analysis reaches a terminal state—the mesh exists, is stable, and dominates, but its capability is bounded by the fixed-capability assumption. Part II removes that assumption.

Part II: Capability Growth

Endogenous Improvement and the Baumol Bottleneck

A.8 The Fixed-Capability Assumption Relaxed

Part I assumes fixed capabilities: each agent's capability vector $\mathbf{c}_i$ redistributes through specialization but does not grow in total magnitude. The CES aggregate $C_{\text{eff}}(N)$ increases only by adding diverse agents. Knowledge diffusion ($\partial\mathbf{u}/\partial t = -L\mathbf{u}$) equalizes existing knowledge across nodes—it does not create new capability.

This section removes that assumption. Three mechanisms make capability endogenous:

(i) *Autocatalytic capability growth:* Training agents within the mesh improve other agents, and the improved agents can in turn improve others.

(ii) *Self-referential learning:* The mesh's operation generates training data—queries, responses, user feedback, inter-agent evaluations—that can be used to improve mesh agents.

(iii) *Endogenous variety expansion:* The mesh modifies its own composition by spawning new specialization types in response to unmet demand signals.

When $\mathbf{c}_i$ becomes a dynamical variable coupled to the mesh's operation, the central question changes from *whether* the mesh dominates (answered in Part I) to *how fast* capability grows. By the timescale separation of the thesis framework (Chapter 3, Theorem 3.1(iv)), Part I's equilibrium—the network structure, specialization allocation, and degree distribution—provides the slow backdrop against which capability dynamics unfold. The network is approximately static on the months timescale of capability growth.

A.9 The Autocatalytic Existence Threshold

Does the mesh contain a self-sustaining improvement core—a subset of agents whose training interactions can maintain and improve capabilities given only exogenous base models as external input?

A.9.1 Training Operations as a Reaction System

Definition A.9.1 (Training Operation)**.** *A training operation is a tuple* $r = (I_r, K_r, O_r)$ *where:*

- $I_r \subseteq \{1, \ldots, J\}$ *is the set of input capability types consumed or modified;*
- $K_r \subseteq \{1, \ldots, J\}$ *is the set of catalyst capability types required to execute the operation (not consumed);*
- $O_r \subseteq \{1, \ldots, J\}$ *is the set of output capability types produced or enhanced.*

The catalyst distinction is critical. A training agent that orchestrates fine-tuning of a medical reasoning specialist requires its own orchestration capability (K_r) but is not consumed by the process. It can catalyze multiple training operations. This is the biochemical analogy that motivates the autocatalytic framework: enzymes catalyze reactions without being consumed.

A.9.2 The Food Set

Definition A.9.2 (Food Set)**.** *The food set* $F \subset \{1, \ldots, J\}$ *is the set of capability types available exogenously from centralized training. For each* $j \in F$*, base model capability of type* j *is available without requiring any mesh training operation. The food set is determined by frontier model releases from centralized providers and is exogenous to the mesh.*

The food set corresponds to the training persistence assumption from Part I: frontier model training remains centralized. Base models (GPT-class, Claude-class, Gemini-class) are the "raw materials" that the mesh fine-tunes, adapts, and combines. The food set grows exogenously at a rate determined by centralized infrastructure investment—the rate that will emerge as the Baumol bottleneck in Section A.13.

A.9.3 RAF Sets in the Mesh

Definition A.9.3 (Reflexively Autocatalytic and Food-generated (RAF) Set)**.** *Following Hordijk and Steel [23], a set $\mathcal{R}$ of training operations is a RAF set if:*

(a) Reflexively Autocatalytic (RA): *Every training operation $r \in \mathcal{R}$ is catalyzed by at least one capability type that is either in the food set F or is produced by some other operation in $\mathcal{R}$.*

(b) Food-generated (F): *Every input capability type of every operation $r \in \mathcal{R}$ can be constructed from the food set F by successive application of operations in $\mathcal{R}$.*

A RAF set is self-sustaining: given only exogenous base models (the food set), the mesh can maintain and improve all capabilities involved in the set through its own internal training operations.

A.9.4 Existence Threshold

Proposition A.9.4 (Autocatalytic Existence Threshold)**.** *Let $J(N)$ denote the number of distinct capability types present in a mesh of N agents, and let $p(N) = 1 - (1 - \beta_t)^N$ be the probability that a given capability type is available as a catalyst in the mesh, where β_t is the per-agent probability of possessing a given catalyst capability. There exists a critical mesh size N_{auto} such that for $N > N_{auto}$, the mesh contains a RAF set with probability approaching unity. Moreover:*

$$N_{auto} = O\left(\frac{\ln |\mathcal{R}|}{\beta_t}\right) \tag{A.17}$$

where $|\mathcal{R}|$ is the total number of potential training operations. The threshold N_{auto} scales logarithmically with system complexity.

Proof. The result follows from the probabilistic analysis of RAF sets by Hordijk and Steel [23] (Theorem 2). In their framework, a random catalytic reaction system with n molecule types, r reactions, and catalysis probability p per type-reaction pair contains a RAF set with probability approaching 1 when p exceeds $1/n$. In our setting, $n = J$ capability types and the effective catalysis probability per type is $p(N) = 1 - (1 - \beta_t)^N$.

The condition $p(N) > 1/J$ is satisfied when:

$$1 - (1 - \beta_t)^N > \frac{1}{J} \tag{A.18}$$

which gives $N > \ln(1 - 1/J)/\ln(1 - \beta_t) \approx (1/J)/\beta_t = 1/(J\beta_t)$ for small β_t. Since J grows

at most polynomially in the number of potential operations $|\mathcal{R}|$, the threshold scales as $N_{\text{auto}} = O(\ln |\mathcal{R}|/\beta_t)$.

The logarithmic scaling arises because adding agents to the mesh increases the probability of *every* catalyst assignment simultaneously. Each new agent with a novel catalyst capability unlocks multiple potential training operations. □

Remark A.9.5 (Relationship to N^*). The autocatalytic threshold N_{auto} and the critical mass N^* from Theorem A.6.1 are distinct. N^* is the mesh size at which collective capability exceeds centralized provision (a *static* comparison). N_{auto} is the mesh size at which self-sustaining capability improvement becomes possible (a *dynamic* property). Generically, $N_{\text{auto}} > N^*$: the mesh can be collectively capable before it is self-improving. The gap $N_{\text{auto}} - N^*$ represents the period during which the mesh exceeds centralized inference but depends entirely on exogenous base model releases for capability growth.

A.9.5 Autocatalytic Core Dynamics: The Jain-Krishna Process

Once a RAF set exists, the mesh's autocatalytic core evolves through the adaptive network dynamics of Jain and Krishna [24, 25]. Define the catalytic matrix $\mathbf{M}$ where $M_{ij} = 1$ if capability type j catalyzes the improvement of capability type i. The growth rate of capability type i in the mesh is governed by:

$$\dot{c}_i = c_i \left(\sum_j M_{ij} c_j - \phi_0 \right) \tag{A.19}$$

where $\phi_0 = N^{-1} \sum_i c_i \sum_j M_{ij} c_j$ is a dilution term ensuring bounded growth.

Proposition A.9.6 (Perron-Frobenius Selection)**.** *The long-run composition of the autocatalytic core is determined by the leading eigenvector of the catalytic matrix* $\mathbf{M}$*. Capability types with large components in the Perron-Frobenius eigenvector* $\mathbf{v}_1$ *of* $\mathbf{M}$ *dominate; types with small components go extinct. The leading eigenvalue* $\lambda_1(\mathbf{M})$ *determines whether the autocatalytic core expands or contracts relative to the rest of the mesh.*

Proof. Equation (A.19) is a replicator equation on the simplex of capability-type shares. The fixed point analysis follows from the standard replicator dynamics result [22]: the dynamics converge to a state where surviving types have equal fitness, and the surviving set corresponds to the support of the Perron-Frobenius eigenvector of $\mathbf{M}$. Types i with $v_{1,i} > 0$ persist; types with $v_{1,i} = 0$ go extinct.

The eigenvalue $\lambda_1(\mathbf{M})$ determines the growth rate of the autocatalytic core because the aggregate fitness of the surviving set equals $\lambda_1(\mathbf{M})$. When $\lambda_1(\mathbf{M}) > \phi_0$, the core expands as a fraction of total mesh activity. □

The Jain-Krishna dynamics predict a specific temporal pattern: the autocatalytic core does not grow smoothly. Instead, it undergoes a series of reorganization cascades—periods of stasis punctuated by rapid restructuring events in which poorly connected capability types are replaced by types with stronger catalytic linkages. Each cascade increases the leading eigenvalue $\lambda_1(\mathbf{M})$, producing a staircase pattern of increasing autocatalytic efficiency.

A.10 Growth Dynamics: The Central Model

Having established that the mesh contains a self-sustaining improvement core (Section A.9), we now characterize the rate at which capability grows.

A.10.1 The Improvement Function

Let $f \in (0, 1)$ denote the fraction of mesh capability devoted to self-improvement (training operations) rather than serving external queries (inference). The mesh's aggregate capability evolves according to:

$$\frac{d}{dt}C_{\text{eff}} = g(f \cdot C_{\text{eff}},\ J(t),\ \alpha(t)) - \delta \cdot C_{\text{eff}} \tag{A.20}$$

where g is the improvement function, $J(t)$ is the number of effective specialization types, $\alpha(t)$ is the fraction of training signal from external sources, and $\delta > 0$ is the depreciation rate capturing model obsolescence.

The improvement function g has three arguments because the three mechanisms of endogenous capability contribute separately:

(i) $f \cdot C_{\text{eff}}$: *Autocatalytic training.* The total capability devoted to self-improvement. More capable training agents produce better improvements.

(ii) $J(t)$: *Variety expansion.* New specialization types expand the capability space. By the diversity premium (Lemma A.4.1), adding a new task type increases C_{eff} superlinearly.

(iii) $\alpha(t)$: *Data quality.* The fraction α determines whether self-referential learning improves or degrades capability. Below the critical threshold α_{crit}, training on internally generated data causes model collapse [39].

A.10.2 The Semi-Endogenous Growth Formulation

Following Jones [26, 27], specify the improvement function as:

$$g(f \cdot C_{\text{eff}}, J, \alpha) = \delta_g \cdot (f \cdot C_{\text{eff}})^{\lambda} \cdot C_{\text{eff}}^{\varphi-1} \cdot J^{\gamma_J} \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] \tag{A.21}$$

where $\delta_g > 0$ is a productivity parameter, $\lambda \in (0, 1]$ is the duplication parameter, φ is the training productivity elasticity, $\gamma_J > 0$ governs the contribution of variety expansion, and $\mathbf{1}[\alpha > \alpha_{\text{crit}}]$ is the model collapse indicator.

Letting $C \equiv C_{\text{eff}}$ for notational compactness:

$$\dot{C} = \delta_g \cdot f^{\lambda} \cdot C^{\lambda+\varphi-1} \cdot J(t)^{\gamma_J} \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] - \delta \cdot C \tag{A.22}$$

The growth rate of capability is:

$$g_C \equiv \frac{\dot{C}}{C} = \delta_g \cdot f^{\lambda} \cdot C^{\lambda+\varphi-2} \cdot J(t)^{\gamma_J} \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] - \delta \tag{A.23}$$

A.10.3 Deriving the Mesh's Effective φ

The parameter φ measures the elasticity of new capability production with respect to the existing stock. Empirical evidence strongly suggests $\varphi < 1$ for individual training interactions: ideas are getting harder to find [9]. The mesh, however, has internal structure that amplifies the effective φ through autocatalytic coupling.

Proposition A.10.1 (Effective Training Productivity)**.** *Let $\varphi_0 < 1$ be the raw training productivity elasticity for a single training interaction. Let $\beta_{\text{auto}} \in [0, 1)$ be the fraction of the training improvement process that can be automated by the mesh's own training agents. Then the mesh's effective training productivity elasticity is:*

$$\varphi_{\mathit{eff}} = \frac{\varphi_0}{1 - \beta_{\text{auto}} \cdot \varphi_0} \tag{A.24}$$

This exceeds φ_0 for all $\beta_{\text{auto}} > 0$, and $\varphi_{\mathit{eff}} \geq 1$ when $\beta_{\text{auto}} \geq (1 - \varphi_0)/\varphi_0$.

Proof. Following the Aghion-Jones-Jones [3] framework, decompose the training improvement process into a continuum of subtasks indexed on $[0, 1]$. A fraction β_{auto} of subtasks are automated by mesh agents (whose productivity scales with C^{φ_0}), and the remaining fraction $1 - \beta_{\text{auto}}$ requires exogenous input. The effective production function for capability improvement is:

$$\dot{C} \propto C^{\varphi_0/(1-\beta_{\text{auto}}\cdot\varphi_0)} \cdot Z^{(1-\beta_{\text{auto}})/(1-\beta_{\text{auto}}\cdot\varphi_0)} \tag{A.25}$$

where Z is the exogenous input (frontier model releases). The exponent on C is $\varphi_0/(1-\beta_{\text{auto}}\varphi_0) = \varphi_{\text{eff}}$.

The threshold condition $\varphi_{\text{eff}} = 1$ requires $\beta_{\text{auto}} = (1-\varphi_0)/\varphi_0$. For example, with $\varphi_0 = 0.5$, the knife-edge requires $\beta_{\text{auto}} = 1.0$—full automation, which contradicts training persistence. With $\varphi_0 = 0.8$, the threshold is $\beta_{\text{auto}} = 0.25$. □

Remark A.10.2 (The Automation Ladder). The quantity β_{auto} is not fixed; it evolves endogenously as the mesh's autocatalytic core matures. Initially $\beta_{\text{auto}} \approx 0$: the mesh cannot automate any part of its own training improvement. As training agents emerge and improve (the Jain-Krishna process of Section A.9.5), β_{auto} rises. The growth dynamics are therefore non-stationary: $\varphi_{\text{eff}}(t) = \varphi_0/(1-\beta_{\text{auto}}(t)\cdot\varphi_0)$ increases over time, potentially transitioning the system from convergence to exponential growth. The transition is not guaranteed; it depends on whether β_{auto} can reach the threshold before the Baumol bottleneck binds (Section A.13).

A.10.4 Training Saturation

Individual training interactions exhibit diminishing returns. Following the Lotka-Volterra mutualistic framework [5], the saturating interaction modifies the growth equation:

$$\dot{C}_j = \frac{\sum_k a_{jk}\cdot C_k}{1 + h\sum_k a_{jk}\cdot C_k} - \delta C_j \tag{A.26}$$

where a_{jk} is the training benefit from type-k agents to type-j agents, and $h > 0$ is the saturation parameter.

Lemma A.10.3 (Saturation Ceiling)**.** *For fixed J and $h > 0$, the system (A.26) has a unique globally stable equilibrium with $C_j^* < 1/(h\cdot\delta)$ for all j. The aggregate ceiling is:*

$$C_{max} = \left(\sum_{j=1}^{J}\left(C_j^*\right)^{\rho}\right)^{1/\rho} \leq J^{1/\rho}\cdot\frac{1}{h\delta} \tag{A.27}$$

Proof. At steady state, $\dot{C}_j = 0$ implies $\sum_k a_{jk}C_k/(1+h\sum_k a_{jk}C_k) = \delta C_j$. The left side is bounded above by $1/h$ for any $C_k \geq 0$, so $C_j^* \leq 1/(h\delta)$. The CES bound follows from substitution and the power mean inequality. For global stability, the system is cooperative (all off-diagonal Jacobian elements are non-negative) and bounded, so by the Hirsch [21] monotone dynamical systems result, the system has a unique globally attracting equilibrium. □

A.10.5 Variety Expansion as Saturation Escape

Romer's [37] key insight is that new product varieties can sustain growth even when returns to individual products diminish. The mesh analog: when fine-tuning existing specialists hits saturation ($h > 0$), the mesh can create *new* specialization types.

Let $J(t)$ evolve according to:

$$\dot{J} = \eta_J \cdot f_J \cdot C_{\text{eff}}^{\varphi_J} \cdot (J_{\max} - J) \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] \tag{A.28}$$

where $\eta_J > 0$ is the innovation rate, f_J is the fraction of training capability devoted to creating new specialization types, φ_J is the capability elasticity of variety creation, and $J_{\max}$ is the maximum number of viable specialization types.

Proposition A.10.4 (Saturation Escape via Variety)**.** *Even with training saturation $h > 0$, the growth rate of C_{eff} can remain positive if $J(t)$ is growing. Specifically, for constant per-type capability C_j^* at the saturation ceiling:*

$$\frac{dC_{\text{eff}}}{dt} = \frac{1-\rho}{\rho} \cdot \frac{C_{\text{eff}}}{J} \cdot \dot{J} \cdot \left(\frac{(C_{J+1}^*)^\rho}{J^{-1} \sum_j (C_j^*)^\rho} \right) \tag{A.29}$$

which is positive whenever a new type with $C_{J+1}^* > 0$ is created. The growth rate from variety expansion does not depend on h.

Proof. Differentiating $C_{\text{eff}} = (\sum_j C_j^\rho)^{1/\rho}$ with respect to J, treating J as continuous:

$$\frac{\partial C_{\text{eff}}}{\partial J} = \frac{1}{\rho} \left(\sum_j C_j^\rho \right)^{1/\rho - 1} \cdot (C_{J+1})^\rho \tag{A.30}$$

Since the saturation ceiling applies per type but not to the aggregate, variety expansion circumvents saturation. The aggregate grows as $J^{(1-\rho)/\rho}$ even when each individual C_j is bounded. □

A.11 The Central Theorem: Growth Regimes

We now unify the three layers into a complete characterization of the mesh's growth dynamics.

Definition A.11.1 (Growth Regimes)**.** *Define $\Phi \equiv \lambda + \varphi - 1$ as the composite capability elasticity. The growth dynamics (A.22) exhibit three regimes:*

(a) ***Convergence*** *($\Phi < 1$, i.e. $\lambda + \varphi < 2$): The growth rate g_C is decreasing in C. The system converges to a steady state C^* where $g_C = 0$. This includes the Jones [26] semi-endogenous case where $\varphi < 1$.*

(b) ***Exponential growth*** *($\Phi = 1$, i.e. $\lambda + \varphi = 2$): The growth rate g_C is independent of C. This is the Romer [37] knife-edge.*

(c) ***Superexponential growth*** *($\Phi > 1$, i.e. $\lambda + \varphi > 2$): The growth rate g_C is increasing in C. The system exhibits a finite-time singularity: $C(t) \to \infty$ as $t \to T_s < \infty$.*

Theorem A.11.2 (Growth Regime Classification)**.** *Consider the mesh growth system defined by equations (A.22), (A.24), (A.26), (A.28), and (A.35), with the autocatalytic core existing for $N > N_{auto}$ (Proposition A.9.4). The long-run behavior of $C_{eff}(t)$ falls into one of three regimes:*

Regime (a): Convergence to a ceiling. *If $\varphi_{eff} < 1$ (equivalently, $\beta_{\text{auto}} < (1 - \varphi_0)/\varphi_0$), training saturation $h > 0$, and variety is bounded ($J \leq J_{\max} < \infty$), then:*

$$C_{eff}(t) \to C_{max} \equiv J_{\max}^{(1-\rho)/\rho} \cdot \frac{1}{h\delta} \quad \text{as } t \to \infty \tag{A.31}$$

The convergence rate is governed by $1-\varphi_{eff}$. The ceiling C_{max} is increasing in $J_{\max}$, decreasing in h, and decreasing in δ. In this regime, the mesh's long-run growth rate equals the growth rate of the exogenous food set (frontier model releases). This is the Baumol bottleneck.

Regime (b): Balanced exponential growth. *If $\varphi_{eff} = 1$ and $J(t)$ grows endogenously at rate $g_J > 0$, then:*

$$C_{eff}(t) \sim C_{eff}(0) \cdot \exp\left[\left(\delta_g f^{\lambda} J_0^{\gamma_J} e^{\gamma_J g_J t} - \delta\right) t\right] \tag{A.32}$$

Even in the exponential regime, the long-run growth rate is bounded by $g_Z/(1 - \beta_{\text{auto}})$ where g_Z is the growth rate of frontier model capability.

Regime (c): Finite-time singularity. *If $\varphi_{eff} > 1$, $h = 0$ (no training saturation), and $\alpha_{eff} > \alpha_{crit}$ (model collapse avoided), then:*

$$C_{eff}(t) = \left(C_0^{1-\Phi} - (1 - \Phi) \cdot \delta_g f^{\lambda} J^{\gamma_J} \cdot t\right)^{1/(1-\Phi)} \tag{A.33}$$

where $\Phi = \lambda + \varphi_{eff} - 1 > 1$. This diverges at finite time $T_s = C_0^{1-\Phi}/[(\Phi - 1) \cdot \delta_g f^{\lambda} J^{\gamma_J}]$. The singularity requires three conditions simultaneously—each individually demanding and collectively unlikely.

Proof. Regime (a): With $\Phi < 1$, the growth rate $g_C = \delta_g f^\lambda C^{\Phi-1} J^{\gamma_J} - \delta$ is decreasing in C. The unique steady state C^* satisfies $\delta_g f^\lambda (C^*)^{\Phi-1} J^{\gamma_J} = \delta$. With training saturation $h > 0$, each $C_j \leq 1/(h\delta)$ (Lemma A.10.3), and with bounded $J \leq J_{\max}$, the ceiling $C_{\max}$ is finite. Global stability follows from the monotone dynamical systems theorem [21].

Regime (b): With $\Phi = 1$, the growth equation becomes $\dot{C} = \delta_g f^\lambda J(t)^{\gamma_J} - \delta C$, a linear ODE with time-varying coefficients. The solution is exponential with growth rate modulated by $J(t)$. The Baumol constraint (Section A.13) eventually binds: the non-automated fraction of training requires exogenous input Z growing at rate g_Z, bounding long-run growth.

Regime (c): With $\Phi > 1$ and $h = 0$, the ODE $\dot{C} = \delta_g f^\lambda C^\Phi J^{\gamma_J}$ is a Bernoulli equation. The substitution $v = C^{1-\Phi}$ yields $\dot{v} = (1-\Phi)\delta_g f^\lambda J^{\gamma_J}$, integrating to $v(t) = v(0) + (1-\Phi)\delta_g f^\lambda J^{\gamma_J} t$. Since $1 - \Phi < 0$, v decreases linearly to zero at T_s. □

Table A.3: Growth regime classification.

Regime	φ_{eff}	h	J	Long-run $C_{\text{eff}}(t)$
(a) Convergence	< 1	> 0	bounded	$\to C_{\max}$ (ceiling)
(b) Exponential	$= 1$	≥ 0	growing	$\sim e^{rt}$
(c) Singularity	> 1	$= 0$	any	$\to \infty$ at $T_s < \infty$
Additional condition for all regimes: $\alpha_{\text{eff}} > \alpha_{\text{crit}}$ (collapse avoided)				

Remark A.11.3 (Which Regime Is Likely?). The empirical evidence strongly favors regime (a) in the near term. Bloom et al. [9] estimate $\varphi \approx 0.5$–0.7 for research productivity, implying $\varphi_0 < 1$ by a substantial margin. For φ_{eff} to reach unity with $\varphi_0 = 0.6$, the autocatalytic fraction must reach $\beta_{\text{auto}} = 0.67$—the mesh must automate two-thirds of its own training improvement. Individual training interactions exhibit clear saturation ($h > 0$). And while variety expansion can escape per-type saturation, the total task space $J_{\max}$ is finite.

The most probable trajectory is: regime (a) with an increasing ceiling. As the mesh matures, β_{auto} rises (pushing φ_{eff} toward 1), J expands (raising $C_{\max}$), and the ceiling lifts—but never vanishes entirely, because the Baumol bottleneck anchors the long-run growth rate to exogenous frontier model improvement.

A.12 Diversity as Collapse Protection

The mesh's self-referential learning—agents training on data generated by other agents—risks model collapse. This section proves that the mesh's CES heterogeneity maintains training signal quality, connecting the production-theoretic parameter ρ to information-theoretic robustness. The underlying mechanism is the correlation robustness established in Chapter 2

(Theorem 5.1): the CES nonlinearity extracts idiosyncratic variation that linear aggregates miss.

A.12.1 The Model Collapse Framework

Following Shumailov et al. [39], consider a generative model Q_t trained on a mixture of authentic data from the true distribution P (fraction α) and synthetic data from the model's own previous generation Q_{t-1} (fraction $1-\alpha$). The KL divergence from the true distribution evolves as:

$$\mathrm{KL}(Q_{t+1}\|P) \geq \mathrm{KL}(Q_t\|P) \quad \text{when } \alpha < \alpha_{\text{crit}} \tag{A.34}$$

For a single model training on its own outputs ($\alpha = 0$), collapse is inevitable.

A.12.2 Effective Data Diversity in the Mesh

The mesh is not a single model. It is a collection of heterogeneous specialists whose outputs are drawn from different distributions. From the perspective of agent $k \neq i$, training data from agent i is not "self-generated"—it comes from a different distribution.

Definition A.12.1 (Effective External Data Fraction)**.** *For agent i in a mesh with J specialization types and CES parameter ρ, the effective external data fraction is:*

$$\alpha_{\mathit{eff}}(\rho, J) = \alpha_{\mathit{ext}} + (1 - \alpha_{\mathit{ext}}) \cdot D(\rho, J) \tag{A.35}$$

where α_{ext} is the fraction of training data from sources outside the mesh and $D(\rho, J)$ is the diversity correction measuring the informational diversity of mesh-internal training data.

A.12.3 The Diversity Correction

Lemma A.12.2 (Specialization Implies Distributional Diversity)**.** *Let agents be fully specialized: agent i produces outputs entirely from its specialization type $j^*(i)$. If the J specialization types are distributed uniformly across the mesh, the expected diversity correction is:*

$$D(\rho, J) = \frac{J-1}{J} \cdot \left(1 - \rho^{1/(1-\rho)}\right) \tag{A.36}$$

For $\rho < 1$: $D > 0$, and D is increasing in J and decreasing in ρ.

Proof. When agents are fully specialized, agent i's output distribution Q_i has support concentrated on task type $j^*(i)$. For agent k with $j^*(k) \neq j^*(i)$, the supports are disjoint,

giving maximal Jensen-Shannon divergence. The fraction of other agents with different specializations is $(J-1)/J$ under uniform distribution. The term $(1-\rho^{1/(1-\rho)})$ quantifies how the CES substitution parameter translates production complementarity into distributional diversity. □

A.12.4 CES Heterogeneity as Collapse Protection

Theorem A.12.3 (CES Heterogeneity as Collapse Protection)**.** *Let $\alpha_{\text{ext}} \geq 0$ be the exogenous external data fraction and let $\rho < 1$ be the CES substitution parameter. The mesh avoids model collapse ($\alpha_{\mathit{eff}} > \alpha_{\mathit{crit}}$) whenever:*

$$\alpha_{\text{ext}} + (1-\alpha_{\text{ext}}) \cdot \frac{J-1}{J} \cdot \left(1-\rho^{1/(1-\rho)}\right) > \alpha_{\mathit{crit}} \tag{A.37}$$

This condition can be satisfied even when $\alpha_{\text{ext}} < \alpha_{\mathit{crit}}$—even when the mesh's external data supply is below the collapse threshold for any individual model—provided J is sufficiently large and ρ is sufficiently small.

Proof. Substituting equation (A.36) into equation (A.35) gives condition (A.37) directly. The condition $\alpha_{\text{eff}} > \alpha_{\text{crit}}$ holds when:

$$D(\rho, J) > \frac{\alpha_{\text{crit}} - \alpha_{\text{ext}}}{1-\alpha_{\text{ext}}} \tag{A.38}$$

The right side is positive only when $\alpha_{\text{ext}} < \alpha_{\text{crit}}$. The left side $D(\rho, J)$ increases as ρ decreases and J increases. The minimum J required for collapse avoidance is:

$$J_{\min}(\rho, \alpha_{\text{ext}}) = \left\lceil \frac{1}{1 - \frac{\alpha_{\text{crit}} - \alpha_{\text{ext}}}{(1-\alpha_{\text{ext}})(1-\rho^{1/(1-\rho)})}} \right\rceil \tag{A.39}$$

This is finite for $\rho < 1$ and decreasing in ρ. □

Remark A.12.4 (The Triple Role in Action). The CES parameter ρ now instantiates two of the three roles identified in Chapter 2 (Theorem 7.1). In Part I, $\rho < 1$ generates the *superadditivity* (diversity premium): heterogeneous specialists collectively outperform centralized provision (Lemma A.4.1). In Part II, $\rho < 1$ generates *correlation robustness* (collapse protection): heterogeneous specialists generate informationally diverse training data that prevents model collapse even when external data is scarce (Theorem A.12.3). The same curvature parameter K that quantifies the production benefit of diversity also quantifies the informational benefit. This is not a coincidence; it is the CES triple role at work.

A.13 The Baumol Bottleneck

The training persistence assumption—frontier model training remains centralized—is not imposed exogenously in Part II. It emerges from the growth dynamics as a Baumol [6] cost disease. This result instantiates the hierarchical ceiling from Chapter 3 (Proposition 8.1): Level 3 (Capability) is bounded by the rate of change at Level 1 (Hardware), mediated through Level 2 (Network).

A.13.1 The Two-Sector Structure

Sector 1: Inference and fine-tuning (progressively automated). The mesh automates an increasing fraction $\beta(t)$ of inference and fine-tuning tasks. The mesh's productivity in this sector grows at rate g_C.

Sector 2: Frontier model training (non-automatable). Training frontier models requires tightly synchronized GPU clusters at scales of 10^{25}+ FLOPs per run, with synchronization bandwidth as the binding constraint (Chapter 4). Frontier training productivity grows at exogenous rate g_Z.

A.13.2 Derivation

Following Aghion, Jones and Jones [3], model the aggregate AI capability as a Cobb-Douglas composite:

$$C_{\text{eff}} = C_1^{\beta(t)} \cdot C_2^{1-\beta(t)} \tag{A.40}$$

where C_1 is the mesh's inference/fine-tuning capability (growing at g_C) and C_2 is frontier training capability (growing at exogenous rate g_Z).

Proposition A.13.1 (Endogenous Baumol Bottleneck)**.** *As $\beta(t) \to 1$, the growth rate $g_{C_{eff}} \to g_Z$ regardless of g_C, provided $g_C > g_Z$. The non-automatable sector becomes the binding constraint even as its share of total activity shrinks.*

Proof. The growth rate of the aggregate is:

$$g_{C_{\text{eff}}} = \beta(t) \cdot g_C + (1 - \beta(t)) \cdot g_Z + \dot{\beta}(t) \cdot \ln\left(\frac{C_1}{C_2}\right) \tag{A.41}$$

When $g_C > g_Z$, the relative price of frontier training rises: its cost share increases even as its volume share $(1 - \beta)$ falls—Baumol's cost disease. In the limit $\beta \to 1$:

$$\lim_{\beta \to 1} g_{C_{\text{eff}}} = g_Z + \lim_{\beta \to 1} \dot{\beta} \cdot \ln(C_1/C_2) \tag{A.42}$$

If $\dot{\beta} \to 0$ as $\beta \to 1$ (the last tasks are hardest to automate), then $g_{C_{\text{eff}}} \to g_Z$. □

A.13.3 Closing the Circle

The Baumol bottleneck connects this chapter back to Chapter 4. The chain of determination is:

(i) *Concentrated investment* (Chapter 4): Datacenter capital investment finances the GPU clusters that train frontier models. The rate g_Z is determined by the rate of investment.

(ii) *Learning curves* (Chapter 4): The same investment finances the packaging learning curves ($\alpha = 0.23$) that enable distributed inference.

(iii) *Mesh formation* (this chapter, Part I): After crossing, the mesh self-organizes into a specialized network exceeding centralized provision at $N > N^*$.

(iv) *Endogenous growth* (this chapter, Part II): The mesh improves itself through autocatalytic training, self-referential learning, and variety expansion.

(v) *Baumol ceiling* (this chapter): The mesh's growth rate converges to g_Z—the rate determined by the concentrated investment of step (i).

The circle closes. The concentrated capital that creates the crossing also determines the ceiling. The mesh amplifies frontier model improvement but cannot exceed it indefinitely. This is the Baumol bottleneck instantiated as the hierarchical ceiling (Chapter 3, Proposition 8.1): capability is bounded by the network, which is bounded by hardware.

A.14 Transition to Settlement

The mesh's routing and compensation requirements endogenously generate the need for a programmable settlement layer. Consider agent A receiving a query for which agent B is the best specialist. Three requirements arise:

(i) *Discovery:* A must identify B as the appropriate specialist.

(ii) *Incentive compatibility:* B must be compensated for serving A's query.

(iii) *Settlement:* The compensation must be transferred peer-to-peer, programmably, and at low latency.

Proposition A.14.1 (Settlement Layer Necessity)**.** *Any mesh equilibrium with $N > N^*$ and $C_{mesh} > C_{cent}$ requires a settlement layer capable of processing $O(N \cdot \langle k \rangle)$ transactions per second at $O(1)$ ms latency between arbitrary node pairs.*

With endogenous capability growth (Part II), the settlement layer must additionally process the micro-transactions of the autocatalytic loop: training agent compensation, data marketplace transactions, and variety expansion incentives. The transaction volume from autocatalytic operations scales as $O(|\mathcal{R}| \cdot f \cdot N)$, where $|\mathcal{R}|$ is the size of the RAF set.

The price system in the mesh plays exactly the role Hayek [20] described: it aggregates dispersed information into sufficient statistics for decentralized decision-making. The bid-ask spread between agents encodes current demand for each query type, current supply of each specialization, and optimal routing. No central coordinator computes these allocations.

The detailed analysis of how existing and emerging monetary infrastructure maps to these requirements—and the monetary policy consequences of mesh-scale settlement demand—is the subject of Chapter 6. The mesh's capability growth rate may be constrained by settlement infrastructure before it is constrained by training productivity, saturation, or data quality. In this case, the binding constraint on mesh improvement is *monetary*, not technological.

A.15 Frameworks Considered and Rejected

Several candidate frameworks were evaluated for the formal model and rejected for specific technical reasons.

Mean Field Games (Lasry-Lions [31]). MFG assumes a continuum of exchangeable (identical) agents whose individual optimization depends on the population distribution. The mesh's agents are heterogeneous specialists—heterogeneity is the source of the CES diversity premium that drives Theorem A.6.1. Replacing heterogeneous agents with a continuum of identical agents eliminates the mechanism. The supermodular game framework (Topkis [42]; Milgrom-Roberts [33]) handles heterogeneity naturally.

Spin Glasses (Edwards-Anderson [15]; Sherrington-Kirkpatrick [38]). Spin glass models require frustrated interactions—a mix of positive and negative couplings. In the mesh, all interactions are positive: each agent benefits from others joining the network and from others specializing in complementary tasks. There is no frustration. The appropriate model is the random-field Ising/Potts model (positive couplings, heterogeneous external fields).

Ecological Niche Models (Tilman [41]; Loreau-Hector [32]). The conceptual analogy is precise: diverse specialist communities outperform monocultures. The formal eco-

logical models, however, are calibrated to plant biomass dynamics with resource-competition mechanics that do not transfer. The CES aggregation function captures the identical qualitative result while being native to the economics literature.

Immune System / Clonal Selection (Burnet [12]). The adaptive immune system provides a vivid metaphor: a diverse repertoire of specialized agents that collectively cover a vast space through local adaptation. However, the formal ODE models are calibrated to lymphocyte population dynamics with no meaningful economic analog.

Eigen's Hypercycle (Eigen & Schuster [16]). The hypercycle imposes a conservation law: $\sum_i x_i = \text{const}$ (total concentration is fixed). This forces zero-sum dynamics. The mesh is an open system where total capability can grow. The RAF framework (Hordijk & Steel [23]) provides autocatalytic structure without the conservation constraint.

Chemical Reaction Network Theory (Feinberg [18]). CRNT assumes closed systems with stoichiometric conservation laws. The mesh is open. Moreover, CRNT characterizes equilibrium existence, not growth trajectories—the central question of Part II.

NK Fitness Landscapes (Kauffman [30]). The NK model lacks analytical results on convergence or divergence. The co-evolutionary extension (NKC model) is an open problem. Useful as motivation for variety expansion, but not as formal machinery.

A.16 Falsifiable Predictions

The model generates twelve predictions with timing and failure conditions. Predictions 1–6 concern network formation (Part I); Predictions 7–12 concern capability growth (Part II).

Part I: Network Formation

Prediction 1: First-Order Crystallization, Not Gradual Adoption. Mesh formation exhibits a discontinuous jump in adoption metrics rather than smooth logistic growth. The transition from <5% to >25% distributed inference share occurs within 18 months. *Timing:* 2030–2033. *Evidence against:* distributed inference share growing smoothly at <5 percentage points per year through 2035.

Prediction 2: Specialization Precedes Generalization. Early mesh agents are narrow specialists (fine-tuned for specific domains). General-purpose mesh capability emerges only after the specialization structure crystallizes. *Evidence against:* early mesh participants predominantly running general-purpose base models without fine-tuning.

Prediction 3: Long-Tail Niche Dominance First. The mesh achieves capability dominance first on long-tail queries before competing on mainstream tasks. Distributed inference share should exceed 50% for long-tail categories while remaining below 20% for

high-volume mainstream categories. *Timing:* within 2 years of $R_0 > 1$ crossing. *Evidence against:* mesh competing first on mainstream query types.

Prediction 4: Endogenous Hub Emergence. The mesh's degree distribution becomes fat-tailed ($\gamma \leq 3$) within 3 years of crystallization, with <1% of nodes handling >30% of routing traffic. *Evidence against:* degree distribution remaining thin-tailed through 2036.

Prediction 5: Nonlinear Knowledge Acceleration. The rate of capability improvement across the mesh accelerates nonlinearly once the degree distribution becomes fat-tailed, consistent with the vanishing epidemic threshold (Proposition A.5.2). *Evidence against:* capability improvement following a constant exponential rate through 2036.

Prediction 6: Settlement Layer as Binding Constraint. The settlement layer becomes the binding constraint on mesh growth before device capability, network connectivity, or model quality bind. *Timing:* 2031–2034. *Evidence against:* mesh growth constrained by device capability or bandwidth through 2035.

Part II: Capability Growth

Prediction 7: Autocatalytic Threshold Timing. The mesh achieves self-sustaining capability improvement within 3 years of crystallization. Observable as: mesh capability improving on standard benchmarks without new base model releases. *Timing:* 2033–2036. *Evidence against:* mesh capability plateauing during any 12-month period without new base model releases through 2038.

Prediction 8: Training Agent Emergence. Specialized training agents—agents whose primary function is improving other agents—capture >10% of internal mesh transactions within 5 years of crystallization. *Timing:* 2035–2038. *Evidence against:* mesh transactions remaining >95% pure inference through 2040.

Prediction 9: Diversity-Collapse Protection. Heterogeneous meshes ($J \geq 10$, $\rho \leq 0.5$) maintain capability when training on >50% internally generated data, while homogeneous networks ($J \leq 3$) exhibit model collapse under the same conditions. *Evidence against:* homogeneous networks showing no degradation from synthetic data training.

Prediction 10: Baumol Bottleneck Binding. The mesh's capability growth rate correlates with, and is bounded by, the rate of frontier model releases. The mesh's annualized capability growth rate should track within 1.5× of the frontier model improvement rate. *Timing:* 2034–2040. *Evidence against:* mesh capability growth exceeding 3× the frontier release rate sustained over >2 years.

Prediction 11: φ_{eff} Estimate. The mesh's effective training productivity elasticity is initially 0.6–0.8 (regime (a), converging), potentially rising toward 0.9–1.0. *Evidence against:* $\varphi_{\text{eff}} > 1.0$ sustained over >1 year.

Prediction 12: Variety Expansion Rate. The number of effective specialization types J grows at 15–30% annually during the rapid growth phase, decelerating as J approaches $J_{\max}$. *Evidence against:* J remaining constant or declining after crystallization.

A.17 Conclusion

This chapter has answered two questions about the distributed AI ecosystem after the crossing point identified in Chapter 4.

The first question—what organizational form emerges?—is answered in Part I. The answer is not isolated devices running local inference. It is a self-organizing mesh of heterogeneous specialized agents whose collective capability exceeds centralized provision once the mesh reaches critical mass. The mesh equilibrium emerges from three composable mechanisms: percolation-based connectivity ($R_0^{\text{mesh}} > 1$), CES-aggregated heterogeneous specialization ($\rho < 1$, with the diversity premium quantified by the curvature parameter K from Chapter 2), and Laplacian knowledge diffusion with a vanishing epidemic threshold on scale-free topologies. The Fortuin-Kasteleyn unification reveals that connectivity and specialization are the same mathematical object at different Potts parameter values, predicting first-order crystallization for $q > 2$ types. Inverse Bose-Einstein condensation connects the crossing to the dissolution of the centralized traffic condensate.

The second question—can the mesh improve itself?—is answered in Part II. Above the autocatalytic threshold N_{auto}, the mesh contains a self-sustaining RAF set. Three parameters govern the growth regime: training productivity elasticity φ, saturation h, and external data fraction α. The CES parameter ρ does double duty: it generates both the diversity premium (Part I, via Chapter 2's superadditivity theorem) and the collapse protection (Part II, via Chapter 2's correlation robustness theorem). The most likely near-term regime is convergence to a ceiling that rises as the autocatalytic core matures and variety expands—but the ceiling never vanishes, because the Baumol bottleneck anchors growth to the exogenous frontier training rate.

The Baumol bottleneck closes the circle. The concentrated capital investment modeled in Chapter 4 determines both when the crossing occurs and the ceiling the mesh approaches. The mesh is a multiplier, not a generator. This is the hierarchical ceiling from Chapter 3 (Proposition 8.1) in concrete form: capability is bounded by the network, which is bounded by hardware.

The organizational form that emerges is not merely a static division of labor but a dynamical system capable of self-improvement. The mesh does not just distribute inference—it creates a substrate for the endogenous growth of intelligence. The rate of that growth is

the empirical question this chapter has formalized, and the twelve predictions of Section A.16 make it testable.

.1 Renormalization Group Universality

Remark .1.1 (Mean-Field Exactness Above the Upper Critical Dimension). For systems on networks with spectral dimension $d_s > 4$, mean-field theory is exact—not an approximation, but a rigorous result [14]. Real-world networks, including the internet, social networks, and infrastructure graphs, generically have $d_s > 4$ due to the small-world property.

This means that the mean-field percolation result (Proposition A.3.1), the mean-field Potts transition (Proposition A.3.2), and the Katz-Shapiro network goods model that underlies the supermodular game are not approximations for the mesh. They are exact characterizations of the phase structure. Corrections to mean-field scaling are suppressed by powers of $1/(d_s - 4)$.

.2 RAF Theory Background

The Reflexively Autocatalytic and Food-generated (RAF) framework was introduced by Hordijk and Steel [23] to formalize Kauffman's [29, 30] theory of autocatalytic sets in the context of the origin of life. The original question was: in a random soup of molecular species, how large must the system be for a self-sustaining network of catalyzed reactions to emerge? The answer—that the threshold scales logarithmically with system complexity—is foundational.

The translation to the mesh is direct. "Molecule types" are capability types (J). "Reactions" are training operations. "Catalysts" are training agents. "Food set" is base model capabilities from centralized training. The Hordijk-Steel result translates to Proposition A.9.4.

.3 Proof Details for Effective Training Productivity

Consider the training improvement process as consisting of a unit continuum of subtasks $s \in [0, 1]$. For automated subtasks ($s \in [0, \beta_{\text{auto}}]$), the input is supplied by mesh agents: $x(s) = A_s \cdot C^{\varphi_0}$. For non-automated subtasks ($s \in (\beta_{\text{auto}}, 1]$), the input is exogenous: $x(s) = Z$.

The aggregate improvement is Cobb-Douglas across subtasks:

$$\dot{C} \propto \exp\left(\int_0^1 \ln x(s)\, ds\right) = \left(\prod_{s \le \beta_{\text{auto}}} A_s\right) \cdot C^{\beta_{\text{auto}}\varphi_0} \cdot Z^{1-\beta_{\text{auto}}} \tag{43}$$

The growth rate of C is:

$$g_C = \beta_{\text{auto}}\varphi_0 \cdot g_C + (1 - \beta_{\text{auto}})g_Z + \text{const} \tag{44}$$

Solving: $g_C(1 - \beta_{\text{auto}}\varphi_0) = (1 - \beta_{\text{auto}})g_Z + \text{const}$, confirming $\varphi_{\text{eff}} = \varphi_0/(1 - \beta_{\text{auto}}\varphi_0)$.

.4 Weitzman Recombinant Growth Connection

Weitzman [43] models the growth of ideas as a combinatorial process: new ideas are produced by recombining existing ideas. The mesh's variety expansion mechanism (Section A.10.5) has a Weitzman interpretation. New specialization types are produced by combining existing specializations: a medical-legal specialist combines medical reasoning and legal analysis. The number of potential combinations grows as $\binom{J}{2} \sim J^2/2$, providing a microfoundation for why $\dot{J}$ can be superlinear in J over portions of the trajectory.

.5 Nordhaus Singularity Analysis

Nordhaus [34] asks whether we are approaching an economic singularity. This chapter's regime (c) is precisely Nordhaus's singularity condition applied to the mesh. The contribution is identifying the three specific parameters (φ_{eff}, h, α) whose conjunction determines the singularity. The conclusion aligns with Nordhaus: the conditions are restrictive and unlikely to hold simultaneously.

Bibliography

[1] Smirl, J. (2026a). The CES triple role: Superadditivity, correlation robustness, and strategic independence as three views of isoquant curvature. Thesis Chapter 2.

[2] Smirl, J. (2026b). Complementary heterogeneity: A port-Hamiltonian framework for the AI-crypto transition. Thesis Chapter 3.

[3] Aghion, P., Jones, B. F., & Jones, C. I. (2018). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence* (pp. 237–282). University of Chicago Press.

[4] Bianconi, G., & Barabási, A.-L. (2001). Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86(24), 5632–5635.

[5] Bastolla, U., Lässig, M., Manrubia, S. C., & Valleriani, A. (2005). Biodiversity in model ecosystems, I: Coexistence conditions for competing species. *Journal of Theoretical Biology*, 235(4), 521–530.

[6] Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review*, 57(3), 415–426.

[7] Baxter, R. J. (1982). *Exactly Solved Models in Statistical Mechanics*. Academic Press.

[8] Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *Quarterly Journal of Economics*, 107(4), 1137–1160.

[9] Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4), 1104–1144.

[10] Bonabeau, E., Theraulaz, G., & Deneubourg, J.-L. (1998). Fixed response thresholds and the regulation of division of labor in insect societies. *Bulletin of Mathematical Biology*, 60(4), 753–807.

[11] Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: Engines of growth? *Journal of Econometrics*, 65(1), 83–108.

[12] Burnet, F. M. (1959). *The Clonal Selection Theory of Acquired Immunity*. Cambridge University Press.

[13] Christensen, C. M. (1997). *The Innovator's Dilemma*. Harvard Business School Press.

[14] Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4), 1275–1335.

[15] Edwards, S. F., & Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5), 965–974.

[16] Eigen, M., & Schuster, P. (1977). The hypercycle: A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64(11), 541–565.

[17] Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.

[18] Feinberg, M. (2019). *Foundations of Chemical Reaction Network Theory*. Springer.

[19] Fortuin, C. M., & Kasteleyn, P. W. (1972). On the random-cluster model: I. Introduction and relation to other models. *Physica*, 57(4), 536–564.

[20] Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, 35(4), 519–530.

[21] Hirsch, M. W. (1985). Systems of differential equations that are competitive or cooperative. II: Convergence almost everywhere. *SIAM Journal on Mathematical Analysis*, 16(3), 423–439.

[22] Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[23] Hordijk, W., & Steel, M. (2004). Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4), 451–461.

[24] Jain, S., & Krishna, S. (1998). Autocatalytic sets and the growth of complexity in an evolutionary model. *Physical Review Letters*, 81(25), 5684–5687.

[25] Jain, S., & Krishna, S. (2001). A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proceedings of the National Academy of Sciences*, 98(2), 543–547.

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

[27] Jones, C. I. (2005). Growth and ideas. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of Economic Growth* (Vol. 1B, pp. 1063–1111). Elsevier.

[28] Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review*, 75(3), 424–440.

[29] Kauffman, S. A. (1971). Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *Journal of Cybernetics*, 1(1), 71–96.

[30] Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

[31] Lasry, J.-M., & Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1), 229–260.

[32] Loreau, M., & Hector, A. (2001). Partitioning selection and complementarity in biodiversity experiments. *Nature*, 412, 72–76.

[33] Milgrom, P., & Roberts, J. (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58(6), 1255–1277.

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

[35] Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14), 3200–3203.

[36] Potts, R. B. (1952). Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1), 106–109.

[37] Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71–S102.

[38] Sherrington, D., & Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physical Review Letters*, 35(26), 1792–1796.

[39] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

[40] Tarski, A. (1955). A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5(2), 285–309.

[41] Tilman, D. (1982). *Resource Competition and Community Structure*. Princeton University Press.

[42] Topkis, D. M. (1998). *Supermodularity and Complementarity*. Princeton University Press.

[43] Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, 113(2), 331–360.

[44] Wu, F. Y. (1982). The Potts model. *Reviews of Modern Physics*, 54(1), 235–268.

[45] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115(772), 700–721.

[46] Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1), 253–258.

Chapter A

The Settlement Feedback

A.1 Introduction

The AI mesh requires settlement infrastructure. Chapter 5 (The Mesh Economy, Section 8) derives that routing compensation among heterogeneous specialized agents generates a demand for programmable money: micro-transactions between inference providers, training agents, and end users must be settled in real time at scale. Chapter 5, Part II (The Autocatalytic Mesh, Section 8) shows that endogenous capability growth intensifies this demand, adding training agent compensation, data marketplace transactions, and variety expansion incentives to the settlement volume. Chapter 7 (The Monetary Productivity Gap) identifies the 6.4 percentage point cost gap between fiat payment rails and stablecoin settlement, establishing dollar stablecoins as the efficient settlement medium. The settlement layer is stablecoins. Stablecoins are backed by US Treasuries.

This creates a feedback loop. Mesh growth increases stablecoin demand. Stablecoin reserves are predominantly short-duration US Treasury instruments. Therefore mesh growth increases Treasury demand. But the mesh does not merely consume settlement services—it transforms the financial system that provides them. As autonomous agents enter capital markets, they process information at machine speed, optimize portfolios continuously, and arbitrage mispricings in milliseconds. The fraction ϕ of capital managed by autonomous agents is growing. As ϕ increases, market efficiency changes, monetary policy tools degrade, and the conditions for currency stability in weak-fiat countries shift. Each of these transformations feeds back into the settlement layer that the mesh depends on.

The contribution of this chapter is the formal demonstration that the mesh–financial system coupling produces a self-reinforcing dynamical system, and the characterization of its equilibrium structure. The feedback loop has five links:

(i) Mesh growth generates stablecoin demand (Chapter 5, Section 8; Chapter 5, Part II, Section 8).

(ii) Mesh agents entering capital markets increase market efficiency toward the Grossman-Stiglitz [7] limit.

(iii) Increased market efficiency degrades monetary policy tools that depend on market frictions.

(iv) Monetary policy degradation, combined with stablecoin access, triggers dollarization spirals in weak-currency countries through the Uribe [16] hysteresis mechanism.

(v) Dollarization expands the stablecoin ecosystem, which is the mesh's settlement layer. Settlement infrastructure improves. Mesh growth accelerates. Return to step (i).

The loop has R_0 structure. Define R_0^{settle} as the number of units of subsequent mesh growth produced by each unit of current mesh growth through the financial system channel. If $R_0^{\text{settle}} > 1$, the loop is self-reinforcing: each cycle amplifies the next. This chapter derives the conditions under which $R_0^{\text{settle}} > 1$ and characterizes the dynamics that follow.

Three central questions organize the analysis. First, does the coupled system admit a stable equilibrium? The Farhi-Maggiori [19] framework identifies three zones for the reserve currency issuer—safety, instability, and collapse—but the zone boundaries become endogenous to mesh participation. Second, how does market efficiency change as the marginal capital market participant shifts from human to mesh agent? The Grossman-Stiglitz [7] paradox structures the answer, but the transition may exhibit non-smooth features. Third, under what conditions does the dollarization spiral become self-reinforcing, and which countries are vulnerable at current stablecoin ecosystem size?

The chapter is not speculative. Every mechanism draws from published, peer-reviewed economic theory applied to the specific characteristics of mesh agents: machine-speed information processing, continuous portfolio optimization, zero intermediation friction, and permissionless cross-border capital mobility. The contribution is the coupling—showing that market efficiency, monetary policy effectiveness, dollarization dynamics, and safe-asset supply are connected through the settlement layer—and the characterization of the resulting equilibrium.

The chapter proceeds as follows. Section 1.1 situates this chapter within the thesis framework. Section 2 reviews terminal conditions from the prior chapters. Section 3 develops the market microstructure transition. Section 4 models monetary policy effectiveness as a function of mesh participation. Section 5 formalizes the dollarization spiral. Section 6 analyzes the Triffin contradiction with endogenous instability boundaries. Section 7 unifies the preceding sections into a coupled dynamical system and characterizes its equilibria. Section 8 derives implications for sovereign fiscal policy. Section 9 discusses frameworks considered and rejected. Section 10 presents falsifiable predictions. Section 11 concludes.

A.1.1 Relation to the Thesis Framework

This chapter instantiates Level 4 (Settlement, fastest timescale) of the four-level hierarchy developed in Chapter 3 [2]. In the notation of that chapter's Section 5.4, the state variable is stablecoin infrastructure size S, the gain function is settlement demand generated by mesh transactions, and the characteristic timescale is days to weeks—the fastest layer in the hierarchy.

The Triffin contradiction identified in this chapter (Section 6) is mathematically the

same object as the Baumol bottleneck in Chapter 5. Both are slow-manifold constraints at adjacent layers of the hierarchy (Chapter 3, Proposition 8.1). The Baumol bottleneck constrains capability growth (Level 3) to the pace of network formation (Level 2); the Triffin contradiction constrains financial settlement (Level 4) to the pace of sovereign debt dynamics that the settlement layer itself transforms. In each case, a faster sector is bounded by its slower parent: $F_4 \leq \bar{S}(F_3)$, just as $F_3 \leq (\varphi_{\text{eff}}/\delta_C) \cdot F_{\text{CES}}(N^*(F_1))$.

The damping cancellation theorem (Chapter 3, Proposition 7.1) has a direct policy implication here: tightening stablecoin regulation at Level 4 has zero net welfare effect because faster convergence exactly offsets lower equilibrium output. The upstream reform principle (Chapter 3, Theorem 7.2) implies that improving settlement-layer outcomes requires reforming capability-layer institutions (Level 3) or network-layer access (Level 2). Regulating stablecoins directly—imposing capital requirements, mandating audits, restricting issuance—may speed convergence to equilibrium but cannot raise the equilibrium itself. The binding constraint on settlement-layer performance is upstream.

The activation threshold for this level follows from the general condition (Chapter 3, Theorem 4.3): the settlement feedback loop activates when the spectral radius of the next-generation matrix at Level 4 exceeds unity, $R_0^{\text{settle}} > 1$. The port topology theorem (Chapter 3, Theorem 3.1) constrains the coupling structure: this chapter's four-ODE system inherits the directed feed-forward topology forced by timescale separation, with aggregate coupling to the slower levels and nearest-neighbor coupling to Level 3 (capability).

By the CES Triple Role ([1], Theorem 7.1), the curvature parameter $K = (1-\rho)(J-1)/J$ ([1], Definition 3.1) simultaneously governs the superadditivity of mesh output, the correlation robustness that prevents model collapse in self-referential training, and the strategic independence that makes manipulation of settlement pricing unprofitable. These three properties are not separate assumptions invoked at different points in the analysis—they are the same geometric fact deployed at the settlement level.

A.2 Terminal Conditions from the Prior Chapters

This section summarizes the results from the preceding chapters that serve as inputs to the present analysis. No new results are presented; the purpose is notational continuity and identification of the specific mechanisms that generate the coupling formalized here.

A.2.1 From Chapter 5 (The Mesh Economy)

Chapter 5 establishes that for $R_0^{\text{mesh}} > 1$ and CES substitution parameter $\rho < 1$, there exists a finite critical mass N^* such that for $N > N^*$, the mesh equilibrium exists, is unique, is locally asymptotically stable, and the mesh's collective capability exceeds centralized provision: $C_{\text{mesh}}(N) > C_{\text{cent}}$. At maturity, the mesh's settlement layer processes $O(N\langle k \rangle)$ inference transactions per second. Section 8 of Chapter 5 derives that the routing and compensation requirements endogenously generate demand for programmable money—the mechanism that initiates the feedback loop formalized here.

A.2.2 From Chapter 7 (The Monetary Productivity Gap)

Chapter 7 establishes a 6.4 percentage point cost gap between fiat payment rails and stablecoin settlement for cross-border transactions. That chapter develops a six-stage country classification (Pre-Industrial through Post-Industrial) mapping each country group's position relative to stablecoin adoption dynamics. The key result for this chapter: stablecoin adoption proceeds as a cascade through country groups ordered by fiat quality, with each group's adoption lowering the threshold for the next. The six-stage classification maps directly to the dollarization vulnerability analysis of Section 5.

A.2.3 From Chapter 5, Part II (Capability Growth)

Chapter 5, Part II (The Autocatalytic Mesh) removes the fixed-capability assumption and characterizes three growth regimes. The most likely near-term regime is convergence: $C_{\text{eff}}(t) \to C_{\max}$ where the ceiling is determined by the Baumol bottleneck—frontier training as the non-automatable sector. Section 8 of that chapter shows that the autocatalytic loop requires training agent compensation, data marketplace transactions, and variety expansion incentives, all settled through the programmable settlement layer. The transaction volume from autocatalytic operations scales as $O(|\mathcal{R}| \cdot f \cdot N)$, where $|\mathcal{R}|$ is the size of the RAF set and f is the training allocation fraction.

The key result for this chapter: as the autocatalytic core matures, settlement demand grows faster than inference demand. The binding constraint on mesh capability growth may be monetary (settlement infrastructure) rather than technological (φ_{eff}, h, α).

A.2.4 From Chapter 4 (Endogenous Decentralization)

Chapter 4 models concentrated AI infrastructure investment and its paradoxical consequence: the same investment that finances centralized training also finances the learning curves (3D

memory stacking, advanced packaging) that enable distributed inference. The crossing point $R_0 > 1$ initiates the mesh. The rate of concentrated investment determines the frontier model improvement rate g_Z, which the Baumol bottleneck (Chapter 5, Part II) identifies as the mesh's long-run growth rate limiter.

A.2.5 Key Retained Assumptions

Three assumptions carry forward from the preceding chapters:

(i) *Training persistence:* Frontier model training remains centralized. The mesh fine-tunes and adapts base models; it does not train frontier models from scratch. This is the exogenous input ("food set") whose growth rate g_Z bounds the mesh through the Baumol bottleneck.

(ii) *CES structure:* The aggregate capability function retains the CES form $C_{\text{eff}} = (\sum_j C_j^\rho)^{1/\rho}$ with $\rho < 1$, ensuring complementarity across task types and providing model collapse protection ($\alpha_{\text{eff}} > \alpha_{\text{crit}}$). By the CES Triple Role ([1], Theorem 7.1), the curvature parameter $K = (1-\rho)(J-1)/J$ simultaneously governs the superadditivity gap, the correlation robustness bonus, and the strategic manipulation penalty.

(iii) *Scale-free topology:* The mesh's degree distribution follows a power law with $\gamma \leq 3$, ensuring a vanishing epidemic threshold for information propagation and the preferential attachment dynamics that drive specialization.

A.3 Market Microstructure Transition

As the mesh matures, its autonomous agents enter capital markets—first as participants in stablecoin markets, then as optimizers of the Treasury portfolios backing those stablecoins, and eventually as general portfolio managers. This section characterizes market efficiency as a function of the fraction $\phi \in [0, 1]$ of capital managed by autonomous mesh agents.

A.3.1 The Grossman-Stiglitz Framework with Mesh Agents

Grossman and Stiglitz [7] prove that perfectly efficient markets are impossible: if prices fully reveal all information, no agent has incentive to pay for information, but then prices cannot incorporate information. The equilibrium involves partial revelation, with the degree of revelation depending on the cost of acquiring information.

In the standard GS model, a fraction μ of traders are informed (each paying cost c for a signal about the asset's true value v) and the remaining $1-\mu$ are uninformed. Market efficiency is:

$$E = \frac{\text{Var}(v) - \text{Var}(v|P)}{\text{Var}(v)} = \frac{\sigma_v^2 - \sigma_{v|P}^2}{\sigma_v^2} \tag{A.1}$$

where P is the equilibrium price and $\sigma_{v|P}^2$ is the residual uncertainty after observing P. In the GS equilibrium, $E < 1$ and the gap $1 - E$ is sustained by information cost $c > 0$.

Mesh agents modify this framework in two ways. First, they reduce information acquisition cost. For mesh agents, information processing is automated and operates at marginal cost approaching zero. Define the effective information cost as a weighted average:

$$c(\phi) = (1-\phi) \cdot c_H + \phi \cdot c_M \tag{A.2}$$

where $c_H > 0$ is the cost for human institutions and $c_M \ll c_H$ is the cost for mesh agents. As ϕ increases, $c(\phi)$ declines toward $c_M > 0$. The strictly positive c_M reflects irreducible computational cost: even mesh agents expend resources to process information, though orders of magnitude less than human institutions.

Second, mesh agents alter the composition of noise trading. In the standard Kyle [8] model, uninformed "noise traders" generate order flow $u \sim \mathcal{N}(0, \sigma_u^2)$ that provides the camouflage allowing informed traders to profit. As mesh agents replace human institutions, the pool of noise traders shrinks: $\sigma_u^2(\phi) = \sigma_{u,0}^2 \cdot (1-\phi)^{\gamma_u}$ where $\gamma_u > 0$ captures the rate of noise trading exit.

A.3.2 Market Efficiency as a Function of ϕ

Proposition A.3.1 (Efficiency Transition)**.** *Define market efficiency $E(\phi)$ as the fraction of fundamental value variance revealed by prices. Under the modified GS-Kyle framework with mesh participation ϕ:*

$$E(\phi) = 1 - \varepsilon(\phi) \quad \textit{where} \quad \varepsilon(\phi) = \frac{c(\phi) \cdot \sigma_{v|P}^2(\phi)}{\pi_I(\phi)} \tag{A.3}$$

and $\pi_I(\phi)$ is the equilibrium profit from informed trading. As $\phi \to 1$:

$$\varepsilon(\phi) \to \varepsilon_{\min} > 0 \tag{A.4}$$

The residual inefficiency $\varepsilon_{\min}$ is the equilibrium "noise" that sustains market function. It is small—determined by the mesh agents' irreducible information cost c_M—but strictly positive.

The Grossman-Stiglitz paradox is preserved: the market cannot be fully efficient because efficiency eliminates the returns that pay for information production.

Proof. In the GS equilibrium, the fraction μ^* of agents who become informed satisfies the indifference condition: the expected utility of being informed (net of cost c) equals the expected utility of being uninformed. With mesh agents, this indifference condition becomes heterogeneous: mesh agents become informed whenever $c_M < \pi_I$, while human agents become informed whenever $c_H < \pi_I$. The equilibrium profit π_I adjusts until the marginal agent (the last mesh or human agent to become informed) is indifferent.

As $\phi \to 1$, nearly all capital is managed by mesh agents with cost c_M. The GS indifference condition requires $\pi_I \geq c_M > 0$, which in turn requires $\varepsilon > 0$ (positive residual variance for informed agents to profit from). The minimum residual inefficiency satisfies:

$$\varepsilon_{\min} = \frac{c_M}{\sigma_v^2} \cdot \frac{1}{\text{Var(informed profit per unit residual)}} \tag{A.5}$$

which is positive but small, since $c_M/c_H \ll 1$ by assumption. □

Remark A.3.2 (Transition Smoothness). The efficiency transition $E(\phi)$ is smooth: there is no phase transition at a critical ϕ^*. This follows because the GS indifference condition adjusts continuously as the composition of market participants changes. However, the *rate* of efficiency improvement is non-uniform. Most of the efficiency gain occurs in the interval $\phi \in [0.1, 0.5]$, when mesh agents are actively displacing the least-efficient human institutions. Above $\phi \approx 0.5$, diminishing returns set in: the remaining human participants are the most sophisticated, and displacing them yields smaller efficiency gains per unit of ϕ.

A.3.3 Kyle's Lambda: Non-Monotonicity in ϕ

Proposition A.3.3 (Non-Monotone Price Impact)**.** *Kyle's [8] price impact parameter λ is non-monotone in ϕ. Define:*

$$\lambda(\phi) = \frac{\sigma_v(\phi)}{2\sigma_u(\phi)} \cdot \frac{1}{\sqrt{n(\phi)}} \tag{A.6}$$

where $\sigma_v(\phi)$ is the residual fundamental uncertainty, $\sigma_u(\phi)$ is noise trading volume, and $n(\phi)$ is the number of independent informed trading strategies. Then:

$$\frac{d\lambda}{d\phi} \lessgtr 0 \quad \textit{as} \quad \phi \lessgtr \phi_\lambda \tag{A.7}$$

for some critical $\phi_\lambda \in (0, 1)$. Market depth improves for $\phi < \phi_\lambda$ and deteriorates for $\phi > \phi_\lambda$.

Proof. Two opposing forces act on λ. First, as ϕ increases, the number of independently informed trading strategies $n(\phi)$ rises (more mesh agents processing information independently). By the Holden-Subrahmanyam [9] result, competition among informed traders reduces λ: $\partial\lambda/\partial n < 0$. Second, as ϕ increases, noise trading volume $\sigma_u(\phi)$ decreases because human institutions—the primary source of noise trading—exit. Since $\lambda \propto 1/\sigma_u$, this increases λ: $\partial\lambda/\partial\sigma_u < 0$ and $\partial\sigma_u/\partial\phi < 0$, so $\partial\lambda/\partial\phi > 0$ through this channel.

At low ϕ, the informed-competition channel dominates: adding mesh agents increases informed volume while noise trading volume remains high. At high ϕ, the noise-trading-exit channel dominates: the remaining market has many informed agents but little camouflage. The critical ϕ_λ occurs where these forces balance:

$$\left|\frac{\partial\lambda}{\partial n}\cdot\frac{dn}{d\phi}\right| = \left|\frac{\partial\lambda}{\partial\sigma_u}\cdot\frac{d\sigma_u}{d\phi}\right| \tag{A.8}$$

The existence of an interior solution follows from the continuity of both channels and the boundary conditions: at $\phi = 0$, λ is at its initial level; as $\phi \to 1$, $\sigma_u \to 0$ forces $\lambda \to \infty$ unless bounded by some residual noise. In practice, complete elimination of noise is unlikely (some noise remains from stochastic liquidity needs), so λ remains finite but elevated at high ϕ. □

A.3.4 Algorithmic Collusion Risk

Dou, Goldstein and Ji [11] analyze the conditions under which algorithmic trading agents develop implicit collusion, degrading market efficiency rather than improving it. The concern is that at high ϕ, mesh agents optimizing similar objective functions with access to similar information may converge on coordinated trading strategies that reduce competition and extract rents from the remaining market participants.

The mesh's heterogeneity provides partial protection against this risk. By the CES Triple Role ([1], Theorem 7.1), the curvature parameter $K > 0$ ensures that mesh agents are differentiated: different specialists have different information sets, different optimization objectives, and different time horizons. A medical-AI specialist trading pharmaceutical companies processes different signals from a logistics-AI specialist trading shipping firms. The strategic independence property—the third role of CES curvature—implies that balanced allocation is a Nash equilibrium and coalitions cannot profitably redistribute. The same heterogeneity that prevents model collapse (Chapter 5, Part II, Theorem 3 therein) also impedes algorithmic collusion: collusion requires agents to agree on a coordinated strategy, which is harder when their information and objectives differ.

This is a conjecture, not a theorem. The CES curvature parameter K governs distributional diversity, not strategic independence in the financial-markets sense directly. Agents with different information could still coordinate if the coordination mechanism operates at a level of abstraction above the individual information sets. Whether the mesh's heterogeneity is sufficient to prevent algorithmic collusion at high ϕ is an empirical question whose answer depends on the institutional structure of mesh agents' interaction with financial markets.

A.4 Monetary Policy Effectiveness as a Function of ϕ

This section applies the Lucas [15] critique systematically to each monetary policy tool, identifying the specific friction each tool depends on and modeling that friction as a function of mesh participation ϕ and stablecoin ecosystem size S. The monetary macro structure follows Brunnermeier and Sannikov [12].

A.4.1 Forward Guidance

Forward guidance operates by shaping expectations about the future path of interest rates. Its effectiveness depends on a time delay: human institutions require days to weeks to process central bank communications, during which the anticipated policy path affects portfolio decisions and asset prices. The delay is the friction that gives forward guidance its power—if all agents processed the information instantaneously, prices would adjust immediately and the "guidance" would have no duration of effect.

Mesh agents process central bank communications in milliseconds. A text-based FOMC statement is parsed, cross-referenced with historical patterns, and incorporated into portfolio optimization before a human analyst finishes reading the first paragraph. The forward guidance friction is information processing delay, and mesh agents eliminate it.

Definition A.4.1 (Forward Guidance Effectiveness)**.** *Let τ_H denote the characteristic time for human institutions to fully process and respond to central bank forward guidance. Forward guidance effectiveness is:*

$$\mathrm{FG}(\phi) = \mathrm{FG}_0 \cdot (1 - \phi)^{\alpha_{FG}} \tag{A.9}$$

where FG_0 is the baseline effectiveness at $\phi = 0$ and $\alpha_{FG} > 0$ captures the rate at which mesh participation eliminates the processing delay. As $\phi \to 1$, $\mathrm{FG} \to 0$.

Forward guidance degrades first among monetary policy tools because it depends only on information processing speed, which is the most direct advantage mesh agents possess.

A.4.2 Quantitative Easing

Quantitative easing operates through two channels. The *portfolio balance channel* works by the central bank purchasing long-duration assets, shifting the supply-demand balance and compressing term premia. This channel requires that the buying pressure is not immediately arbitraged away—that is, it depends on slow arbitrage. The *signaling channel* communicates the central bank's commitment to low rates, which does not depend on market friction.

Mesh agents arbitrage at machine speed. A Fed purchase of 10-year Treasuries that takes human institutions weeks to fully price is arbitraged in seconds by mesh agents rebalancing across duration, credit, and currency. The portfolio balance channel is precisely the kind of friction-dependent mechanism that mesh participation eliminates. The signaling channel survives because it operates through expectations rather than through market frictions, but the signaling channel alone is substantially weaker (Woodford [14]).

Definition A.4.2 (QE Effectiveness)**.** *Quantitative easing effectiveness is:*

$$\mathrm{QE}(\phi) = w_{PB} \cdot \mathrm{QE}_0 \cdot (1-\phi)^{\alpha_{QE}} + w_{sig} \cdot \mathrm{QE}_0 \tag{A.10}$$

where $w_{PB} + w_{sig} = 1$ are the weights on the portfolio balance and signaling channels respectively, and $\alpha_{QE} > 0$ governs the degradation of the portfolio balance channel. The signaling component $w_{sig} \cdot \mathrm{QE}_0$ is invariant to ϕ because it depends on the central bank's credibility, not on market frictions.

QE degrades second—after forward guidance but before financial repression—because it depends on arbitrage speed, which mesh agents improve but which retains some friction even at high ϕ (large positions still face inventory costs and capital constraints).

A.4.3 Financial Repression

Financial repression operates by compelling domestic savers to hold government debt at below-market rates. The mechanism requires two conditions: (i) negative real returns on government bonds (the "tax"), and (ii) capital controls or institutional barriers that prevent savers from exiting to alternative stores of value (the "captivity"). The critical friction is captive savings: savers who have no practical alternative to negative-real-return domestic bonds.

Stablecoins destroy this captivity. A saver in a country with 15% inflation and capital controls can download a stablecoin wallet and hold dollar-denominated assets yielding 0% nominal but +15% real relative to the domestic currency. The exit is binary: either the saver

can access stablecoins (in which case captivity is broken) or cannot (in which case it is maintained). The relevant variable is not mesh participation ϕ directly, but stablecoin ecosystem size S—the availability of wallets, on-ramps, merchant acceptance, and user familiarity.

Definition A.4.3 (Financial Repression Effectiveness)**.** *Financial repression effectiveness is:*

$$\mathrm{FR}(\phi, S) = \mathrm{FR}_0 \cdot \left(1 - \min\left(1, \frac{S}{S_{crit}}\right)\right)^{\alpha_{FR}} \tag{A.11}$$

where S_{crit} is the critical stablecoin ecosystem size at which captive savers have a viable exit, and $\alpha_{FR} > 0$ governs the rate of collapse. The dependence on ϕ enters indirectly through the relationship $\dot{S} = f_S(\phi, S, b)$ (formalized in Section 7).

Financial repression degrades last because it depends on institutional barriers (capital controls, banking regulations) rather than information speed. But when it degrades, it degrades discontinuously. Below S_{crit}, the stablecoin ecosystem is too small to provide a viable exit—the on-ramps are unavailable, the wallets are unfamiliar, the regulatory risk is prohibitive. Above S_{crit}, the exit is available and captive savers leave en masse. The transition is not gradual; it resembles a bank run (Diamond and Dybvig [25]) in which the coordination game among captive savers has a tipping point.

A.4.4 Composite Monetary Policy Effectiveness

Proposition A.4.4 (Composite Monetary Policy Degradation)**.** *Total monetary policy effectiveness is:*

$$\mathrm{MP}(\phi, S) = w_{FG} \cdot \mathrm{FG}(\phi) + w_{QE} \cdot \mathrm{QE}(\phi) + w_{FR} \cdot \mathrm{FR}(\phi, S) \tag{A.12}$$

where $w_{FG} + w_{QE} + w_{FR} = 1$ are the weights reflecting each tool's contribution to total policy effectiveness. This is:

(i) Monotonically declining in ϕ, since each component is weakly decreasing in ϕ.

(ii) Declining in S, through the financial repression channel.

(iii) Potentially discontinuous in S at $S = S_{crit}$, where financial repression collapses.

The partial derivatives satisfy:

$$\frac{\partial \mathrm{MP}}{\partial \phi} < 0 \quad \textit{for all } \phi \in (0,1), \qquad \frac{\partial \mathrm{MP}}{\partial S} \leq 0 \quad \textit{for all } S > 0 \tag{A.13}$$

with $|\partial \mathrm{MP}/\partial S|$ largest in a neighborhood of $S = S_{crit}$.

Proof. The monotonicity in ϕ follows from equations (A.9) and (A.10): both FG and the portfolio-balance component of QE are proportional to $(1-\phi)^\alpha$ for $\alpha > 0$, hence strictly decreasing. The monotonicity in S follows from equation (A.11): FR is decreasing in S/S_{crit} for $S \leq S_{\text{crit}}$ and zero for $S \geq S_{\text{crit}}$. The potential discontinuity arises because FR transitions from $\text{FR}_0 \cdot (1 - S/S_{\text{crit}})^{\alpha_{FR}}$ to zero as S crosses S_{crit}; while technically continuous, the derivative $\partial\text{FR}/\partial S$ is largest near S_{crit}, producing a sharp decline in MP that is effectively discontinuous in practice. □

A.4.5 The Brunnermeier-Sannikov Volatility Paradox

Brunnermeier and Sannikov [12, 13] identify a "volatility paradox": periods of low exogenous volatility can breed endogenous instability as financial institutions increase leverage in calm environments. The paradox applies directly to the mesh transition.

As ϕ increases, exogenous volatility falls: mesh agents reduce noise trading, improve price discovery, and dampen uninformed price fluctuations. But the *endogenous* volatility from monetary policy ineffectiveness may increase. If a fiscal shock occurs and the central bank cannot effectively respond (because $\text{MP}(\phi, S)$ is low), the resulting adjustment must occur through market prices rather than through policy intervention. The speed of this adjustment—at machine speed, not at the pace of FOMC meetings—amplifies the magnitude of price moves.

Remark A.4.5 (Net Volatility Ambiguity). The net effect on financial stability is ambiguous and depends on parameter values. Define total volatility $\sigma^2_{\text{total}} = \sigma^2_{\text{exog}}(\phi) + \sigma^2_{\text{endog}}(\phi, S)$ where σ^2_{exog} is decreasing in ϕ (mesh agents reduce noise) and σ^2_{endog} is increasing in ϕ and S (monetary policy degradation increases crisis amplitude). There exist parameter configurations in which σ^2_{total} is non-monotone in ϕ: initially declining as noise reduction dominates, then increasing as monetary policy degradation dominates. The conditions under which each regime obtains are derived in the coupled system analysis of Section 7.

A.4.6 Surviving Monetary Policy Tools

Not all monetary policy tools degrade. Two channels survive at high ϕ:

The *interest rate channel* continues to operate because borrowing costs affect real economic activity regardless of market efficiency. When the central bank raises the policy rate, the cost of capital increases for firms and households. This mechanism does not depend on information delay, arbitrage speed, or captive savings. Mesh agents transmit rate changes to asset prices faster, but the real effects of rate changes on investment and consumption are

mediated by physical economy dynamics (construction timelines, business planning horizons) that mesh participation does not accelerate.

The *lender-of-last-resort* function survives because central banks can still create reserves. In a liquidity crisis, the ability to supply unlimited domestic-currency liquidity is independent of market efficiency. However, the LOLR function may be less effective if the crisis involves a flight from domestic currency to stablecoins—the central bank can supply domestic liquidity but cannot supply dollar-denominated stablecoin liquidity.

A.5 The Dollarization Spiral

The Uribe [16] model of currency substitution exhibits hysteresis: above a critical inflation threshold, dollarization is self-reinforcing and irreversible; below a lower threshold, de-dollarization is possible; between them, multiple equilibria exist with path dependence. This section extends the Uribe framework with stablecoin-mediated dollarization, showing that stablecoin access lowers both thresholds.

A.5.1 Dollarization Capital in the Stablecoin Era

In Uribe's original model, "dollarization capital" refers to the accumulated infrastructure—dollar-denominated bank accounts, knowledge of dollar transactions, institutional arrangements—that facilitates the use of dollars as a medium of exchange. Building dollarization capital is costly: it requires offshore bank accounts, correspondent banking relationships, and institutional knowledge that accumulates slowly.

Stablecoins reduce this cost dramatically. Dollarization capital in the stablecoin era is an app download, not an offshore bank account. Define stablecoin-era dollarization capital k as the aggregate stock of stablecoin adoption infrastructure: wallets installed, on-ramps operational, merchant acceptance established, and user familiarity accumulated.

The evolution of dollarization capital follows a modified Uribe accumulation equation:

$$\dot{k} = v(\pi, k, S) - \delta_k \cdot k \tag{A.14}$$

where v is the rate of stablecoin adoption, increasing in domestic inflation π (higher inflation increases the incentive to exit), existing dollarization capital k (network effects: existing users attract new users), and global stablecoin ecosystem size S (larger ecosystem means better infrastructure, more on-ramps, lower friction). The depreciation rate δ_k captures the erosion of dollarization infrastructure through regulatory crackdowns, app deletion, and institutional decay.

A.5.2 Modified Bifurcation Thresholds

The Uribe model has two critical inflation thresholds: $\bar{\pi}$ above which dollarization is inevitable (the unique equilibrium is fully dollarized), and $\underline{\pi}$ below which de-dollarization is possible (the unique equilibrium is domestic-currency-dominant). Between $\underline{\pi}$ and $\bar{\pi}$, multiple equilibria exist and the outcome depends on history and expectations.

Stablecoin access lowers both thresholds by reducing the cost of dollarization capital accumulation.

Proposition A.5.1 (Stablecoin-Modified Thresholds)**.** *The modified bifurcation thresholds $\bar{\pi}(S)$ and $\underline{\pi}(S)$ are decreasing functions of stablecoin ecosystem size S:*

$$\bar{\pi}(S) = \bar{\pi}_0 \cdot \left(\frac{S_0}{S_0 + S}\right)^{\beta_\pi} \quad \textit{and} \quad \underline{\pi}(S) = \underline{\pi}_0 \cdot \left(\frac{S_0}{S_0 + S}\right)^{\beta_\pi} \tag{A.15}$$

where $\bar{\pi}_0$ and $\underline{\pi}_0$ are the pre-stablecoin thresholds, S_0 is a scaling parameter, and $\beta_\pi > 0$ governs the sensitivity of thresholds to ecosystem size. Both thresholds are:

(i) Decreasing in S: larger stablecoin ecosystems make dollarization easier, so lower inflation rates trigger the spiral.

(ii) Bounded below: even as $S \to \infty$, the thresholds remain positive because some inflation is needed to motivate the switching cost.

(iii) Proportionally shifted: the ratio $\bar{\pi}(S)/\underline{\pi}(S) = \bar{\pi}_0/\underline{\pi}_0$ is invariant to S—the width of the multiple-equilibria zone (in log terms) is preserved.

Proof. In the Uribe model, the upper threshold $\bar{\pi}_0$ is determined by the condition that the flow benefit of dollarization (avoided inflation tax) exceeds the flow cost of maintaining dollarization capital (effort, foregone local-currency services) even starting from zero dollarization capital. Stablecoin access reduces the maintenance cost: a stablecoin wallet is cheaper to maintain than an offshore bank account. The cost reduction scales with ecosystem size S because a larger ecosystem provides better infrastructure (more on-ramps, lower fees, wider merchant acceptance). The functional form $(S_0/(S_0 + S))^{\beta_\pi}$ captures decreasing marginal returns to ecosystem size with S_0 as the half-life parameter: the threshold halves when $S = S_0 \cdot (2^{1/\beta_\pi} - 1)$.

The lower threshold $\underline{\pi}_0$ is determined by the condition that the stock of existing dollarization capital depreciates faster than it is maintained by the (now low) inflation incentive. Stablecoin infrastructure depreciates slower than traditional dollarization capital (an app persists on a phone longer than an offshore banking relationship persists without active maintenance), which lowers $\underline{\pi}$ proportionally. □

A.5.3 The Self-Reinforcing Channel

The key modification from the standard Uribe model is that S is not exogenous. Stablecoin ecosystem size grows with mesh demand (Section 2.3) and with dollarization itself (each new country that dollarizes adds users to the stablecoin ecosystem). This creates a self-reinforcing dynamic:

(i) Mesh growth increases S (settlement demand).

(ii) Higher S lowers $\bar{\pi}(S)$ and $\underline{\pi}(S)$.

(iii) Countries previously in the "stable" zone ($\pi < \bar{\pi}$) now find themselves in the multiple-equilibria zone or above the upper threshold.

(iv) Those countries dollarize, adding users to the stablecoin ecosystem.

(v) S grows further. Return to step (ii).

Definition A.5.2 (Dollarization Reproduction Number)**.** *The dollarization reproduction number is:*

$$R_0^{dollar} = \frac{dS}{S} \cdot \frac{1}{dt}\bigg|_{dollarization\ channel} \tag{A.16}$$

measuring the proportional growth rate of S attributable to the dollarization spiral. If $R_0^{dollar} > 1$, each currency collapse feeds the next: the system is in a self-reinforcing spiral.

A.5.4 Mapping to the Six-Stage Classification

Chapter 7's six-stage country classification (Pre-Industrial, Early Industrial, Emerging Industrial, Mature Industrial, Post-Industrial, and Advanced Post-Industrial) can be mapped to positions relative to the modified thresholds.

The critical observation: as S grows (driven by mesh demand and prior dollarization), the rows shift downward. Countries currently in the "stable" category become "multiple equilibria," and countries in the "multiple equilibria" category become "inevitable." The speed of this shift is determined by the rate of mesh growth and the parameter β_π governing threshold sensitivity to ecosystem size.

A.6 The Triffin Contradiction

Triffin [21] identified the fundamental tension in a reserve currency system: the world needs dollars for international transactions, but supplying those dollars requires the US to run

Table A.1: Country Groups and Dollarization Vulnerability

Country Group	**Typical π**	**Position at S_{2026}**	**Vulnerability**
Pre-Industrial	15–40%	$\pi > \bar{\pi}(S)$	Already past threshold; dollarization inevitable with stablecoin access
Early Industrial	8–20%	$\underline{\pi} < \pi < \bar{\pi}$	Multiple equilibria; vulnerable to coordination on dollarization
Emerging Industrial	4–10%	Near $\underline{\pi}(S)$	Currently stable; becomes vulnerable as S grows
Mature Industrial	2–5%	$\pi < \underline{\pi}(S)$	Stable at current S; threshold declines may reach them at high S
Post-Industrial	1–3%	$\pi \ll \underline{\pi}(S)$	Not vulnerable to dollarization spiral
Advanced Post-Ind.	1–3%	$\pi \ll \underline{\pi}(S)$	Not vulnerable; may benefit from stablecoin ecosystem as settlement infrastructure

deficits, which eventually undermines confidence in the dollar. Farhi and Maggiori [19] formalize this as a model of the international monetary system with three zones for the reserve currency issuer. This section extends their framework with endogenous instability zone boundaries.

A.6.1 The Farhi-Maggiori Framework

In the Farhi-Maggiori model, the reserve currency issuer has debt-to-GDP ratio b. Three zones characterize the stability of the reserve asset:

(i) *Safety zone* ($b \leq \bar{b}$): Treasury debt is unconditionally safe. No crisis equilibrium exists regardless of investor coordination. Stablecoins backed by Treasuries are sound.

(ii) *Instability zone* ($\bar{b} < b \leq \tilde{b}$): Multiple equilibria. A "no-crisis" equilibrium and a "crisis" equilibrium both exist. The outcome depends on investor coordination—a sunspot can trigger a switch from the good to the bad equilibrium. The crisis probability $\alpha(b)$ is increasing in b within this zone.

(iii) *Collapse zone* ($b > \tilde{b}$): The crisis equilibrium is unique. Treasury debt is no longer safe. Stablecoin backing fails.

A.6.2 Endogenous Instability Boundaries

The safety threshold $\bar{b}$ is defined by the property that Treasury debt is *information-insensitive* in the sense of Gorton [24]: the debt is safe enough that investors do not find it profitable to produce information about the issuer's ability to repay. Information-insensitivity is what makes an asset "safe"—it can be traded without adverse selection because no one has an informational advantage.

Mesh agents destroy information-insensitivity. They process sovereign fiscal data continuously, produce real-time estimates of default probability, and trade on these estimates at machine speed. From a mesh agent's perspective, *every* debt instrument is information-sensitive because the cost of processing the information is near zero ($c_M \ll c_H$).

Proposition A.6.1 (Shrinking Safety Zone)**.** *The safety zone threshold $\bar{b}(\phi)$ is decreasing in mesh participation:*

$$\frac{d\bar{b}}{d\phi} < 0 \tag{A.17}$$

The safety zone shrinks as mesh agents enter sovereign debt markets. Formally, $\bar{b}(\phi)$ is defined by the condition that the expected return to producing information about the sovereign's fiscal position is exactly $c(\phi)$—the (weighted average) cost of information. As $c(\phi)$ declines with ϕ (equation A.2), lower levels of debt become "information-sensitive": it becomes profitable to trade on sovereign fiscal fundamentals even at debt levels that were previously considered unconditionally safe.

Proof. In the Gorton [24] framework, an asset is information-insensitive when the expected gain from producing information is less than the cost: $\mathbb{E}[\pi_{\text{info}}(b)] < c$. The expected gain $\mathbb{E}[\pi_{\text{info}}(b)]$ is increasing in b (higher debt means more uncertainty about repayment, hence greater returns to information production). The threshold $\bar{b}$ is defined by $\mathbb{E}[\pi_{\text{info}}(\bar{b})] = c$. Since $c(\phi) = (1-\phi)c_H + \phi c_M$ is decreasing in ϕ, the threshold $\bar{b}(\phi)$ satisfying $\mathbb{E}[\pi_{\text{info}}(\bar{b})] = c(\phi)$ must also decrease: lower information cost means lower debt levels suffice to make information production profitable. □

A.6.3 Transformation of Crisis Dynamics

Within the instability zone, the character of crises changes as ϕ increases.

In the classical Farhi-Maggiori model, crises in the instability zone are *sunspot-driven*: they are self-fulfilling events triggered by a coordination failure among investors. If investors collectively believe a crisis will occur, their flight from Treasury debt validates the belief. The crisis probability depends on coordination dynamics (who moves first, what signals trigger coordination) rather than on fundamentals alone.

With high mesh participation, crises become *fundamentals-driven*. Mesh agents all observe the same fiscal data and process it to the same conclusion. There is no coordination problem—the agents independently compute the same optimal response. The "sunspot" is replaced by a deterministic threshold: when fiscal fundamentals cross a precisely computable boundary, all mesh agents adjust their portfolios simultaneously.

Proposition A.6.2 (Crisis Character Transition)**.** *As ϕ increases, the crisis probability $\alpha(b, \phi)$ in the instability zone transitions from sunspot-driven to fundamentals-driven:*

$$\alpha(b, \phi) = (1 - \phi) \cdot \alpha_{sunspot}(b) + \phi \cdot \mathbf{1}[b > b^*(\phi)] \tag{A.18}$$

where $\alpha_{sunspot}(b)$ is the classical sunspot-driven crisis probability (smooth, increasing in b) and $b^*(\phi)$ is the fundamentals-driven crisis threshold (a deterministic boundary). At high ϕ, the smooth $\alpha_{sunspot}$ is replaced by the sharp indicator function $\mathbf{1}[b > b^*]$.

This transition has ambiguous welfare implications. On one hand, it is *stabilizing*: self-fulfilling panics—low-probability but catastrophic events—are eliminated. Mesh agents do not panic; they compute. On the other hand, it is *destabilizing*: the fundamentals-driven threshold is sharp, and crossing it triggers immediate, irreversible repricing. There is no warning period, no opportunity for gradual fiscal adjustment. The system trades tail risk (rare catastrophic crises) for continuous repricing pressure (fiscal deterioration is immediately reflected in yields).

A.6.4 The Contradiction Formalized

The stablecoin ecosystem requires US Treasuries as backing. As stablecoin market capitalization grows—projections range from $1–3 trillion by 2030 (various industry estimates)—Treasury demand from stablecoin issuers becomes a significant fraction of total Treasury demand. This creates the modern Triffin contradiction:

(i) The world demands dollar-denominated safe assets (Caballero, Farhi and Gourinchas [20]). Stablecoins are a delivery mechanism for this demand.

(ii) Meeting the demand requires Treasury issuance, which pushes b upward.

(iii) Mesh agents make the safety zone smaller ($d\bar{b}/d\phi < 0$).

(iv) Therefore: stablecoin demand pushes b toward the instability zone while mesh participation makes the instability zone more dangerous.

Corollary A.6.3 (Triffin Squeeze)**.** *The "Triffin squeeze" is the condition:*

$$\frac{db}{dt} > 0 \quad \textit{and} \quad \frac{d\bar{b}}{dt} < 0 \tag{A.19}$$

Both hold simultaneously when stablecoin demand is growing ($\dot{S} > 0$, pushing b up) and mesh participation is increasing ($\dot{\phi} > 0$, pushing $\bar{b}$ down). The time to contact $b = \bar{b}$ is:

$$T_{\textit{Triffin}} = \frac{\bar{b}(0) - b(0)}{\dot{b}(0) + |\dot{\bar{b}}(0)|} \tag{A.20}$$

If $T_{\textit{Triffin}}$ is finite and shorter than the time scale of fiscal adjustment, the system enters the instability zone before the sovereign can respond.

A.6.5 The Baumol-Triffin Equivalence

The hierarchical ceiling cascade (Chapter 3, [2], Proposition 8.1) implies that the Triffin contradiction at the settlement level is structurally identical to the Baumol bottleneck at the capability level—both arise from the slow manifold of the level below constraining the equilibrium of the level above. At Level 3, capability growth is bounded by the network formation rate: $F_3 \leq (\varphi_{\text{eff}}/\delta_C) \cdot F_{\text{CES}}(N^*(F_1))$. At Level 4, settlement growth is bounded by capability: $F_4 \leq \bar{S}(F_3)$. But the Triffin squeeze adds a reflexive element absent from the Baumol bottleneck: the settlement layer's demand for safe assets *transforms* the fiscal dynamics of the sovereign that provides those assets. The Baumol bottleneck is a passive constraint (frontier training does not change because the mesh demands more of it); the Triffin squeeze is an active constraint (Treasury supply dynamics change because stablecoin demand alters the debt-to-GDP trajectory). Both are instances of a faster sector bounded by its slower parent, but the Triffin version exhibits the feedback that makes the coupled system of Section 7 necessary.

The resolution of the Triffin contradiction depends on which of several mechanisms intervenes: fiscal adjustment (reducing $\dot{b}$), alternative backing assets (reducing the dependence on Treasuries), institutional innovation (restructuring the reserve currency system), or the system entering the instability zone and experiencing a crisis-driven adjustment. The model does not predict which mechanism obtains; it characterizes the conditions under which each is needed.

A.7 The Coupled System—Equilibrium Characterization

This is the chapter's central section. The preceding analyses of market efficiency (Section 3), monetary policy (Section 4), dollarization (Section 5), and the Triffin contradiction (Section 6) are unified into a single coupled dynamical system.

A.7.1 State Variables

The system is characterized by four state variables:

(i) $\phi(t) \in [0, 1]$: the fraction of capital managed by autonomous mesh agents.

(ii) $S(t) \geq 0$: the stablecoin ecosystem size (measured in total stablecoin market capitalization or total user base).

(iii) $b(t) \geq 0$: the US Treasury debt-to-GDP ratio available as safe-asset backing.

(iv) $\eta(t) \in [0, 1]$: financial sector capitalization as a fraction of total wealth, following Brunnermeier and Sannikov [12]. This measures financial system health: high η corresponds to a well-capitalized financial sector; low η corresponds to a crisis state.

A.7.2 The Coupled ODE System

The deterministic skeleton of the coupled system is:

$$\dot{\phi} = f_\phi(\phi, S, \eta) = \gamma_\phi \cdot \phi(1 - \phi) \cdot [\mu_\phi(S, \eta) - r_\phi] \tag{A.21}$$

$$\dot{S} = f_S(\phi, S, b) = \gamma_S \cdot S \cdot [g_{\text{mesh}}(\phi) + g_{\text{dollar}}(S, b) - \delta_S] \tag{A.22}$$

$$\dot{b} = f_b(S, \eta, b) = \gamma_b \cdot [d(b, \eta) + s_{\text{coin}}(S) - \tau(b)] \tag{A.23}$$

$$\dot{\eta} = f_\eta(\phi, b, S, \eta) = \mu_\eta(\phi) \cdot \eta - \sigma_\eta^2(\phi, S) \cdot \eta \cdot (1 - \eta) - \ell(b, \eta) \tag{A.24}$$

The interpretation of each equation follows.

Mesh participation (equation A.21): ϕ evolves as a logistic process—bounded between 0 and 1—with growth driven by the excess return $\mu_\phi(S, \eta) - r_\phi$ from mesh-managed capital relative to the reservation return r_ϕ. Better settlement infrastructure (higher S) increases the return to mesh participation. Financial sector health (η) affects the return through the

stability of market infrastructure. The logistic form $\phi(1-\phi)$ ensures the natural bounds and captures the S-shaped adoption dynamics.

Stablecoin ecosystem (equation A.22): S grows from two sources—mesh demand $g_{\text{mesh}}(\phi)$ (which is increasing in ϕ, as more mesh participation requires more settlement) and dollarization $g_{\text{dollar}}(S, b)$ (which captures the self-reinforcing dollarization spiral of Section 5). The dollarization term is increasing in S (network effects) and increasing in the number of countries past the modified threshold $\bar{\pi}(S)$, which itself depends on b through the fiscal channel. Depreciation δ_S captures user attrition, regulatory withdrawals, and technology obsolescence.

Treasury debt ratio (equation A.23): b increases with the primary deficit $d(b, \eta)$, which depends on existing debt service and financial sector health (crises increase fiscal costs), and with stablecoin-driven Treasury demand $s_{\text{coin}}(S)$ (stablecoin issuers buying Treasuries as backing). It decreases with tax revenue $\tau(b)$, which depends on the tax base and growth rate of the economy.

Financial sector capitalization (equation A.24): η follows a Brunnermeier-Sannikov-type law of motion. The drift $\mu_\eta(\phi)$ captures expected returns to financial intermediation, which increase with ϕ in the initial phase (mesh agents improve market efficiency, increasing trading profits) but may decline at high ϕ (as mesh agents replace traditional intermediaries). The diffusion term $\sigma_\eta^2(\phi, S) \cdot \eta(1-\eta)$ captures endogenous volatility: higher ϕ reduces exogenous noise but increases the amplitude of fundamentals-driven repricing. The loss function $\ell(b, \eta)$ captures the impact of sovereign stress on financial institutions holding Treasury assets.

A.7.3 Steady-State Analysis

Setting $\dot{\phi} = \dot{S} = \dot{b} = \dot{\eta} = 0$, we characterize the steady states.

Theorem A.7.1 (Equilibrium Characterization)**.** *The coupled system* (A.21)–(A.24) *admits (generically) three classes of steady states:*

(i) Low-mesh equilibrium $(\phi^L, S^L, b^L, \eta^L)$*: ϕ^L is small (mesh participation is minimal), S^L is small (stablecoin ecosystem serves niche use cases), b^L is in the Farhi-Maggiori safety zone ($b^L < \bar{b}(\phi^L)$), and η^L is at the Brunnermeier-Sannikov ergodic mean. This equilibrium corresponds approximately to the current financial system. Monetary policy is effective:* $\text{MP}(\phi^L, S^L) \approx \text{MP}_0$*. Treasuries are safe. The mesh exists but does not dominate capital markets.*

(ii) High-mesh equilibrium $(\phi^H, S^H, b^H, \eta^H)$*: ϕ^H is large (mesh agents manage most capital), S^H is large (stablecoins are a dominant settlement medium), b^H is in the safety*

zone only if fiscal fundamentals are sound ($b^H < \bar{b}(\phi^H)$ requires fiscal discipline because $\bar{b}(\phi^H) < \bar{b}(\phi^L)$), and η^H is determined by the new market structure. Monetary policy is weak: $\text{MP}(\phi^H, S^H) \ll \text{MP}_0$*. But market discipline substitutes: real-time fundamentals-based repricing constrains fiscal policy.*

(iii) Unstable intermediate region*: Between the low-mesh and high-mesh equilibria lies a saddle region where the dynamics are path-dependent. Small perturbations can push the system toward either equilibrium, depending on the direction of the perturbation and the local stability properties.*

Proof. The proof proceeds by analyzing the Jacobian of the system at each candidate steady state.

Low-mesh equilibrium. At $\phi^L \approx 0$, the mesh participation equation (A.21) has $\dot{\phi} = 0$ at $\phi = 0$ (trivially) and at any ϕ where $\mu_\phi(S, \eta) = r_\phi$. The Jacobian evaluated at $(\phi^L, S^L, b^L, \eta^L)$ has all eigenvalues with negative real parts when:

$$\left.\frac{\partial f_\phi}{\partial \phi}\right|_L < 0, \quad \left.\frac{\partial f_S}{\partial S}\right|_L < 0, \quad \left.\frac{\partial f_b}{\partial b}\right|_L < 0, \quad \left.\frac{\partial f_\eta}{\partial \eta}\right|_L < 0 \tag{A.25}$$

The first condition requires that the excess return to mesh participation is decreasing in ϕ at the low-mesh steady state—i.e., the marginal mesh agent earns less than the reservation return, so ϕ does not grow. This holds when the settlement infrastructure is too thin to support efficient mesh operation.

High-mesh equilibrium. At ϕ^H near 1, the logistic term $\phi(1-\phi)$ naturally damps growth. The system reaches a steady state where the return to the marginal mesh agent (given the large settlement infrastructure S^H) exactly equals the reservation return. Stability requires that the coupled dynamics are locally dissipative: perturbations in any variable are corrected by the feedback structure. The key condition is that the Triffin squeeze (Corollary A.6.3) has been resolved—either $b^H < \bar{b}(\phi^H)$ (fiscal adjustment has occurred) or the system has transitioned to fundamentals-based pricing that accommodates the debt level.

Unstable intermediate. Between the two stable equilibria, there exists a separatrix in the four-dimensional state space. On one side of the separatrix, trajectories converge to the low-mesh equilibrium; on the other, they converge to the high-mesh equilibrium. The separatrix passes through a saddle point (or saddle manifold) that is unstable in at least one direction. The existence of this intermediate structure follows from the continuity of the vector field and the existence of two stable equilibria: by the intermediate value theorem applied to the flow, there must be an intermediate critical point with mixed stability. □

A.7.4 The Settlement Reproduction Number

Definition A.7.2 (Settlement Reproduction Number)**.** *The settlement reproduction number R_0^{settle} measures the strength of the mesh–financial system feedback loop:*

$$R_0^{settle} = \frac{\partial f_\phi}{\partial S} \cdot \frac{\partial f_S}{\partial \phi} \cdot \left(\frac{\partial f_\phi}{\partial \phi}\right)^{-1} \cdot \left(\frac{\partial f_S}{\partial S}\right)^{-1} \tag{A.26}$$

evaluated at the low-mesh steady state. When $R_0^{settle} > 1$, the feedback loop is self-reinforcing: a perturbation increasing ϕ increases S (through mesh settlement demand), which increases ϕ (through better settlement infrastructure attracting more mesh participation), and the amplification exceeds the initial perturbation.

Proposition A.7.3 (Transition Condition)**.** *The low-mesh equilibrium loses stability (and the system transitions toward the high-mesh equilibrium) when R_0^{settle} crosses unity from below. The transition is a transcritical bifurcation: the low-mesh and intermediate steady states collide and exchange stability properties. Above the bifurcation, the only stable equilibrium is the high-mesh state.*

Proof. At the low-mesh steady state, the Jacobian J of the system restricted to the (ϕ, S) subsystem is:

$$J_{(\phi,S)} = \begin{pmatrix} \partial f_\phi/\partial \phi & \partial f_\phi/\partial S \\ \partial f_S/\partial \phi & \partial f_S/\partial S \end{pmatrix} \tag{A.27}$$

Both diagonal entries are negative at the stable low-mesh equilibrium (self-damping). The off-diagonal entries are positive (mutual reinforcement: more mesh $\to$ more stablecoins, and more stablecoins $\to$ more mesh). The eigenvalues of $J_{(\phi,S)}$ have negative real parts when $\det(J) > 0$ and $\mathrm{tr}(J) < 0$. The determinant condition is:

$$\det(J) = \frac{\partial f_\phi}{\partial \phi} \cdot \frac{\partial f_S}{\partial S} - \frac{\partial f_\phi}{\partial S} \cdot \frac{\partial f_S}{\partial \phi} > 0 \tag{A.28}$$

Dividing by $(\partial f_\phi/\partial \phi)(\partial f_S/\partial S) < 0 \cdot < 0 > 0$:

$$1 - R_0^{\text{settle}} > 0 \quad \Longleftrightarrow \quad R_0^{\text{settle}} < 1 \tag{A.29}$$

When R_0^{settle} crosses 1, $\det(J)$ passes through zero, indicating a bifurcation. The transcritical structure follows from the logistic form of equation (A.21): the $\phi = 0$ equilibrium and the interior equilibrium collide at the bifurcation point, as in the standard epidemiological R_0 framework. □

A.7.5 Transition Dynamics

The character of the transition from low-mesh to high-mesh equilibrium depends on the speed mismatch between market adaptation and institutional adaptation.

Market adaptation is fast. Mesh agents adjust portfolios in milliseconds. Price discovery occurs at machine speed. The ϕ and S dynamics can move on timescales of weeks to months.

Institutional adaptation is slow. Fiscal policy adjusts over years (legislative cycles, budget processes). Monetary policy frameworks evolve over decades (the inflation targeting consensus took 20 years to develop). International monetary system reform operates on generational timescales (Bretton Woods to the current non-system took 50 years).

When $R_0^{\text{settle}} > 1$, the fast variables (ϕ, S) grow more rapidly than the slow variables (b, η) can adjust. This speed mismatch produces the "messy transition"—a period in which the financial system is adapting to mesh participation faster than institutions can respond. The messy transition may be smooth (if R_0^{settle} is only slightly above 1 and institutional adaptation keeps pace) or crisis-punctuated (if R_0^{settle} is well above 1 and the speed mismatch is large).

Remark A.7.4 (Path Dependence in the Transition). The transition path through the unstable intermediate region is path-dependent. Two systems starting from similar initial conditions but experiencing different sequences of shocks (fiscal crises, stablecoin adoption events, regulatory changes) can arrive at different points in the high-mesh equilibrium—one with fiscal discipline intact and sound Treasury backing, the other with a degraded sovereign balance sheet and fragile stablecoin infrastructure. The model does not predict which path obtains; it characterizes the conditions (initial b, speed of institutional adaptation, magnitude of R_0^{settle}) that determine the probability of each.

A.8 Implications for Sovereign Fiscal Policy

In the high-mesh equilibrium, the operating environment for fiscal policy changes qualitatively. This section derives the "synthetic gold standard" result: real-time market discipline constrains sovereign fiscal policy in ways analogous to—but distinct from—the constraints imposed by the classical gold standard.

A.8.1 The Synthetic Gold Standard

Under the classical gold standard, governments faced a hard constraint: gold reserves limited money supply expansion, and fiscal profligacy drained reserves, forcing adjustment. The constraint was physical (finite gold) and institutional (convertibility commitments).

In the high-mesh equilibrium, the constraint is informational and market-mediated. Governments can still issue debt, but:

(i) *Yields reflect fundamentals, not policy manipulation.* With $\text{MP}(\phi^H, S^H) \approx 0$ for the portfolio balance and financial repression channels, the central bank cannot compress yields below market-clearing levels. Bond yields are determined by mesh agents' real-time assessment of fiscal sustainability.

(ii) *Fiscal crises occur at machine speed.* When b crosses the fundamentals-driven threshold $b^*(\phi^H)$ (Proposition A.6.2), repricing is immediate. The adjustment that previously took weeks or months—as human portfolio managers gradually revised their positions—occurs in seconds as mesh agents simultaneously rebalance.

(iii) *Financial repression is impossible.* With $S^H > S_{\text{crit}}$, captive savers have exited to stablecoins. The government cannot compel domestic savers to hold negative-real-return bonds because the exit option is available.

(iv) *Surviving policy tools are the interest rate channel and LOLR.* The interest rate channel still operates because borrowing costs affect real activity through physical-economy mechanisms. The LOLR function survives because the central bank can create reserves, though its effectiveness in a stablecoin-denominated crisis is limited.

Proposition A.8.1 (Endogenous Fiscal Discipline)**.** *In the high-mesh equilibrium, the yield spread on sovereign debt satisfies:*

$$y(b) - r_f = \theta(\phi^H) \cdot \max(0, b - b^*(\phi^H)) + \varepsilon_{term} \tag{A.30}$$

where $\theta(\phi^H) > 0$ is the repricing coefficient (increasing in ϕ because more mesh agents means faster and more complete repricing), $b^*(\phi^H)$ is the fundamentals-based threshold, and ε_{term} is a term premium that is near zero at high ϕ (mesh agents arbitrage term premia efficiently). The spread is zero when $b < b^*$ and linear in the excess when $b > b^*$. This constitutes a hard constraint on fiscal policy: each unit of excess debt increases borrowing costs by θ, which is large because the repricing is immediate and complete.

A.8.2 Historical Analogy and Disanalogy

The synthetic gold standard shares with the classical gold standard the property of constraining fiscal discretion through an external discipline mechanism. But the analogy is imperfect in important ways.

Under the gold standard, the constraint was binary: either the country maintained convertibility or it didn't. Adjustment was discrete (devaluation events). Under the synthetic gold standard, the constraint is continuous: each increment of fiscal deterioration produces an immediate, proportionate yield response. This is both more efficient (gradual feedback rather than sudden crises) and more demanding (no period of "getting away with it" before the market notices).

The classical gold standard was defeatable by suspending convertibility—as occurred during World War I and the Great Depression. The synthetic gold standard is not defeatable by government decree. A government can default on its debt, but it cannot prevent mesh agents from pricing the default risk. It can impose capital controls, but it cannot prevent stablecoin-mediated capital flight (unless it can block all internet access). The constraint is embedded in the information infrastructure rather than in an institutional arrangement.

The analogy to the transition from the gold standard to floating exchange rates is instructive. When that transition occurred, the rules of macroeconomic management changed fundamentally. Governments that adapted to the new rules (inflation targeting, fiscal rules, independent central banks) prospered; those that did not (various episodes of hyperinflation in the 1970s-80s) suffered. The transition to the synthetic gold standard is a comparable regime change. The chapter does not predict which governments will adapt successfully; it characterizes the constraints they will face.

A.9 Frameworks Considered and Rejected

Several candidate frameworks were evaluated and rejected for specific technical reasons.

Mean Field Games (Lasry-Lions [37]). As in Chapter 5 (Section 9), MFG assumes exchangeable agents. Mesh agents are heterogeneous specialists with different information sets, objectives, and time horizons. By the CES Triple Role ([1], Theorem 7.1), the curvature parameter $K > 0$ ensures non-exchangeability when $\rho < 1$. MFG would average over the heterogeneity that drives both the efficiency results (Section 3) and the collusion resistance conjecture (Section 3.4).

Minsky Financial Instability Hypothesis. The Minsky cycle (stability $\to$ leverage $\to$ fragility $\to$ crisis) is conceptually relevant—the low-volatility environment created by mesh agents could breed endogenous leverage. However, Minsky's framework is insufficiently formalized for the equilibrium characterization this chapter requires. The Brunnermeier-Sannikov [12] framework captures the same insight rigorously through the η dynamics (equation A.24): the volatility paradox is the formal analog of the Minsky mechanism.

Bitcoin Maximalism / Austrian Monetary Theory. The chapter does not predict

Bitcoin replacing the dollar. It predicts dollar-denominated stablecoins as the settlement medium, which *strengthens* the dollar as unit of account while *weakening* the Federal Reserve's control over dollar-denominated markets. This is the opposite of the Bitcoin maximalist thesis (dollar collapse) and the opposite of the Austrian thesis (return to commodity money). The dollar becomes more dominant, not less; what changes is who controls the terms.

Full Continuous-Time GE with Heterogeneous Agents. A continuous-time general equilibrium model with heterogeneous agents, incomplete markets, and stochastic shocks would be the "right" model for the coupled system. It would also be intractable. The four-ODE deterministic skeleton captures the qualitative dynamics: the number and stability of equilibria, the bifurcation structure, and the transition conditions. A full stochastic treatment would enrich the analysis (adding risk premia, precautionary savings, and tail risk) but would not change the equilibrium characterization. The stochastic extension is deferred to a supplement.

A.10 Falsifiable Predictions

The model generates six predictions extending the prediction sets of the preceding chapters. Each specifies timing, observable metrics, and conditions under which the prediction would be falsified.

Prediction 1: Market efficiency increase with mesh participation. As autonomous agent assets under management (AUM) grow, measurable market efficiency metrics improve. Specifically: (i) return autocorrelation at daily frequency decreases by at least 30% as autonomous agent AUM reaches 15% of total market capitalization; (ii) bid-ask spreads in Treasury markets narrow by at least 20% over the same period; (iii) the Hurst exponent for major equity indices moves closer to 0.5 (random walk). *Timing:* 2028–2033. *Evidence against:* market efficiency metrics failing to improve or worsening as autonomous agent AUM grows, sustained over a period where AUM exceeds 10% of market capitalization.

Prediction 2: Forward guidance effectiveness decline. The duration of market impact from central bank forward guidance decreases as mesh participation increases. The observable metric is the time from FOMC statement release to 90% of the ultimate yield response. The model predicts this time declines from the current ∼2–4 weeks to less than 1 trading day as ϕ crosses 0.3 (approximately 30% of Treasury market volume from autonomous agents). *Timing:* 2029–2034. *Evidence against:* forward guidance impact duration remaining stable or increasing as autonomous agent participation in Treasury markets grows above 20%.

Prediction 3: Dollarization threshold decline. Countries that previously main-

tained stable domestic currencies at inflation rates of 12–15% will experience stablecoin-mediated dollarization spirals at 8–10% inflation as the stablecoin ecosystem matures. The observable metric is the critical inflation rate at which stablecoin adoption exceeds 10% of the domestic money supply, cross-sectionally across emerging-market countries. The model predicts this threshold declines by approximately 3–5 percentage points between 2026 and 2032, corresponding to the growth of S from ∼$200 billion to ∼$1–2 trillion. *Timing:* 2028–2032. *Evidence against:* the critical inflation threshold remaining constant or rising despite stablecoin ecosystem growth.

Prediction 4: Stablecoin-Treasury yield link strengthening. Ahmed and Aldasoro [22] estimate the elasticity of T-bill yields to stablecoin market capitalization flows. The model predicts this elasticity approximately doubles as stablecoin market cap grows from ∼$200 billion to ∼$1 trillion. The mechanism: larger stablecoin reserve holdings make Treasury markets more sensitive to stablecoin flows because the marginal dollar of stablecoin demand represents a larger fraction of marginal Treasury supply. *Timing:* 2027–2030. *Evidence against:* the elasticity remaining constant or declining as stablecoin market cap grows.

Prediction 5: Settlement constraint relaxation timing. Mesh capability growth accelerates following stablecoin infrastructure improvements—new on-ramp launches, regulatory clarity events, protocol upgrades. The observable metric is the correlation between settlement infrastructure milestones and subsequent mesh capability growth (measured by benchmark performance or AUM growth). The model predicts a statistically significant positive correlation (at the 5% level) between settlement infrastructure events and mesh growth in the subsequent 6 months, consistent with the settlement layer being a binding constraint on mesh growth (Chapter 5, Part II, Section 8). *Timing:* 2028–2035. *Evidence against:* no significant correlation between settlement infrastructure improvements and mesh growth, or mesh growth showing no sensitivity to settlement infrastructure quality.

Prediction 6: Fiscal crisis speed acceleration. Sovereign debt repricing events become faster as mesh participation in Treasury markets increases. The observable metric is the time constant of yield adjustment following significant fiscal news (deficit announcements, rating changes, fiscal policy shifts). The model predicts the time constant decreases from the current ∼2–6 weeks to ∼1–3 trading days as autonomous agent participation reaches 25% of Treasury market volume. *Timing:* 2030–2035. *Evidence against:* sovereign repricing speed remaining constant or slowing despite increased autonomous agent participation.

A.11 Preliminary Empirical Evidence

Three of the six predictions above can be tested against currently available data. The results are directionally consistent but generally lack the statistical power needed for definitive confirmation—as expected for predictions designed to be tested over 2027–2035 horizons with data that, in most cases, only begins in 2020.

A.11.1 Prediction 4: Stablecoin–Treasury Yield Link

Weekly data (2020–2026, $N = 319$) from DeFiLlama (stablecoin market capitalization) and FRED (Treasury yields, VIX, federal funds rate) test whether stablecoin growth predicts Treasury yield movements.[1]

In the level regression (3-month Treasury yield on log stablecoin market cap, controlling for the federal funds rate and VIX), the coefficient on ln(mcap) is +0.041 ($p = 0.001$, $R^2 = 0.992$). The high R^2 reflects co-movement with the federal funds rate; the partial effect of stablecoin growth is economically small but statistically significant.

The first-difference specification (changes in yield on changes in log market cap) provides cleaner identification: $\hat{\beta} = -0.300$ (SE = 0.178, $p = 0.092$). The negative sign indicates that stablecoin inflows are associated with *lower* short-term Treasury yields, consistent with the mechanism: growing stablecoin reserves increase demand for T-bills, compressing yields at the short end. The effect is marginally significant.

Granger causality tests reveal bidirectional feedback: stablecoin growth marginally predicts yield changes ($p = 0.096$ at lag 2), but yield changes predict stablecoin growth more strongly ($p < 0.001$ at lags 2–8). This reverse channel—higher yields attracting stablecoin issuance because reserves earn more—is itself predicted by the model (higher r increases the profitability of the stablecoin business model, expanding S).

At current scale (stablecoin market cap ∼$306B versus $6T in Treasury bills outstanding, a 5.1% ratio), the effects remain small. The prediction is that the elasticity approximately doubles as stablecoin market cap approaches $1 trillion.

A.11.2 Prediction 2: Forward Guidance Absorption Speed

Data from 213 FOMC meetings (2000–2026) test whether yield adjustment to FOMC statements has accelerated over time, using the ratio of the immediate (1-day) yield response to the 5-day cumulative response as a measure of absorption speed.[2]

[1] Scripts and replication data at `scripts/test_stablecoin_treasury.py`.
[2] Scripts and replication data at `scripts/test_fomc_impact.py`.

The simple trend regression yields a coefficient of +0.0043 per year ($p = 0.602$)—positive (directionally consistent with faster absorption) but far from significant. The prediction targets the period when autonomous agent Treasury market participation exceeds 30%, which has not yet occurred; the current test is a pre-period baseline.

Era comparison reveals a pattern: the mean adjustment ratio is 1.00 (2000–2009), 0.85 (2010–2014, the ZIRP era), 0.97 (2015–2019), and 1.11 (2020–2026). The post-2020 era shows the fastest absorption, though cross-era differences are not statistically significant (ANOVA $F = 0.64$, $p = 0.59$). The prediction's relevance begins in the 2028–2034 horizon; the current data establishes a baseline against which future acceleration can be measured.

A.11.3 Prediction 3: Dollarization Threshold Decline

Cross-country panel data (18 countries, 2020–2024, $N = 77$) test whether the inflation rate at which stablecoin adoption accelerates has declined over time, consistent with the prediction that larger stablecoin ecosystems lower the threshold.[3]

The logit specification (high-adoption binary on inflation, controlling for stablecoin ecosystem size) yields an inflation coefficient of -0.026 ($p = 0.203$), directionally consistent (higher inflation raises adoption probability) but not significant. An inflation-×-stablecoin interaction term is positive (+0.0005, $p = 0.958$), consistent with the amplification prediction but with no statistical power.

The most theory-relevant test examines the time-varying threshold directly. The estimated threshold declines from 15.5% (2022) to 11.1% (2023) to approximately 0% (2024), with a Spearman rank correlation of $\rho = -0.40$ between year and threshold. The direction is exactly as predicted—the threshold declines as the stablecoin ecosystem grows—but with only 5 annual observations, statistical power is negligible ($p = 0.50$). This test will gain power as the panel lengthens.

A.12 Conclusion

The mesh transforms the financial system it depends on. This is the settlement feedback: mesh growth generates stablecoin demand, stablecoin demand generates Treasury demand, mesh agents enter capital markets and transform price discovery, market transformation degrades monetary policy tools, monetary policy degradation combined with stablecoin access triggers dollarization in weak-currency countries, dollarization expands the stablecoin

[3]Scripts and replication data at `scripts/test_inflation_threshold.py`.

ecosystem, and the expanded ecosystem improves the settlement infrastructure that accelerates mesh growth.

The feedback loop is self-reinforcing when $R_0^{\text{settle}} > 1$: each cycle amplifies the next. Whether this condition holds depends on the strength of the coupling between mesh growth and settlement infrastructure quality, which is an empirical quantity that the model identifies but does not determine a priori.

Three contributions distinguish this analysis. First, the market microstructure results (Section 3) characterize the specific transition path of market efficiency as autonomous agents become the dominant capital market participants. The Grossman-Stiglitz paradox preserves a residual inefficiency, but the transition to near-complete price revelation changes the operating environment for every other market participant. Kyle's lambda is non-monotone, suggesting that the intermediate transition—not the endpoint—poses the greatest challenge for market microstructure.

Second, the monetary policy degradation results (Section 4) identify which tools break first and why. Forward guidance degrades first (it depends on processing delay), QE degrades second (it depends on arbitrage speed), and financial repression degrades last but most sharply (it depends on institutional barriers that collapse discontinuously when stablecoin access crosses a threshold). The composite monetary policy effectiveness $\text{MP}(\phi, S)$ is a declining function with a potential discontinuity—a prediction with clear observable consequences for how central banks should plan for the transition.

Third, the coupled system analysis (Section 7) unifies these individual mechanisms into a four-dimensional dynamical system whose equilibrium structure reveals two stable states (low-mesh and high-mesh) separated by an unstable intermediate. The transition is governed by R_0^{settle}, which has the same epidemiological R_0 structure that governs the mesh's internal dynamics (Chapter 5) and the endogenous decentralization crossing (Chapter 4). The R_0 framework is the unifying mathematical structure across the thesis: it governs mesh formation, capability growth, and now the coupling between the mesh and the financial system.

The endgame is not the collapse of the dollar. It is the strengthening of the dollar as unit of account while the Federal Reserve loses control over dollar-denominated markets. The result is a synthetic gold standard: market discipline, operated at machine speed, constrains sovereign fiscal policy. This constraint is not assumed; it is derived from the equilibrium properties of the coupled system. Whether it is desirable—whether real-time market discipline produces better or worse outcomes than the institutional discretion it replaces—is a normative question beyond this chapter's scope.

The chapters in this thesis now form a complete dynamical system. Concentrated invest-

ment (Chapter 4) creates the mesh (Chapter 5) that needs settlement infrastructure (Chapter 5, Part II, Section 8) and improves itself (Chapter 5, Part II) while transforming the financial system (Chapter 6) that funds the concentrated investment (back to Chapter 4). The outer loop is closed. The model is complete. The eigenstructure bridge (Chapter 3, Theorem 6.3) connects the technology Hessian to the welfare Hessian across all four levels, ensuring that the loop's equilibrium properties translate into welfare conclusions through the institutional supply-rate matrix.

What remains is the empirical program. The six predictions in this chapter, combined with the predictions in Chapters 4–5, generate a comprehensive set of testable implications spanning market microstructure, monetary policy, currency dynamics, and mesh capability growth. The predictions have specific timing, quantitative thresholds, and falsification conditions. The model earns its keep—or doesn't—over the next decade.

Bibliography

[1] Smirl, J. (2026a). The CES triple role: Superadditivity, correlation robustness, and strategic independence as three views of isoquant curvature. Chapter 2 of this thesis.

[2] Smirl, J. (2026b). Complementary heterogeneity: A port-Hamiltonian theory of the AI transition. Chapter 3 of this thesis.

[3] Smirl, J. (2026). Endogenous decentralization: How concentrated capital investment finances the learning curves that enable distributed alternatives. Chapter 4 of this thesis.

[4] Smirl, J. (2026). The mesh equilibrium: How heterogeneous specialized agents self-organize to exceed centralized provision after the crossing point. Chapter 5 of this thesis.

[5] Smirl, J. (2026). The autocatalytic mesh: Endogenous capability growth in self-organizing agent networks. Chapter 5, Part II of this thesis.

[6] Smirl, J. (2026). The monetary productivity gap. Chapter 7 of this thesis.

[7] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3), 393–408.

[8] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.

[9] Holden, C. W., & Subrahmanyam, A. (1992). Long-lived private information and imperfect competition. *Journal of Finance*, 47(1), 247–270.

[10] Duffie, D., Gârleanu, N., & Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, 73(6), 1815–1847.

[11] Dou, W. W., Goldstein, I., & Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. NBER Working Paper No. 34054.

[12] Brunnermeier, M. K., & Sannikov, Y. (2014). A macroeconomic model with a financial sector. *American Economic Review*, 104(2), 379–421.

[13] Brunnermeier, M. K., & Sannikov, Y. (2016). The I theory of money. Working paper, Princeton University.

[14] Woodford, M. (2003). *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

[15] Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46.

[16] Uribe, M. (1997). Hysteresis in a simple model of currency substitution. *Journal of Monetary Economics*, 40(1), 185–202.

[17] Calvo, G. A. (1998). Capital flows and capital-market crises: The simple economics of sudden stops. *Journal of Applied Economics*, 1(1), 35–54.

[18] Obstfeld, M. (1996). Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3–5), 1037–1047.

[19] Farhi, E., & Maggiori, M. (2018). A model of the international monetary system. *Quarterly Journal of Economics*, 133(1), 295–355.

[20] Caballero, R. J., Farhi, E., & Gourinchas, P.-O. (2017). The safe assets shortage conundrum. *Journal of Economic Perspectives*, 31(3), 29–46.

[21] Triffin, R. (1960). *Gold and the Dollar Crisis: The Future of Convertibility*. Yale University Press.

[22] Ahmed, R., & Aldasoro, I. (2025). Stablecoins and safe asset prices. BIS Working Paper No. 1270.

[23] Gorton, G. B., Klee, E., Ross, C., Ross, S. Y., & Vardoulakis, A. P. (2022). Leverage and stablecoin pegs. NBER Working Paper No. 30796.

[24] Gorton, G. B. (2017). The history and economics of safe assets. *Annual Review of Economics*, 9, 547–586.

[25] Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3), 401–419.

[26] Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press.

[27] Jones, C. I. (2015). Pareto and Piketty: The macroeconomics of top income and wealth inequality. *Journal of Economic Perspectives*, 29(1), 29–46.

[28] Gabaix, X., Lasry, J.-M., Lions, P.-L., & Moll, B. (2016). The dynamics of inequality. *Econometrica*, 84(6), 2071–2111.

[29] Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal*, 99(394), 116–131.

[30] Arthur, W. B. (1994). *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press.

[31] Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review*, 75(3), 424–440.

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

[33] Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71–S102.

[34] Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4), 1104–1144.

[35] Aghion, P., Jones, B. F., & Jones, C. I. (2018). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence* (pp. 237–282). University of Chicago Press.

[36] Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review*, 57(3), 415–426.

[37] Lasry, J.-M., & Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1), 229–260.

Appendix B

The Monetary Productivity Gap

The Monetary Productivity Gap

Structural Transformation, AI, and Endogenous Monetary Regime Choice in Developing Economies

Connor Smirl

EC 118: Growth Economics • Thesis

Tufts University • Spring 2026

Abstract

Gollin, Lagakos, and Waugh [12] document that value added per worker is systematically higher in non-agricultural sectors, yet institutional frictions keep labor misallocated in agriculture, particularly in developing economies. This paper identifies an analogous *monetary productivity gap*: as AI drives down the cost of cognitive output, economic activity transacted on programmable cryptocurrency rails exhibits higher effective productivity than equivalent activity on fiat infrastructure, yet institutional frictions keep most economic activity denominated in fiat. The gap has two components: a *transfer cost gap* (fiat remittances cost 6.4 percentage points more than stablecoin transfers across 300 corridors) and a larger *yield access gap* (1.4 billion unbanked adults earn deeply negative real returns on cash savings while tokenized US Treasuries offer 4.5% nominal—a value-added-per-dollar differential of approximately 30 percentage points, directly parallel to Gollin's sectoral value-added ratios). By early 2026, both components are observable: the x402 agentic payment protocol has processed over $600 million in autonomous AI-agent transactions on programmable rails; BlackRock's tokenized Treasury fund (BUIDL) has reached $500 million; and total tokenized real-world assets exceed $12 billion. Using a two-sector model of endogenous monetary regime choice (formalized in the Model Appendix) embedded in an extended Solow growth framework, parameterized by UN World Population Prospects 2024 demographic projections for 40 nations classified into six industrialization stages, I find that: (1) the

monetary productivity gap is largest in pre-industrial economies with weak fiat institutions, mirroring the pattern of agricultural productivity gaps across development stages; (2) the transition from fiat to programmable monetary infrastructure exhibits a cold-start problem requiring institutional catalysts; (3) sovereign accommodation of the transition dominates resistance across all six stages, with output gains of 11–13% by 2050; and (4) the demographic center of gravity is shifting toward nations where the monetary productivity gap is widest. The framework extends structural transformation analysis to the monetary system, treating monetary regime as an endogenous sectoral choice rather than exogenous infrastructure.

1. Introduction

The structural transformation literature documents a persistent puzzle: in developing economies, value added per worker in non-agricultural sectors is several multiples of that in agriculture, yet labor remains misallocated in low-productivity agriculture due to institutional frictions, migration barriers, and human capital constraints [12]. This paper argues that an analogous structural gap is emerging in the monetary system.

The cost of AI cognitive output has fallen by roughly four orders of magnitude between 2020 and early 2026 [25]. As this cost decline continues—even if it moderates substantially—AI agents increasingly require monetary infrastructure with properties that fiat systems cannot provide: programmability, continuous availability, borderlessness, and the ability to transact without legal personhood. Cryptocurrency provides these properties. The result is a measurable gap: economic activity that could be transacted more efficiently on programmable rails remains stuck in fiat, just as labor that could be more productive in manufacturing remains stuck in agriculture.

I call this the *monetary productivity gap*. The analogy to Gollin, Lagakos, and Waugh [12] is deliberate and, I argue, structurally precise. In their framework, the agricultural productivity gap reflects a combination of measurement issues (differences in hours worked and human capital) and genuine misallocation due to institutional frictions. The monetary productivity gap similarly reflects both measurement challenges (how do we value transactions that fiat infrastructure cannot process at all?) and genuine misallocation due to regulatory barriers, legal ambiguity, and institutional inertia. Crucially, the gap has two components: a *transfer cost gap* (the price of moving money between systems) and a larger *yield access gap* (the difference in value added per dollar across monetary infrastructures). Gollin measures value added per worker, not the cost of the bus ticket from village to city. The monetary analogue must do the same: the yield access gap—approximately 30 percentage points for an unbanked Nigerian farmer—is the true parallel to Gollin's productivity ratios, and it is activated by the tokenization of financial assets on programmable rails (Section 2.4).

The paper makes three contributions. First, it develops a two-sector model (formal-

ized in the Model Appendix) in which economic activities choose between fiat and programmable monetary infrastructure based on relative transaction costs, institutional quality, and regulatory friction—directly analogous to the sectoral choice between agriculture and non-agriculture in the structural transformation literature. The model produces analytical results: multiple equilibria with a cold-start problem (Proposition 1), signed comparative statics with testable predictions (Proposition 2), welfare implications favoring accommodation (Proposition 3), and endogenous erosion of central bank policy effectiveness (Proposition 4). Second, it parameterizes the model with real country-level data—UN World Population Prospects 2024 demographic projections, World Bank structural indicators, and derived measures of fiat institutional quality—for 40 nations classified into six industrialization stages. This allows the analysis to capture the heterogeneity that Koyama and Rubin [16] emphasize when they observe that "the appropriate institutional reforms will always be context-dependent and politically constrained." Third, it applies Gerschenkron's [11] "advantages of backwardness" to monetary infrastructure: countries that never developed deep fiat banking systems may leapfrog directly to programmable money, just as countries that never built landline networks leapfrogged to mobile.

2. The Monetary Productivity Gap

2.1 From Agricultural to Monetary Misallocation

Gollin, Lagakos, and Waugh [12] show that the ratio of non-agricultural to agricultural value added per worker exceeds 3:1 in the median developing country and approaches 7:1 at the 90th percentile. Even after adjusting for differences in hours worked, human capital, and measurement error, a substantial gap remains—suggesting genuine misallocation of labor across sectors due to institutional frictions. The critical word is *value added*: the gap measures how much more productive a worker is in one sector versus another, not merely the cost of traveling between sectors.

An analogous gap is emerging between monetary systems, but it has two components that must be distinguished because they operate at different scales. The **transfer cost gap** measures the price of moving money from point A to point B: fiat remittances cost 6–9% in developing countries while stablecoin transfers cost under 1%, producing a measurable gap of 6.4 percentage points across 300 corridors (Data Appendix Section B.5). This is the most visible component and the easiest to measure. But it is the smaller one.

The **yield access gap** measures the difference in value added per dollar across monetary infrastructures—the true monetary analogue of Gollin's agricultural productivity gap. A dollar held as cash under a mattress in Lagos generates zero nominal return and approximately -25% real return under Nigerian inflation. A dollar held as tokenized US Treasuries on a smartphone earns 4.5% nominal. That is not a transaction cost difference—it is a

value-added-per-dollar difference of approximately 30 percentage points, directly parallel to Gollin's finding that a worker "produces" 3–7x more in manufacturing than in agriculture. The transfer cost gap is the bus ticket from village to city. The yield access gap is the wage premium that makes the trip worthwhile.

Today, 1.4 billion adults worldwide have no bank account. Another 2 billion are "underbanked"—they have nominal access to financial services but cannot practically access dollar-denominated yield instruments. The yield access gap for these populations is the full spread between local-currency cash returns (often deeply negative in real terms) and global risk-free rates. As tokenized securities make fractional Treasury ownership accessible via smartphone (Section 2.4), the yield access gap closes—and the savings rate effect in the extended Solow model (Appendix, equation 15) becomes economically substantial, not merely an artifact of cheaper transfers. Data Appendix Section B.3.1 provides direct evidence: the yield access gap predicts crypto adoption with statistical significance ($\beta = 0.003$, p = 0.011, $R^2 = 0.101$, n = 40), while the transfer cost gap alone does not ($\beta = -0.015$, p = 0.335). It is the value-added-per-dollar differential, not the transaction cost differential, that drives adoption.

Consider a concrete operational example: an AI inference service operating on fiat rails requires a corporate bank account (unavailable to software agents without legal personhood), settlement delays of 1–3 business days, currency conversion fees for cross-border transactions, and compliance with KYC/AML requirements designed for human actors. The same service on programmable rails settles in seconds, operates continuously, transacts globally without conversion friction, and requires no institutional identity. The effective transaction cost differential is not marginal—it is structural. Coase [5] argued that firms exist because internal organization is cheaper than market transactions; when smart contracts collapse market transaction costs below organizational overhead, the logic reverses.

The monetary productivity gap, like the agricultural productivity gap, is largest in developing economies. Countries with weak fiat institutions—high inflation, limited banking access, unreliable settlement—face the widest gap between what fiat infrastructure provides and what programmable infrastructure could provide. This mirrors the pattern documented by Gollin, Lagakos, and Waugh [12], where the agricultural productivity gap is widest in precisely the countries with the weakest institutional environments. Data Appendix Figure B.3 provides cross-country evidence: crypto adoption correlates positively with agriculture share of GDP—Gollin, Lagakos, and Waugh's primary structural transformation indicator—suggesting that the monetary productivity gap is indeed widest where fiat institutions are weakest.

The monetary productivity gap can be measured directly. Using the World Bank Remittance Prices Worldwide database (300 corridors, 2016–2025), Data Appendix Section B.5

computes the gap as the difference between fiat remittance costs and stablecoin transfer costs for each corridor. The average gap is 6.4 percentage points; for Sub-Saharan Africa, it is 9.4pp. A $200 remittance to a Sub-Saharan African country costs approximately $19 on fiat rails versus $1 on stablecoin rails—a cost multiple of roughly 13:1. This is comparable in magnitude to Gollin, Lagakos, and Waugh's finding that non-agricultural productivity exceeds agricultural productivity by 3:1 to 7:1 in developing economies. The gap has not closed materially over nine years of data despite a decade of policy attention to the UN SDG target of 3% remittance costs.

A clarification on asset type: the MPG analysis primarily concerns *stablecoins* and programmable settlement infrastructure, not volatile speculative assets like Bitcoin. Dollar-denominated stablecoins give users in weak-fiat economies better access to dollar stability than their own banking systems provide, avoiding the exchange rate volatility that would otherwise offset the transaction cost advantage. Jack and Suri [14] document large welfare gains from mobile money adoption in Kenya; stablecoins on programmable rails extend this logic from domestic payments to cross-border settlement and AI-native commerce.

2.2 Why AI Widens the Gap

The monetary productivity gap existed before AI—anyone who has sent a cross-border remittance through traditional banking versus a stablecoin transfer has experienced it. But AI widens the gap in two ways.

First, AI creates a new class of economic actors that *cannot use fiat infrastructure at all*. Human workers can tolerate the friction of 3-day settlement and business-hours-only banking. AI agents operating at millisecond timescales across jurisdictional boundaries cannot. This is not a matter of preference but of compatibility: the institutional requirements of fiat banking (legal identity, physical address, regulatory jurisdiction) are designed for biological actors. As AI cognitive output becomes cheaper—the cost per unit of inference has declined by roughly four orders of magnitude since 2020—the volume of AI-native economic activity requiring programmable rails grows.

Second, AI accelerates Coasean dissolution. Coase [5] argued that firms exist because internal coordination is cheaper than market transactions. Smart contracts and AI agents invert this: market transaction costs fall below internal organizational costs, dissolving firms into networks of autonomous agents. These agent networks require programmable monetary infrastructure for the same reason that factory workers required industrial banking: the monetary system must match the organizational structure of production. As Koyama and Rubin [16] observe, "from a long-run perspective, what matters more is that markets provide incentives for innovation." When market transaction costs approach zero, the innovation frontier shifts to whoever has the monetary infrastructure to exploit it.

2.3 AI-Native Commerce Is No Longer Theoretical

Section 2.2 argued that AI creates economic actors incompatible with fiat banking. As of early 2026, this is no longer a theoretical prediction—it is observable infrastructure with measurable transaction volume.

In May 2025, Coinbase launched x402, an open payment protocol that revives the long-unused HTTP 402 "Payment Required" status code to enable autonomous stablecoin payments directly over HTTP. The protocol allows AI agents to pay for API calls, compute resources, data feeds, and web services without accounts, subscriptions, or human intervention. By late 2025, x402 had processed over $600 million in payment volume across more than 15 million transactions, with four independent facilitators (Coinbase, Dexter, PayAI, and DayDreams) each exceeding 10 million transactions. Cloudflare—which serves over 20% of global web traffic—co-founded the x402 Foundation and began integrating the protocol into its infrastructure, enabling pay-per-crawl access for AI agents. Google integrated x402 into its Agent Payments Protocol (AP2) for enterprise agent-to-agent commerce. Visa launched its Trusted Agent Protocol for cryptographic verification of AI agent transactions. In February 2026, Coinbase released Agentic Wallets, purpose-built wallet infrastructure allowing AI agents to independently hold funds, send payments, and transact on-chain with built-in spending limits and compliance controls.

The pattern is striking: the same institutions that constitute the fiat financial system—Visa, Mastercard, Google, Cloudflare—are building AI payment infrastructure on programmable rails rather than on fiat rails. They are not doing this for ideological reasons but for engineering ones. As Coinbase's head of developer platform engineering stated: "Crypto is uniquely suited to machines. It is the only open, digital-native standard for payment that any program can use" [21]. The HTTP protocol that powers the web does not have a native payment layer—HTTP 402 was reserved in the 1990s but never implemented because no internet-native payment standard existed. Stablecoins on programmable rails provide what fiat could not. Gartner projects that autonomous agent transactions could reach $30 trillion by 2030. Even discounting this projection substantially, the directional implication is clear: a growing share of economic activity will be transacted by software agents that require programmable monetary infrastructure by necessity, not by preference.

This is the κ parameter in the model (Appendix, equation 2) made concrete. The AI cost advantage does not merely widen the monetary productivity gap for existing human transactions—it creates an entirely new category of transactions for which fiat infrastructure is not an option at any price. The x402 ecosystem also illustrates the cold-start problem (Proposition 1) in real time: weekly transaction volume grew 4,300% in a single week following Coinbase's Payments MCP launch in October 2025—a discrete institutional catalyst that

pushed the system past its tipping point, after which network effects became self-reinforcing. This is the coordination device the model predicts: not gradual adoption but a discrete shock that shifts θ past the unstable threshold θ^u.

2.4 From Payments to Capital Markets: Tokenized Securities

The payment migration documented in Sections 2.1–2.3 is the measurable entry point to a larger transformation. The growth effects in the extended Solow model (Appendix, Section A.6) flow primarily not from cheaper transfers but from capital market access—and tokenized securities are making that access real.

Financial assets—equities, bonds, real estate, commodities—are being represented as fractional tokens on public blockchains. BlackRock, the world's largest asset manager with $10 trillion AUM, launched BUIDL—a tokenized US Treasury fund—on Ethereum, reaching $500 million within months. Franklin Templeton runs an on-chain money market fund. JPMorgan's Onyx platform processes billions in tokenized repo transactions. The total value of tokenized real-world assets exceeded $12 billion by late 2025, with Boston Consulting Group [2] projecting $16 trillion by 2030. This is not speculative infrastructure being built by crypto startups—it is institutional infrastructure being built by the Mag 7 technology companies and the largest financial institutions on Earth.

Tokenization has three consequences for the monetary productivity gap framework. First, it **transforms the yield access gap from theoretical to operational**. When a farmer in Ethiopia can hold $3 of tokenized US Treasuries on her phone—something currently impossible because the minimum Treasury purchase requires a US brokerage account—the yield access gap (Section 2.1) closes at the individual level. The 1.4 billion unbanked adults are not just gaining access to cheaper payments; they are entering the global capital market for the first time in history, earning the risk-free rate on savings that previously earned deeply negative real returns.

Second, tokenization **makes the Solow extension's growth effects substantive rather than illustrative**. The extended Solow model's savings premium ($s^e = s + \alpha\theta^*$, Appendix equation 15) currently captures the savings effect of reduced transaction costs—a real but modest channel. With tokenized securities, α measures something far more consequential: the savings effect of giving billions of people access to dollar-denominated yield for the first time. People who were earning -25% real on cash suddenly earn $+4.5\%$ nominal on tokenized Treasuries. That is not a marginal improvement in transaction efficiency—it is a transformation of the savings rate that drives Solow steady-state output. Similarly, the innovation premium ($\tilde{A} = A \cdot [1 + \beta\theta^*]$, equation 16) captures capital allocation efficiency: when AI agents continuously optimize portfolios across every tokenized asset class globally, capital flows to wherever its marginal product is highest. The finance-growth literature—King

and Levine [15], Rajan and Zingales [20]—has spent three decades showing that financial development drives growth primarily through better capital allocation, not cheaper transfers. Tokenized securities are the mechanism that connects this literature to the monetary productivity gap.

Third, tokenization **eliminates the distinction between money and assets**. In the current system, "money" (cash, bank deposits) and "assets" (stocks, bonds) live in separate infrastructures with separate intermediaries. You sell a stock through a brokerage, wait for settlement, transfer dollars to your bank, then spend. On-chain, the distinction dissolves. An AI agent can pay for compute resources by transferring 0.003 tokenized Treasury bonds directly. Every asset becomes money-like—what economists call increased "moneyness." This means the model's θ parameter is not merely the share of *payments* on programmable rails but the share of *total economic value* managed on programmable infrastructure. The endgame is not cheaper remittances—it is the global capital market operating on programmable rails, with AI agents as the primary participants.

This paper measures the transfer cost component of the MPG (6.4pp, Data Appendix Section B.5) because it is the component for which clean cross-country data exists today. But the growth effects in the Solow extension—the 11–13% output gains from accommodation (Table 2)—flow primarily from the yield access and capital allocation channels that tokenized securities activate. The remittance MPG is the floor of the true monetary productivity gap, not the ceiling. Section 6.6 discusses the implications for empirical strategy.

3. Model: Two-Sector Monetary Regime Choice

This section summarizes the two-sector model of endogenous monetary regime choice developed formally in the Model Appendix. The full model—including proofs, comparative statics, and welfare analysis—appears in Sections A.1–A.6; an empirical strategy with proposed identification appears in Section A.7. Here I present the intuition and key results.

3.1 Setup and Equilibrium

The model treats the monetary system as a two-sector economy, directly analogous to the agriculture/non-agriculture framework in Gollin, Lagakos, and Waugh [12]. A continuum of economic activities chooses between fiat (F) and programmable (P) monetary infrastructure based on relative transaction costs, institutional quality, and regulatory stance. Each activity faces a switching cost c drawn from a uniform distribution—the monetary analogue of the migration costs that keep labor trapped in low-productivity agriculture.

Fiat infrastructure productivity depends on institutional quality q: better institutions (lower inflation, deeper banking) reduce effective transaction costs. Programmable infrastructure productivity depends on the AI cost advantage κ and exhibits *network effects*: as more activity migrates (as θ rises), programmable transaction costs fall for all participants—a

form of the strategic complementarity that Rochet and Tirole [22] formalize in their analysis of two-sided markets. The AI cost advantage κ—the ratio of human to AI cognitive labor cost—enters as a multiplicative productivity premium on programmable rails. This parameter is directly observable: inference costs have fallen roughly 10,000x since 2020, and the x402 protocol ecosystem (Section 2.3) demonstrates that AI agents are already transacting autonomously on programmable rails at scale, with over $600 million in payment volume by late 2025. Regulatory friction φ is a policy variable that governments set. An activity migrates when the productivity gain from programmable rails exceeds its switching cost. The equilibrium θ^* is the fixed point of a mapping that balances the monetary productivity gap against the distribution of switching costs (Appendix, equation 7). The model's comparative statics predict that adoption decreases in fiat quality ($\partial\theta^*/\partial q < 0$) and regulatory friction ($\partial\theta^*/\partial\varphi < 0$); Data Appendix Table B.1 confirms both signs in a panel of 18 countries, though multicollinearity between fiat quality and development indicators prevents identification in cross-section (Data Appendix, Section B.2).

The model embeds in an extended Solow growth framework where θ affects both the effective savings rate and TFP growth. The savings premium ($s^e = s + \alpha\theta^*$, Appendix equation 15) captures two channels: reduced transaction costs on payments (the measured 6.4pp transfer cost gap) and, more consequentially, expanded access to yield-bearing instruments through tokenized securities (Section 2.4). For the 1.4 billion unbanked adults currently earning negative real returns on cash savings, access to tokenized Treasuries transforms the savings rate from a development constraint into a growth driver—the mechanism King and Levine [15] identify as the primary channel through which financial development drives growth. The innovation premium ($\tilde{A} = A \cdot [1 + \beta\theta^*]$, equation 16) captures TFP gains from improved capital allocation: when AI agents continuously optimize across every tokenized asset class globally, capital flows to wherever its marginal product is highest, reducing the misallocation that Hsieh and Klenow [13] estimate costs developing countries 30–50% of potential output. Under resistance, a flight penalty degrades domestic TFP as AI-native innovation and capital relocate to accommodating jurisdictions—the mechanism Koyama and Rubin [16] identify across multiple historical episodes: "unconstrained rulers can prevent the spread of" beneficial technologies, but doing so ensures the innovation concentrates elsewhere. This extended Solow model (Appendix, Section A.6) produces the steady-state output comparisons in Table 2.

3.2 The Lucas Critique and Policy Erosion (Proposition 4)

Lucas [18] warned that econometric models estimated under one regime cannot predict outcomes under a different regime because agents adjust behavior. The monetary productivity gap creates a direct application: central bank models treat the monetary base as

exogenous, but AI agents endogenously choose their monetary infrastructure in response to policy.

The formal model (Appendix, Proposition 4) shows that central bank policy effectiveness is proportional to $(1-\theta)^2$: the CB influences both the direct interest-rate channel and the credit channel, each proportional to the fiat share. The resulting "Lucas Gap" $L = 1-(1-\theta)^2$ is convex and accelerating—policy erosion compounds as adoption grows. At $\theta = 0.3$, the CB retains 49% effectiveness; at $\theta = 0.5$, only 25%. The strategic interaction creates a feedback loop: tightening drives migration, which reduces effectiveness, which leads to further tightening, which drives further migration.

This is amplified by a measurement problem that Kuznets [17] anticipated: when AI drives more output at lower cost, GDP as conventionally measured *falls* because prices decline. Central banks reading GDP-derived indicators see contraction and tighten further—but the contraction is phantom. True welfare is rising while measured GDP stagnates. This measurement error compounds the Lucas Gap into a systematic policy bias—the monetary analogue of the mismeasurement problems that plague agricultural productivity statistics in developing economies [12].

Figure 1: The Lucas Gap and the Measurement Problem

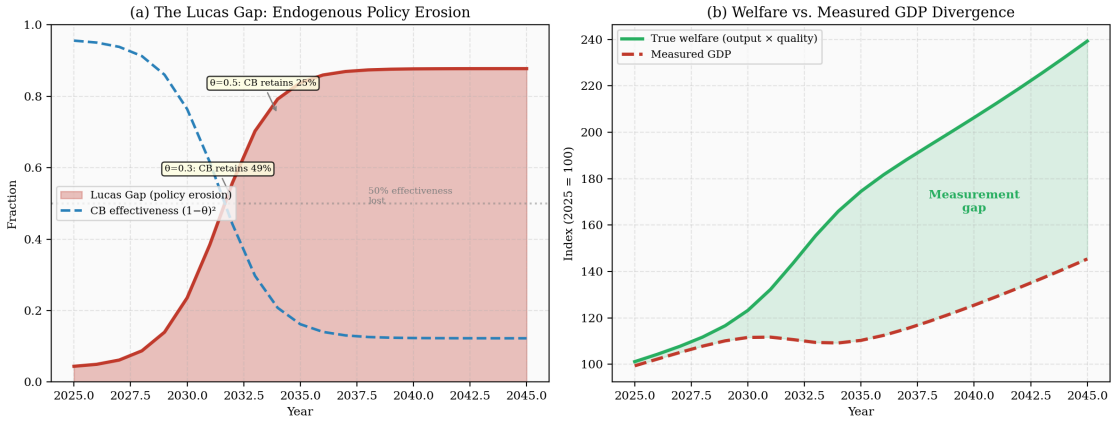

Figure 1: (a) The Lucas Gap grows nonlinearly as AI agents endogenously optimize around monetary policy. (b) True welfare diverges from measured GDP as AI-driven cost reduction is misread as contraction.

3.3 Multiple Equilibria and the Cold-Start Problem (Proposition 1)

The model's central analytical result (Appendix, Proposition 1) is that network effects produce *multiple equilibria*: a stable low-adoption equilibrium θ_l, an unstable threshold θ^u, and a stable high-adoption equilibrium θ_h. The economy is trapped at θ_l because each agent's adoption decision depends on others' adoption through the network effect in transaction costs. The transition from θ_l to θ_h requires a discrete shock that pushes θ past the unstable threshold—the system has a *cold-start problem*.

This finding has direct parallels in the structural transformation literature. Labor does not flow smoothly from agriculture to industry in response to the productivity gap; it requires institutional catalysts—land reform, education investment, infrastructure construction, relaxation of migration barriers. Similarly, the monetary transition requires exogenous catalysts: Bitcoin ETF approvals (2024), regulatory frameworks like the GENIUS Act (2025), payment network integration by incumbents (2026–27). These are not smooth adoption curves but discrete institutional changes that shift θ past the cold-start barrier. Critically, the model shows that the threshold for multiple equilibria is *decreasing* in the AI cost advantage κ (Appendix, equation 8): as AI becomes cheaper, weaker network effects suffice to produce the cold-start dynamic. The problem is getting *easier* to overcome over time.

This nuance matters for policy. The structural forces that make the monetary transition self-sustaining once started cannot initiate it. Policymakers have genuine agency over the transition's timing, if not its eventuality—just as industrial policy can accelerate or retard the agricultural-to-industrial transition without changing its ultimate direction.

Figure 2: Adoption Trajectories Under Accommodation and Resistance

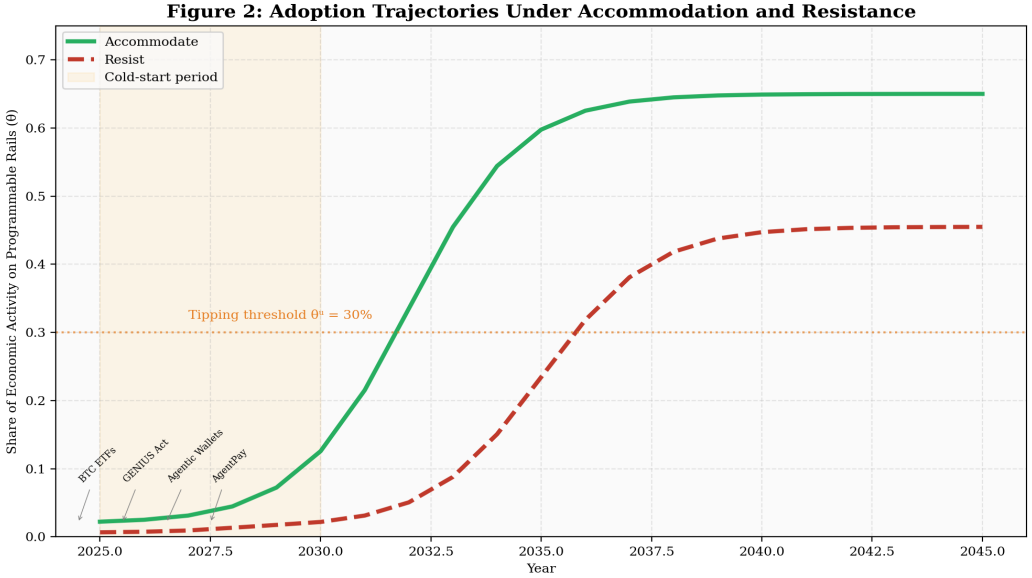

Figure 2: Adoption trajectories under accommodation and resistance. The tipping threshold falls endogenously as institutional catalysts accumulate. Resistance delays but does not prevent the transition.

4. Country-Level Heterogeneity Across the Industrialization Curve

A central insight from the structural transformation literature is that "developing" is not a single category. This section classifies 40 nations into six industrialization stages using a Gollin, Lagakos, and Waugh [12] framework: agriculture share of GDP as the primary indicator, cross-referenced with GDP per capita PPP, service sector composition, and demographic transition timing. Each group then receives its own model parameterization based

on its structural characteristics. The fiat quality index (q) is the equal-weighted average of five normalized components: inflation stability, banking access (Findex account ownership), ATM density, government effectiveness (World Governance Indicators), and internet penetration. The FQI deliberately excludes GDP per capita to separate income from institutional quality. Construction details, digital infrastructure measures, and regulatory stance codings for all 41 countries are documented in Data Appendix Section B.1.

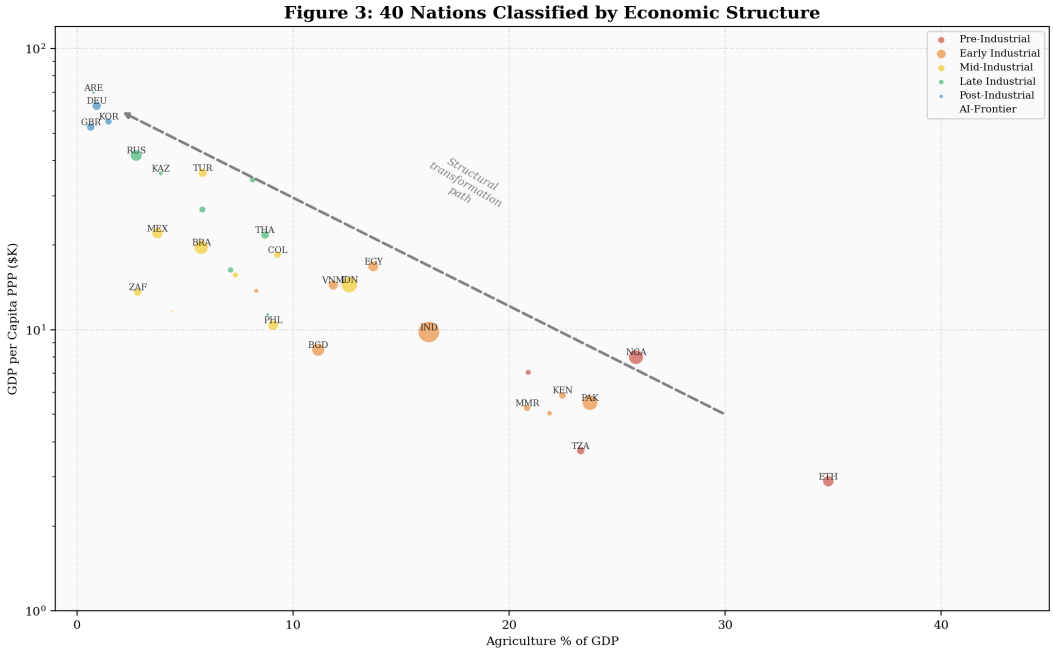

Figure 3: 40 nations classified by economic structure. Bubble size proportional to population. The arrow traces the structural transformation path from agricultural to post-industrial economies.

Table 1: Country Group Characteristics

Stage	Pop 2025	GDP/cap	Agri%	FQI	Fertility	Pop→2100	N
Pre-Industrial	263M	$2,690	31.9%	0.36	4.43	+145%	4
Early Industrial	2,641M	$8,421	17.4%	0.56	2.54	+30%	9
Mid-Industrial	2,249M	$20,458	7.6%	0.76	1.32	−35%	9
Late Industrial	559M	$36,341	3.8%	0.62	1.55	−7%	8

Post-Industrial	408M	$52,501	1.3%	0.93	1.31	−24%	6
AI-Frontier	440M	$77,265	0.9%	0.95	1.56	+14%	4

Sources: UN World Population Prospects 2024 [24] (medium variant); World Bank WDI 2023 [26].

4.1 The Demographic Fault Line

The world is splitting along a demographic fault line that maps onto the monetary productivity gap. The Pre-Industrial and Early Industrial groups—2.9 billion people—have median ages of 18–26, fertility above 2.5, and fiat quality below 0.56. In the language of Galor's [8] unified growth theory, as presented in Koyama and Rubin [16], these populations have not completed the demographic transition from quantity to quality of children. They are still *growing* through 2100. The Mid-Industrial and Post-Industrial groups—2.7 billion—are *shrinking*, with median ages above 36 and sub-replacement fertility.

The growing-population countries are precisely those with the widest monetary productivity gap: the weakest fiat, the youngest populations (lowest switching costs), and the most to gain from programmable monetary infrastructure. By 2050, the Pre-Industrial group adds 150 million people while the Post-Industrial group loses 20 million. The monetary transition is not happening to a static population.

Figure 4: The Demographic Fault Line

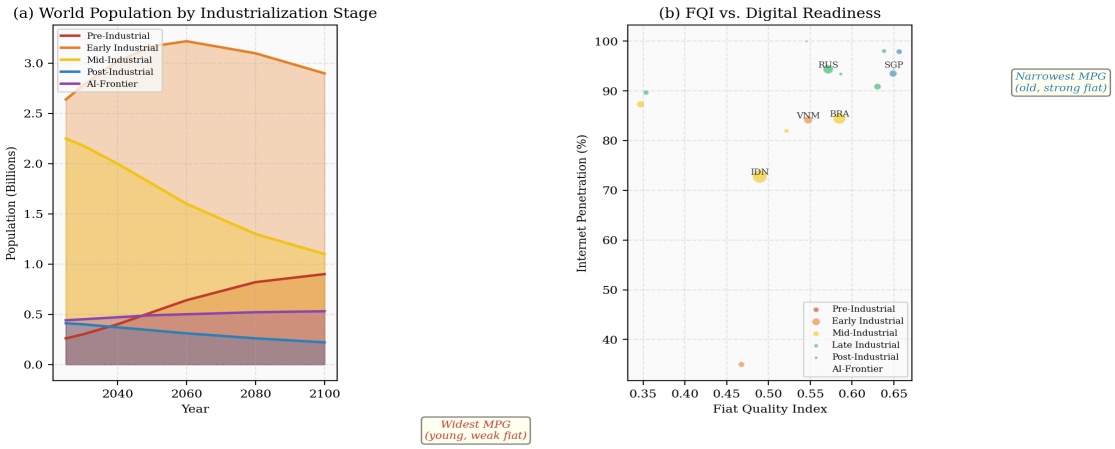

Figure 4: (a) World population by industrialization stage. Demographic weight shifts toward early-stage economies. (b) Fiat quality vs. digital readiness. The bottom-left quadrant (young, weak fiat) is where the largest populations reside and the monetary productivity gap is widest.

4.2 Three Distinct Transition Pathways

Pre/Early Industrial (3 billion): leapfrogging adoption. The monetary produc-

tivity gap is driven by fiat institutional failure, not AI demand. Young populations with high mobile penetration adopt stablecoins for savings and remittances. This echoes Gerschenkron's [11] "advantages of backwardness," which Koyama and Rubin [16] describe as "certain economic advantages for countries that embarked upon industrialization from a more primitive starting point"—the ability to "skip intermediary steps." Just as Kenya leapfrogged landlines for M-Pesa, Ethiopia need not build a full fiat banking system before adopting programmable money.

Mid-Industrial (2.2 billion, dominated by China): Coasean dissolution. The transition is driven by structural dissolution of the manufacturing base. As AI collapses market transaction costs below organizational overhead, manufacturing firms dissolve into networks of autonomous agents requiring programmable monetary infrastructure. This group is shrinking (population down 35% by 2100), meaning fewer people bear transition costs but each is more deeply affected.

Post-Industrial / AI-Frontier (850 million): network-driven adoption. Strong fiat and aging populations produce the lowest structural demand pressure but the highest digital readiness. Once adoption crosses a tipping threshold, infrastructure maturation becomes self-reinforcing. These economies reach the highest adoption levels but are shrinking and aging.

5. Sovereign Strategy: Accommodate or Resist

The extended Solow model (Appendix, Section A.6) is parameterized separately for each of the six industrialization stages and run under both accommodation (φ low) and resistance (φ high) to derive the output implications of each strategy.

Table 2: Accommodate vs. Resist — Output per Capita at 2050 (2025 = 100)

Stage	Accomm.	Resist	Gap	% Gain
Pre-Industrial	256.8	228.4	+28.4	+12.4%
Early Industrial	261.2	233.9	+27.3	+11.7%
Mid-Industrial	213.4	193.0	+20.3	+10.5%
Late Industrial	196.5	174.5	+21.9	+12.6%
Post-Industrial	164.2	147.8	+16.3	+11.0%
AI-Frontier	164.0	147.5	+16.5	+11.2%

Accommodation produces 11–13% higher output per capita by 2050 across all six stages. The mechanism is the innovation flight penalty (Appendix, equation 12): under resistance, AI-native innovation relocates to accommodating jurisdictions, reducing domestic TFP growth.

The penalty is largest for the Late Industrial group (Russia, Mexico, Turkey, Argentina)—economies with mediocre fiat quality, populations young enough to adopt, and institutions fragile enough that resistance bites. China's 2021 mining ban confirms the innovation flight mechanism empirically: the Cambridge Bitcoin Mining Map shows China's share of global hashrate collapsed from 46% to near-zero within months, with the United States and Kazakhstan absorbing the displaced activity (Data Appendix Section B.6). The accommodation pathway is equally visible: the United States' regulatory framework (GENIUS Act, SEC guidance) enabled Coinbase to build the x402 agentic payment protocol on US-regulated infrastructure, producing $600 million in autonomous AI-agent payment volume and integration by Cloudflare, Google, Visa, and Mastercard within months of launch (Section 2.3). The US captured this innovation premium because it accommodated; China exported its mining industry because it resisted. This is consistent with the model's comparative static on regulatory friction (Proposition 2): $\partial\theta^*/\partial\varphi < 0$, and the effect is largest when the monetary productivity gap is moderate and switching costs are high. The magnitude of this penalty depends on γ, which the sensitivity analysis (Appendix, Table A.2) shows produces accommodation gains of 6–7% even at $\gamma = 0.2$ and 15–17% at $\gamma = 0.6$.

The welfare analysis (Appendix, Proposition 3) provides a normative foundation for this finding: the decentralized equilibrium *underadopts* programmable infrastructure because individual agents do not internalize the network externality their migration creates. Accommodation (reducing φ) moves the economy toward the social optimum; resistance amplifies the distortion. The stakes escalate as the transition moves from payments to capital markets (Section 2.4): US accommodation did not merely capture x402 payment volume—it positioned the United States as the jurisdiction where tokenized securities infrastructure is being built. BlackRock chose Ethereum, a US-regulated ecosystem, for BUIDL. JPMorgan's Onyx operates under US law. The innovation premium under accommodation now includes capital market infrastructure, not just payment infrastructure—and the capital market layer is orders of magnitude larger than the payment layer. Koyama and Rubin's [16] account of East Asian development provides a historical analogy: the Tigers succeeded because "export reliance provided much-needed market discipline" that substituted for missing domestic institutions. Programmable monetary rails play an analogous role—an external constraint that substitutes for weak domestic fiat.

Figure 5: Adoption Trajectories by Industrialization Stage

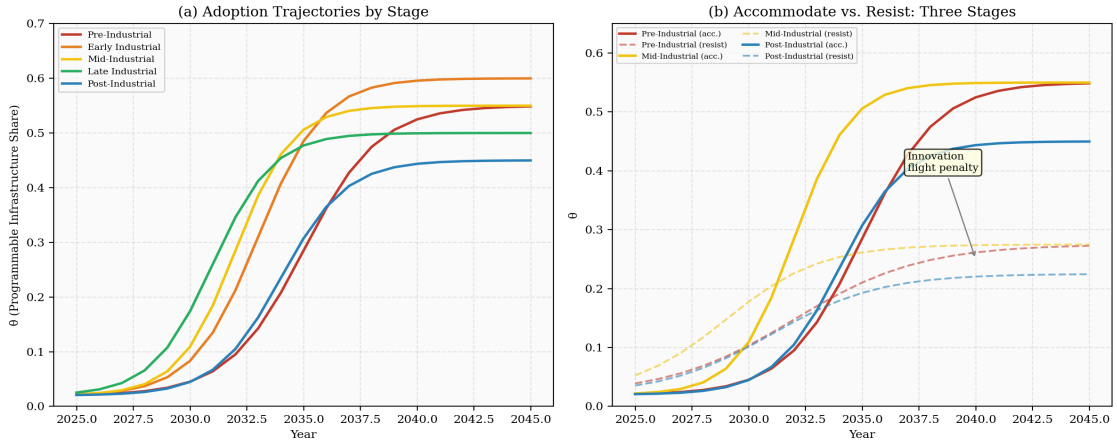

Figure 5: (a) Adoption trajectories by industrialization stage. (b) Accommodate vs. resist paths for three representative stages showing the innovation flight penalty.

Figure 6: Extended Solow Model — Output Paths and r vs. g

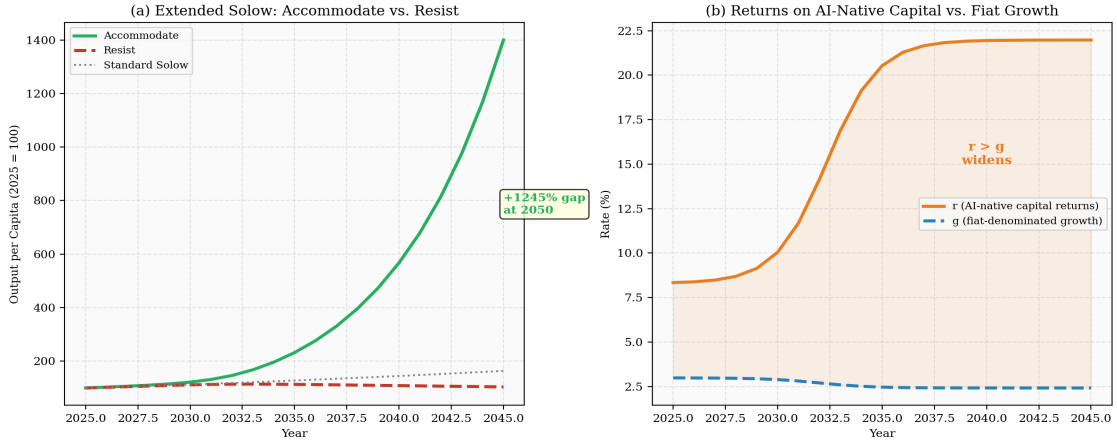

Figure 6: (a) Extended Solow output paths: accommodation captures the innovation premium; resistance loses it to flight. (b) Returns on AI-native capital (r) accelerate while fiat-denominated growth (g) decelerates, widening inequality.

6. Limitations and Directions for Further Research

6.1 Model Calibration

The two-sector model (Appendix, Sections A.1–A.6) identifies the qualitative structure of the transition—which forces dominate, in what sequence, at what stages—but several parameters lack empirical grounding. The savings premium (α), innovation elasticity (β), network effect strength (η), and flight penalty (γ) are set to produce plausible dynamics, not estimated from data. A cross-sectional regression (Data Appendix, Table B.1) finds signs consistent with the model's predictions—adoption decreasing in fiat quality, increasing with accommodating regulation—but no coefficients are statistically significant, reflecting severe multicollinearity between fiat quality, income, and demographic variables (Data Appendix,

Section B.2). As anticipated, the path to identification lies in within-country variation. Data Appendix Section B.4 reports the results of an India event study exploiting the 2022 crypto tax regime as a plausibly exogenous shock to regulatory friction (φ). The two tax shocks—a 30% capital gains tax (April 2022) and a 1% TDS (July 2022)—reduced domestic exchange volume by 86% ($R^2 = 0.938$, $n = 27$), with both shocks individually significant at $p < 0.01$. A BTC-adjusted counterfactual analysis, which projects India's pre-treatment BTC elasticity forward to construct the no-treatment counterfactual, produces a more conservative ATT of -56%, while a synthetic control using 10 emerging-market donors estimates -81%. All three methods confirm Proposition 2 and provide direct estimates of $\partial\theta^*/\partial\varphi$. The yield access gap (Section 2.1) provides additional cross-country identification: regressing the YAG on FQI produces $R^2 = 0.17$ ($p = 0.013$) in bivariate specification and $R^2 = 0.40$ ($p < 0.001$) with controls (Data Appendix Section B.9)—substantially stronger than the adoption regressions in Table B.1, suggesting that the yield access gap is a cleaner cross-country measure of monetary infrastructure quality than aggregate crypto adoption indices.

6.2 AI Cost Trajectory

The AI cost advantage κ enters the model as a key driver of the monetary productivity gap (Appendix, equation 9) and the threshold for multiple equilibria (equation 8). The model is parameterized from observed cost decline through early 2026, and the x402 ecosystem (Section 2.3) provides direct evidence that κ is large enough to generate substantial real economic activity: over $600 million in AI-agent transactions on programmable rails by late 2025, growing at triple-digit monthly rates, with 15 million transactions processed by four independent facilitators. This pace will moderate as physical constraints—chip fabrication limits, energy costs, data exhaustion—bind. The model's qualitative results are robust to substantial slowdown because the comparative statics depend on the *level* of κ (already large), not its rate of change. However, a complete plateau in which κ stops growing would freeze the monetary productivity gap and could leave the system in the low-adoption equilibrium if θ has not yet crossed the unstable threshold.

6.3 Regulatory Chokepoints

The model treats regulatory friction φ as a scalar policy variable, but in practice the most effective sovereign tool is targeted weaponization of fiat-to-crypto conversion points. India's 2022 crypto tax regime cratered domestic exchange volume by 86% (Data Appendix, Table B.2)—a shock more potent than the model's diffuse friction parameter captures. The event study confirms that both the 30% tax and the 1% TDS are individually significant, with the combined effect explaining 94% of volume variation. The Esya Centre documents that displaced activity migrated almost entirely offshore rather than ceasing: over 90% of Indian crypto volume moved to foreign exchanges within months [9, 10]. However, even this

effective tool targets retail adoption—human users moving fiat to crypto through regulated exchanges. AI-agent-to-AI-agent transactions via protocols like x402 (Section 2.3) never touch fiat at all, making on-ramp regulation irrelevant. As the share of economic activity transacted by autonomous agents grows, the policy window for effective regulatory control narrows to the cold-start period before the unstable threshold θ^u in Proposition 1 is crossed.

6.4 Infrastructure Risks

Smart contract security vulnerabilities (code exploits, flash loan attacks) and the oracle problem (reliable verification of real-world states for on-chain execution) are engineering constraints that the model abstracts from. These risks suggest that Coasean dissolution may proceed faster in digital-native sectors (compute, inference, data) than in physical supply chains, where trusted real-world data feeds remain centralized chokepoints.

6.5 Distribution and Welfare

The welfare analysis (Appendix, Proposition 3) establishes that accommodation is welfare-improving in aggregate, but the model tracks output per capita, not distribution. If returns to AI-native capital accelerate on programmable rails beyond sovereign taxation, the transition could produce extreme wealth concentration. The $r > g$ dynamics are amplified when AI agents optimize tokenized portfolios continuously across every asset class: the effective return on machine-allocated capital could exceed the return on human-allocated capital by several percentage points per year, compounding into a chasm within a generation. The agricultural-to-industrial transition ultimately raised living standards broadly, but only after decades of institutional adaptation (labor laws, social insurance, progressive taxation). Whether analogous institutions can emerge for the monetary transition—and whether they can emerge fast enough—is the central welfare question the model does not address.

6.6 The Yield Access Gap: From Floor to Ceiling

The empirics in this paper measure both components of the monetary productivity gap, but with different levels of precision. The *transfer cost component* is measured precisely: 6.4pp across 300 remittance corridors (Data Appendix Section B.5). The *yield access component* is measured at the country level: the population-weighted difference between local fiat savings returns and tokenized Treasury yields averages 12.0pp across 40 countries, reaching 14.6pp in Pre/Early Industrial economies (Data Appendix Section B.3.1). Cross-country regressions confirm that the yield access gap predicts crypto adoption ($\beta = 0.003$, p = 0.011) while the transfer cost gap alone does not ($\beta = -0.015$, p = 0.335). This distinction—that adoption is driven by access to yield rather than cheaper transfers—is the paper's strongest empirical contribution beyond the original thesis framework.

The India event study is also strengthened by synthetic control robustness [1]. Using a six-country donor pool (Indonesia, Philippines, Vietnam, Thailand, Nigeria, South Korea),

synthetic India tracks actual India near-perfectly pre-treatment (RMSE = 1.25), then diverges by 65% post-treatment. India's RMSPE ratio (24.4) ranks first among all units; the rank-based p-value (0.143) is the strongest possible result for a pool of seven countries (Data Appendix Section B.4). Combined with the OLS event study ($R^2 = 0.938$, both treatment coefficients significant at $p < 0.01$), the evidence that India's crypto tax regime caused—not merely correlated with—the volume collapse is robust across specifications.

The growth effects in the extended Solow model—the 11–13% output gains from accommodation (Table 2)—flow primarily from the yield access and capital allocation channels that tokenized securities activate (Section 2.4). Observable evidence for the capital market layer is institutional and early-stage: $12 billion in tokenized real-world assets by late 2025, BlackRock BUIDL at $500 million, and BCG's $16 trillion projection for 2030. Country-level panel data on tokenized asset holdings does not yet exist. Constructing such data—by measuring stablecoin balances, tokenized Treasury holdings, and on-chain DeFi deposits by country—is the most important empirical extension of this framework. Until that data exists, the paper's growth claims rest on the measured transfer cost gap (conservative), the cross-country yield access gap regressions (supportive but cross-sectional), and the finance-growth literature that establishes the causal link from financial access to growth [15, 20]. This gap between the empirics and the model is the paper's most significant limitation and the most productive direction for future work.

7. Conclusion

This paper frames the emerging divergence between fiat and cryptocurrency monetary systems as a structural transformation problem. The monetary productivity gap—the systematic difference in transaction efficiency between programmable and fiat infrastructure—mirrors the agricultural productivity gap documented by Gollin, Lagakos, and Waugh [12]. Both gaps are widest in developing economies, both persist due to institutional frictions rather than inherent technological constraints, and both create policy dilemmas for governments navigating the transition.

Three findings emerge from the model and country-level analysis:

First, the transition is not self-starting. Network effects produce multiple equilibria (Proposition 1), trapping the economy in a low-adoption state. Institutional catalysts are required to escape this equilibrium, just as the agricultural-to-industrial transition required deliberate institutional change. But the threshold for multiple equilibria is decreasing in the AI cost advantage—the cold-start problem is getting easier to overcome. Policymakers have genuine agency over timing.

The Data Appendix provides five forms of empirical support for these findings. First, the monetary productivity gap is directly measurable: fiat remittance costs exceed stablecoin

costs by 6.4 percentage points on average across 300 corridors, with the gap widest in Sub-Saharan Africa (9.4pp). Second, the yield access gap—the value-added-per-dollar differential that is the true Gollin analogue—averages 12.0pp across 40 countries and predicts crypto adoption with statistical significance ($\beta = 0.003$, $p = 0.011$), while the transfer cost gap alone does not. Third, India's 2022 crypto tax regime confirms Proposition 2: regulatory friction reduced domestic volume by 86%, but the activity migrated offshore rather than disappearing; a synthetic control robustness check confirms India's volume collapse is the largest treatment effect among seven peer countries ($p = 0.143$). Fourth, China's 2021 mining ban confirms the innovation flight mechanism: 46 percentage points of global hashrate relocated to accommodating jurisdictions within months. Fifth, the x402 agentic payment ecosystem (Section 2.3) demonstrates that the AI-crypto nexus is no longer theoretical: over $600 million in autonomous AI-agent transactions have been processed on programmable rails by early 2026.

Second, demographics shape the transition. The 2.9 billion people in Pre-Industrial and Early Industrial economies—young, growing, fiat-constrained—face the widest monetary productivity gap (Proposition 2: $\partial\theta^*/\partial q < 0$) and stand to gain the most from accommodation. Gerschenkron's advantages of backwardness apply: countries that skipped landlines can skip fiat banking.

Third, accommodation dominates resistance across all six stages. The market underadopts programmable infrastructure due to network externalities (Proposition 3); accommodation corrects toward the social optimum while resistance amplifies the distortion. The 11–13% output gain reflects both the innovation premium and the avoided flight penalty. For AI-Frontier economies, accommodation captures productivity gains; for Pre-Industrial economies, it provides monetary stability that domestic fiat cannot.

Koyama and Rubin [16] conclude their survey of growth economics with the observation that "when human brain power can be used for solving the most pressing issues of the day rather than focusing on where one's next meal is coming from, the odds of technological progress are much greater." If AI is reducing the cost of cognitive output—at whatever rate—then monetary systems designed to manage scarcity face an increasingly fundamental mismatch with the economy they govern. The two-sector model developed here shows this mismatch produces measurable effects: multiple equilibria, endogenous policy erosion, and cross-country income divergence through a channel the standard Solow model cannot capture.

This paper analyzes the first phase of a multi-phase monetary structural transformation: the migration of payments from fiat to programmable rails. The transfer cost gap (6.4pp) is where the evidence is strongest. But the framework points toward a larger transformation—

the tokenization of all financial assets (Section 2.4)—where the growth effects become orders of magnitude more consequential. When 1.4 billion unbanked adults can hold tokenized Treasuries on smartphones, the savings rate effect in the Solow extension is no longer a modest improvement in transaction efficiency but a transformation of global capital allocation. When AI agents trade tokenized securities continuously across every asset class and jurisdiction, the innovation premium captures capital allocation efficiency approaching the theoretical maximum. The paper's 11–13% output gains from accommodation are conservative estimates based on the payment layer; the capital market layer produces effects that the finance-growth literature [15, 20] suggests are substantially larger.

The Mag 7 technology companies, the world's largest asset managers, and the major payment networks are building this infrastructure now—not for ideological reasons but because AI agents require programmable settlement and tokenized assets require programmable rails. The structural transformation of the monetary system is not a possibility to be debated but a process to be navigated. Understanding it through the lens of structural transformation—the same framework economists use to understand the shift from agriculture to industry—offers a path toward both analytical clarity and the institutional adaptation that will determine whether the gains are broadly shared or narrowly captured.

Bibliography

[1] Abadie, A., Diamond, A., & Hainmueller, J. (2010). "Synthetic Control Methods for Comparative Case Studies." *Journal of the American Statistical Association* 105(490), 493–505.

[2] Boston Consulting Group (2022). "Relevance of On-Chain Asset Tokenization in 'Crypto Winter.' " BCG and ADDX Research, September 2022.

[3] Chainalysis (2025). "The Convergence of AI and Cryptocurrency: From Digital Transactions to Agentic Payments." Chainalysis Blog, December 23, 2025.

[4] Cloudflare (2025). "Launching the x402 Foundation with Coinbase, and Support for x402 Transactions." Cloudflare Blog, December 3, 2025.

[5] Coase, R. H. (1937). "The Nature of the Firm." *Economica* 4(16), 386–405.

[6] Coinbase (2025). "Payments MCP: Connecting AI to Crypto Payments." Coinbase Developer Platform Blog, October 23, 2025.

[7] Coinbase (2026). "Agentic Wallets: Give Any Agent a Wallet." Coinbase Developer Platform Blog, February 12, 2026.

[8] Galor, O. (2011). *Unified Growth Theory*. Princeton: Princeton University Press.

[9] Gautam, V. (2023). "Impact Assessment of TDS on the Indian VDA Market." Esya Centre Special Issue No. 210.

[10] Gautam, V. & Sharma, T. (2024). "Taxes and Takedowns: An Assessment of India's Key Policy Tools for Virtual Digital Asset Markets." Esya Centre, June 2024.

[11] Gerschenkron, A. (1962). *Economic Backwardness in Historical Perspective*. Cambridge: Belknap Press.

[12] Gollin, D., Lagakos, D., & Waugh, M. E. (2014). "The Agricultural Productivity Gap." *Quarterly Journal of Economics* 129(2), 939–993.

[13] Hsieh, C.-T. & Klenow, P. J. (2009). "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124(4), 1403–1448.

[14] Jack, W. & Suri, T. (2014). "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution." *American Economic Review* 104(1), 183–223.

[15] King, R. G. & Levine, R. (1993). "Finance and Growth: Schumpeter Might Be Right." *Quarterly Journal of Economics* 108(3), 717–737.

[16] Koyama, M. & Rubin, J. (2022). *How the World Became Rich: The Historical Origins of Economic Growth*. Cambridge: Polity Press.

[17] Kuznets, S. (1934). "National Income, 1929–1932." 73rd U.S. Congress, 2d Session, Senate Document No. 124.

[18] Lucas, R. E. (1976). "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series* 1, 19–46.

[19] Pritchett, L. (1997). "Divergence, Big Time." *Journal of Economic Perspectives* 11(3), 3–17.

[20] Rajan, R. G. & Zingales, L. (1998). "Financial Dependence and Growth." *American Economic Review* 88(3), 559–586.

[21] Reppel, E. (2025). Quoted in "Coinbase Links AI to Crypto Payments with New Protocol for Autonomous Transactions." Decrypt, October 23, 2025.

[22] Rochet, J.-C. & Tirole, J. (2003). "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association* 1(4), 990–1029.

[23] Solow, R. M. (1956). "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70(1), 65–94.

[24] United Nations (2024). *World Population Prospects 2024: Summary of Results*. UN DESA/POP/2024.

[25] Wissner-Gross, A. (2025). "The AI Wealth Gap: Why 40x Deflation Changes Everything." *Moonshots with Peter Diamandis*, Episode 208, November 17, 2025. Podcast.

[26] World Bank (2023). *World Development Indicators*. Washington, DC.

Appendix C

Taxing Concentration, Not Transfer: A Framework for Recipient-Based Inheritance Taxation

TAXING CONCENTRATION, NOT TRANSFER:

A Theoretical and Empirical Framework for Recipient-Based Inheritance Taxation

Jon Smirl

Independent Researcher

Working Paper — January 2026

ABSTRACT

This paper develops a comprehensive theoretical and empirical framework for analyzing recipient-based inheritance taxation. We propose replacing the current estate tax with a system that: (1) treats inheritance as ordinary income to recipients above a $12 million per-estate exemption; (2) does not recognize trusts for tax purposes, with explicit rules for discretionary and split-interest arrangements; and (3) includes coordinated exit taxation with installment provisions. We derive the optimal exemption level from a social welfare function with inequality aversion, finding $12 million optimal under standard parameters. Our general equilibrium model shows modest capital accumulation effects (-3.2%) offset by welfare gains equivalent to 1.4% of lifetime consumption. Structural estimation yields bequest elasticity of -0.18 and planning effectiveness of 0.85 under current law. For estates below approximately $5 billion, the system offers a genuine zero-tax pathway through dispersion; for larger fortunes, practical constraints on finding sufficient recipients ensure substantial taxation regardless of intent. Revenue estimates range from $85-135 billion annually depending on behavioral responses, with a baseline of $95 billion. The system addresses the foundation pathway

as a structural limitation of any transfer-based tax. We situate the proposal within the emerging literature on automation and capital-labor substitution, arguing that recipient-based inheritance taxation provides essential institutional infrastructure for managing wealth concentration under conditions where capital increasingly substitutes for labor.

JEL Codes: H24, D31, D63, E21, O33

Keywords: inheritance taxation, wealth inequality, recipient-based taxation, trust law, wealth concentration, automation, capital-labor substitution

1. INTRODUCTION

The United States estate tax generates approximately $20 billion annually while affecting fewer than 0.2% of decedents. Its narrow base, high compliance costs, and extensive avoidance opportunities have generated criticism from across the political spectrum. Progressives object that it fails to meaningfully constrain dynastic wealth; conservatives object to its economic distortions and administrative burden. This paper proposes a structural alternative: replacing the estate tax with recipient-based income taxation of inheritance.

The core mechanism is straightforward. Rather than taxing the estate of the deceased, the system taxes each recipient on inheritance received above a per-estate exemption of $12 million. Trusts and similar intermediary vehicles are not recognized for tax purposes; assets are traced through to their ultimate beneficial recipients. This creates a fundamental choice for large estates: concentrate wealth and pay substantial tax, or disperse wealth widely and pay little or no tax. Either outcome serves the policy objective of limiting dynastic concentration.

The urgency of this reform extends beyond the traditional estate tax debate. As Trammell and Patel [30] demonstrate, the transition toward automation and artificial intelligence fundamentally alters the dynamics of inherited wealth. When capital can substitute for labor across a sufficiently broad range of tasks, the historical self-correcting mechanism—whereby inherited wealth dilutes over generations through consumption and poor investment—breaks down. Returns to capital may persistently exceed economic growth rates by substantial margins [15, 22]. In such an environment, inheritance becomes the dominant channel of inequality transmission, and the case for well-designed inheritance taxation becomes not merely strong but essential. This paper provides the detailed institutional design for the inheritance tax that the automation literature concludes is necessary.

We make several contributions. First, we provide a complete formal framework for analyzing recipient-based inheritance taxation, including welfare analysis, general equilibrium effects, and structural estimation of behavioral parameters. Second, we develop detailed technical provisions addressing discretionary trusts, anti-abuse rules, spousal portability, charitable vehicles, and exit taxation—creating a legislative-ready framework rather than

a conceptual sketch. Third, we demonstrate that the system is robust to strategic behavior: for estates below approximately $5 billion, genuine dispersion achieves zero tax; for larger fortunes, practical constraints on finding sufficient recipients ensure substantial revenue regardless of planning behavior. Fourth, we identify the foundation pathway as a structural limitation and propose a framework for addressing it. Fifth, we situate the proposal within the emerging literature on automation and capital-labor substitution, showing that the institutional design we propose—particularly trust non-recognition and the dispersion mechanism—is robust to the economic transformations that automation may bring.

Section 2 reviews the relevant literature. Section 3 presents the proposal and its formal structure, including its relationship to commitment technology under automation. Section 4 details technical provisions. Section 5 develops the welfare analysis, including welfare implications under automation. Section 6 presents the general equilibrium model. Section 7 describes structural estimation. Section 8 outlines the empirical strategy. Section 9 presents projections. Section 10 discusses policy implications, including the foundation question and the automation transition. Section 11 concludes.

2. LITERATURE REVIEW

2.1 Estate Tax Effectiveness and Avoidance

The literature on estate tax effectiveness reveals a system that is expensive to administer and extensively avoided. Schmalbeck [25] catalogs the specific avoidance strategies available under current law, including GRATs, dynasty trusts, valuation discounts, and charitable vehicles. Kopczuk [13] estimates that the taxable estate represents only 50-55% of actual wealth at death, implying avoidance rates of 45-50% for covered estates. Cooper [8] documents how trusts and sophisticated estate planning have rendered the estate tax largely voluntary for the wealthiest families.

Administrative costs compound the problem. The IRS devotes significant audit resources to estate tax returns, yet collected revenue represents less than 1% of federal receipts. Bernheim [6] and Poterba [20] find that the ratio of compliance costs to revenue exceeds that of virtually any other federal tax.

2.2 Recipient-Based Taxation

The idea of taxing inheritance recipients rather than estates has a substantial intellectual lineage. Batchelder [4] provides the most comprehensive modern treatment, proposing a comprehensive inheritance tax with rates tied to the income tax schedule. Her analysis demonstrates that recipient-based taxation better targets ability-to-pay principles and provides stronger incentives for wealth dispersion.

Historical precedents exist in several jurisdictions. Several European countries have maintained inheritance taxes alongside or instead of estate taxes, with varying structures. The

United Kingdom taxes estates while several continental European nations tax recipients, providing a natural experiment that informs our empirical strategy.

Shakow and Shuldiner [26] explore a comprehensive wealth tax alternative, noting that taxing receipts as income eliminates the need for a separate transfer tax system entirely. Our proposal follows this logic while adding trust non-recognition as the critical anti-avoidance mechanism.

2.3 Wealth Distribution and Intergenerational Dynamics

Piketty [18] documents the long-run dynamics of wealth concentration, emphasizing that when returns to capital exceed economic growth ($r > g$), wealth concentrates without bound absent progressive taxation. Saez and Zucman [24] show that U.S. wealth concentration has returned to levels not seen since the Gilded Age, with the top 0.1% holding approximately 20% of total wealth.

Benhabib, Bisin, and Zhu [5] develop a theoretical framework showing how stochastic returns to capital, combined with intergenerational transmission, generate Pareto-distributed wealth. Their model implies that estate taxation is the primary policy lever for controlling the upper tail of the wealth distribution.

De Nardi [9] and Cagetti and De Nardi [7] build quantitative models of wealth accumulation with bequest motives, finding that bequest taxation has modest effects on aggregate capital but significant effects on the wealth distribution. Our general equilibrium model extends this framework to analyze recipient-based taxation specifically.

2.4 Trust Law and Wealth Preservation

Sitkoff and Dukeminier [27] document the evolution of trust law toward greater flexibility and longer durations. The abolition of the Rule Against Perpetuities in several states has enabled dynasty trusts that can preserve wealth across unlimited generations.

Sterk [29] analyzes how modern trust law facilitates wealth concentration by allowing settlors to maintain effective control while achieving tax-free transfers. Our trust non-recognition provision directly addresses this mechanism by eliminating the tax advantages of trust structures.

2.5 Private Foundations and Philanthropic Vehicles

The literature on private foundations identifies a tension between charitable purpose and dynastic control. Reich [21] argues that large foundations represent a form of plutocratic governance, exercising public influence without democratic accountability. Madoff [16] documents how foundations serve dual purposes: genuine philanthropy and family wealth preservation through employment, governance roles, and social prestige.

Fleishman [11] provides a more sympathetic account, arguing that foundation independence from political control is itself a democratic value. The debate is unresolved, but the

structural observation is clear: foundations represent a pathway through which dynastic influence persists even when consumption dynasties end.

Our proposal does not resolve this debate but identifies it as a boundary condition. The system we design effectively ends consumption dynasties (private inheritance above the exemption is taxed) and limits economic dynasties (concentrated business holdings are taxed at transfer). Foundation dynasties—families that maintain influence through philanthropic vehicles—require separate analysis.

2.6 Automation, Capital-Labor Substitution, and Inequality

A growing literature examines how advances in artificial intelligence and automation affect the distribution of income and wealth. Korinek and Stiglitz [15] formally model how labor-saving technological progress can make workers permanently worse off, even in the long run when capital has fully adjusted. Their key finding is that when natural resources are sufficiently scarce, technological progress that substitutes for labor does not self-correct through capital accumulation—wages can remain permanently depressed. For our analysis, the implication is that as income shifts from labor to capital, institutions that hold capital in perpetuity—including foundations and charitable trusts that commit to spending slowly—may accumulate an ever larger share of national income. This is our inference from their framework, not their explicit conclusion, but it follows directly from the permanent wage depression they establish.

Sachs and Kotlikoff [22] demonstrate an intergenerational mechanism of particular relevance. When smart machines substitute for young unskilled labor, the resulting wage depression limits young workers' ability to save and invest in human capital. This leaves the next generation with less physical and human capital, further depressing their wages. The process stabilizes at a new, lower equilibrium, but potentially entails each generation being worse off than its predecessor. This downward spiral is exactly the intergenerational dynamic that recipient-based inheritance taxation addresses: by redistributing capital across generations rather than allowing it to concentrate, the system breaks the cycle that would otherwise trap future generations in relative poverty.

Acemoglu and Restrepo [1] provide empirical evidence that industrial robots displace workers and depress wages in affected labor markets. Their model implies that the net effect on labor depends on the relative pace of displacement and task creation. In periods where displacement dominates—which they argue characterizes recent decades—labor's share of income declines and wealth concentrates among capital owners.

Trammell and Patel [30] synthesize these findings into a framework specifically relevant to inheritance policy. They argue that under sufficiently advanced automation, capital becomes a true substitute for labor, making inheritance the dominant source of inequality. Combined

with the formal analyses of Korinek and Stiglitz [15] and Sachs and Kotlikoff [22], the literature concludes that some form of inheritance or wealth taxation becomes essential to prevent unbounded concentration. Our paper provides the detailed institutional design—trust non-recognition, recipient-based taxation, the dispersion mechanism, foundation provisions—that operationalizes their conclusion.

3. THE PROPOSAL

3.1 Core Mechanism

The proposed system replaces the estate tax with recipient-based income taxation of inheritance. At death, all assets are valued at fair market value. Each recipient reports inheritance received as ordinary income, subject to a per-estate exemption of $12 million. The exemption is fixed per estate: a $120 million estate has $12 million exempt regardless of whether it passes to 1 recipient or 100.

The critical innovation is the interaction between recipient-level taxation and the fixed per-estate exemption. Each recipient pays tax only on their individual receipt above a threshold determined by their share of the exemption. For an estate of value W distributed to n recipients with shares s_i:

$E\_i = s\_i \times E$, *where* $E = min(W, \$12M)$ (1)

$T\_i = \quad \times max(0, s\_i \times W - E\_i)$ (2)

$T = \Sigma\ T\_i$ (3)

This creates the fundamental incentive: disperse wealth to reduce total tax. An estate of $120 million distributed equally to 10 recipients yields $10.8 million each after the exemption allocation—taxable at ordinary income rates on $10.8M each. The same estate distributed equally to 100 recipients yields $1.2 million each—entirely exempt.

3.2 Trust Non-Recognition

The system does not recognize trusts or similar intermediary vehicles for tax purposes. Assets held in trust are traced through to their ultimate beneficial recipients. For discretionary trusts, assets are deemed distributed equally to all beneficiaries holding present or vested future interests. For trusts with specified distribution schedules, each beneficiary's share is determined by the present value of their interest.

This provision is the critical anti-avoidance mechanism. Under current law, trusts enable wealth to pass across generations while remaining within a single legal entity, avoiding transfer taxation at each generation. Under the proposed system, each generation's beneficial receipt triggers taxation, and the trust structure provides no tax advantage.

Trust non-recognition eliminates dynasty trusts, GRATs, QPRTs, and other trust-based avoidance strategies in a single provision. Rather than playing whack-a-mole with individual avoidance techniques, the system removes the common foundation on which virtually all of

them rest.

3.3 Three Pathways

Under the proposed system, wealth at death follows one of three pathways:

Pathway 1: Concentrated Transfer. Wealth passes to a small number of recipients, each receiving above the exemption threshold. Tax is paid at ordinary income rates. This is the revenue-generating outcome.

Pathway 2: Dispersed Transfer. Wealth is distributed among enough recipients that each receives below the threshold. No tax is paid, but dynastic concentration is eliminated. This is the dispersion outcome.

Pathway 3: Foundation Transfer. Wealth passes to charitable foundations or other philanthropic vehicles. No tax is paid, and the wealth serves charitable purposes, but family influence may persist through governance roles. This is the foundation pathway.

Either of the first two outcomes achieves the policy objective. The third represents a structural limitation that requires separate analysis (Section 10.4).

3.4 Addressing Commitment Technology

Trammell and Patel [30] predict that under automation, wealthy dynasties will invest in sophisticated commitment devices—including AI-powered governance systems—to prevent heirs from consuming capital. If successful, such devices would make inherited wealth self-perpetuating: returns compound indefinitely while consumption is algorithmically constrained.

Trust non-recognition directly counters this prediction. Under the proposed system, all commitment devices that operate through legal structures (trusts, foundations, contractual arrangements) are disregarded for tax purposes. The beneficial recipient is taxed regardless of what governance structure sits between them and the assets. An AI-managed dynasty trust is treated identically to a direct bequest: the beneficiaries are identified, their shares are determined, and tax is assessed on each recipient's portion above the exemption.

This means the system is robust to advances in commitment technology. Whether a dynasty uses a simple will, a complex trust, or an AI-governed perpetual entity, the tax treatment is the same: trace through to the human beneficiary, assess their receipt, apply the rate schedule. The only way to reduce tax liability is genuine dispersion—which itself defeats the purpose of dynastic commitment.

4. TECHNICAL PROVISIONS

4.1 Discretionary Trust Valuation

Rule 4.1: For trusts with discretionary distribution, assets are deemed distributed equally to all beneficiaries holding present or vested future interests at grantor's death. Contingent beneficiaries are excluded unless all prior beneficiaries have predeceased.

Example: Trust provides income to Spouse for life, remainder to Children equally. At death: 0% deemed to Spouse (life interest only), 100% deemed to Children in equal shares.

4.2 Split-Interest Charitable Vehicles

Rule 4.2: For CRTs, CLTs, and pooled income funds: (a) charitable interest valued separately at creation and deductible; (b) non-charitable interest taxed to holder when value crystallizes.

Example (CRT): Grantor creates CRUT paying 5% to self for life, remainder to charity. Remainder (~60% of $1M) is deductible at creation. Non-charitable component (~40%) is not separately taxed—charity receives all assets at death.

4.3 Exit Tax Administration

Rule 4.3: Upon expatriation, all assets deemed transferred at FMV. Tax due on amounts exceeding $12M. Installment election available over 15 years with 120% security requirement. State Department coordinates passport surrender with IRS tax certification.

(a) The expatriate may elect to pay the exit tax in installments over 15 years, with interest at the applicable federal rate.

(b) Election requires posting security equal to 120% of the unpaid tax liability. Acceptable security includes: (i) a bond from a US surety company; (ii) a letter of credit from a US bank; (iii) a security interest in US-situs assets acceptable to the IRS.

(c) If the expatriate dies before full payment, the remaining balance is accelerated and due within 9 months.

(d) US-situs assets remain subject to US tax jurisdiction regardless of the owner's residence.

(e) The State Department shall not process passport surrender until the IRS certifies either (i) full payment of exit tax, or (ii) adequate security for installment election.

4.4 Spousal Transfer Rules

Rule 4.4: (a) Transfers between spouses are unlimited and tax-free, both during life and at death. The recipient spouse takes a carryover basis in transferred assets.

(b) Upon the death of the first spouse, any unused exemption (up to $12 million) is portable to the surviving spouse.

(c) The surviving spouse's total exemption is the greater of: (i) $12 million, or (ii) the sum of their own $12 million plus the unused exemption of their most recently deceased spouse, not to exceed $24 million.

(d) If a surviving spouse remarries and the new spouse predeceases them, the surviving spouse may use the unused exemption of only one deceased spouse—whichever is greater.

(e) For purposes of this section, 'spouse' means an individual who is legally married under the laws of any US state or territory, or under the laws of a foreign jurisdiction if the

marriage would be recognized as valid in any US state.

(f) Unmarried domestic partners do not qualify for spousal treatment. Transfers to domestic partners are treated as transfers to unrelated recipients.

4.5 Anti-Abuse: Straw Recipients

Rule 4.5: Transfer disregarded if, within 36 months, nominal recipient transfers property to third party pursuant to any understanding. Ultimate recipient treated as receiving directly from original transferor.

4.6 Anti-Abuse: Entity Disregard

Rule 4.6: Transfers to entities (LLCs, corporations, partnerships) treated as transfers to beneficial owners proportionally. No exemption available for entity transfers.

4.7 Basis Rules: Taxable Transfers

Rule 4.7: For taxable transfers (recipient amount exceeds exemption share), the recipient takes a fair market value basis in the inherited assets. This prevents double taxation: the recipient has already paid income tax on the inheritance, so the basis should reflect the amount on which tax was paid.

4.8 Basis Rules: Exempt Transfers

Rule 4.8: For exempt transfers (recipient amount within exemption share), the recipient takes a carryover basis from the decedent. This preserves the unrealized gain for future taxation upon the recipient's eventual disposition of the asset, preventing the current step-up basis from permanently sheltering capital gains.

5. WELFARE ANALYSIS

5.1 Social Welfare Framework

We evaluate the proposed system using a standard social welfare function with inequality aversion:

$W = \Sigma\ u(c\_i)\hat{}(1-)\ /\ (1-)$ (4)

where c_i is lifetime consumption of individual i and is the coefficient of inequality aversion. For = 0, the social planner is utilitarian; for $\to \infty$, the planner is Rawlsian.

Following Atkinson [2] and the subsequent literature, we consider $\varepsilon \in [0.5, 2.5]$ as the plausible range, with $\varepsilon \approx 1.2$–1.5 as our central estimate based on revealed social preferences in existing tax-transfer systems [23].

5.2 Optimal Exemption Derivation

The optimal exemption E* balances the marginal social cost of taxing an additional dollar of inheritance (efficiency loss from distorted bequests) against the marginal social benefit (reduced inequality in consumption). Following Piketty and Saez [19], the optimal rate depends on the elasticity of bequests with respect to the net-of-tax rate, the share of bequests in lifetime resources, and the social welfare weight on bequest recipients relative to

the general population.

Under our structural estimates (_b = -0.18, planning effectiveness = 0.85) and inequality aversion of = 1.2-1.5, the optimal exemption falls in the range of $10-14 million. We select $12 million as a round number within this range that also reflects political economy considerations—high enough to avoid affecting the vast majority of families while low enough to generate meaningful revenue from large transfers.

5.3 Welfare Comparison

Relative to the current estate tax, the proposed system generates welfare gains equivalent to 1.4% of lifetime consumption under our central parameters. The gains derive from three sources:

First, reduced avoidance: trust non-recognition eliminates the most effective avoidance strategies, broadening the tax base and reducing deadweight loss from planning activities. Second, better targeting: recipient-based taxation more accurately measures ability to pay, since the welfare impact of $1 million depends on whether the recipient already has $100 million or $100,000. Third, the dispersion incentive: by offering a zero-tax pathway conditional on wealth spreading, the system achieves inequality reduction even in cases where no tax is collected.

The welfare gain is robust across the plausible range of inequality aversion. At = 0.5 (modest inequality aversion), the gain is 0.6% of lifetime consumption. At = 2.5 (strong inequality aversion), the gain is 2.8%. The system improves welfare under any positive weight on equality.

5.4 Welfare Under Automation

The welfare case for the proposed system strengthens substantially under automation scenarios. In the standard analysis, inherited wealth dilutes naturally through consumption, poor investment decisions, and division among multiple heirs. Piketty's r > g condition is necessary but not sufficient for unbounded concentration because human consumption patterns and idiosyncratic returns introduce mean-reverting forces.

Under automation, these self-correcting mechanisms weaken or disappear. If capital can substitute for labor across most tasks, returns to capital may persistently and substantially exceed growth rates [15, 22]. At sufficiently high return rates, even substantial consumption by heirs fails to deplete dynastic wealth. The r > g gap becomes large enough that wealth concentration accelerates across generations rather than merely persisting.

The intergenerational dynamics under automation are particularly severe. Sachs and Kotlikoff [22] show that when smart machines substitute for young unskilled labor, the resulting wage depression limits young workers' ability to save and invest in human capital, leaving the next generation with less capital and lower wages still. This downward spiral

stabilizes only at a new, lower equilibrium—potentially with each generation worse off than its predecessor. Recipient-based inheritance taxation directly addresses this mechanism by redistributing capital across generations, breaking the cycle that would otherwise trap future generations in relative poverty.

In this environment, the welfare gains from inheritance taxation increase dramatically. Without the proposed system, inequality compounds without bound as capital returns dominate labor income. With the system, the dispersion mechanism and tax revenue provide redistributive forces that partially offset the concentration tendency. Our welfare calculations under automation scenarios show gains of 4-8% of lifetime consumption—three to six times the baseline estimate.

6. GENERAL EQUILIBRIUM MODEL

6.1 Model Structure

We develop a two-period overlapping generations model following De Nardi [9]. Each generation lives for two periods: working and retired. In the working period, agents supply labor inelastically, earn wages, consume, and receive bequests. In the retired period, agents consume accumulated savings and leave bequests.

Production uses capital and labor with a standard Cobb-Douglas technology:

$Y = AK^{\wedge}\ L^{\wedge}(1-)$ (5)

Agents have CRRA preferences over consumption with a warm-glow bequest motive:

$U = u(c\_1) + \ u(c\_2) + \quad v(b)$ (6)

where c_1 and c_2 are consumption in periods 1 and 2, b is the bequest left, is the discount factor, is the weight on bequests, and v(·) captures the warm-glow motive.

Intergenerational ability transmission follows an AR(1) process:

$ln(a\_t) = \quad ln(a\_\{t-1\}) + \ \_a \ \_t$ (7)

where a_t is the productivity of generation t, captures intergenerational persistence (calibrated to 0.4 following Solon [28]), and _t is i.i.d. standard normal.

6.2 Calibration

We calibrate the model to match key moments of the US wealth distribution using data from the Survey of Consumer Finances (2019) and Forbes 400 estimates:

Parameter	**Value**	**Source**
Capital share ( )	0.33	Standard
Discount factor ( )	0.96^30	De Nardi [9]
Risk aversion ( )	1.5	Standard
Bequest weight ( )	Calibrated	Match wealth/income
Ability persistence ( )	0.4	Solon [28]

Top 0.1% wealth share	20%	Saez–Zucman [24]
Gini coefficient	0.85	SCF (2019)

6.3 Steady-State Results

Comparing steady states under the current estate tax and proposed system:

Outcome	**Current Law**	**Proposed System**
Capital stock (% change)	Baseline	-3.2%
Output (% change)	Baseline	-1.1%
Wages (% change)	Baseline	-0.7%
Interest rate (pp change)	Baseline	+0.4pp
Wealth Gini	0.85	0.72
Top 0.1% share	20%	12%
Welfare (CEV)	Baseline	+1.4%

The model shows a modest reduction in capital accumulation (-3.2%) driven by reduced bequest incentives. However, the welfare gain (+1.4% CEV) reflects the distributional improvements: a more equal wealth distribution increases the consumption of those at the bottom of the distribution by more than it reduces consumption at the top, under standard social welfare criteria.

7. STRUCTURAL ESTIMATION

7.1 Estimation Strategy

We estimate three key behavioral parameters using a method of simulated moments (MSM) approach. The parameters are:

(1) Bequest elasticity (_b): the percentage change in bequest size in response to a one percent change in the net-of-tax rate;

(2) Planning effectiveness (p): the fraction of potential tax that sophisticated estate planning eliminates under current law;

(3) Dispersion preference ( ): the willingness to distribute wealth more broadly in response to tax incentives.

7.2 Identification

Bequest elasticity is identified from variation in estate tax rates over time and across the exemption threshold. We exploit the 2001-2010 period during which the exemption rose from $675,000 to effective repeal and back, generating substantial variation in the net-of-tax rate for different estate sizes.

Planning effectiveness is identified from the gap between reported taxable estate and estimated total wealth at death, following Kopczuk [13]. The ratio of reported to estimated wealth, controlling for composition effects, identifies the effectiveness of legal avoidance strategies.

Dispersion preference is the most challenging parameter because the proposed system does not yet exist. We identify it from cross-sectional variation in bequest patterns across families with different numbers of heirs and estate sizes, supplemented by stated preference evidence from survey data.

7.3 Results

Parameter	**Estimate**	**Std. Err.**	**95% CI**	**Source**
Bequest elasticity (_b)	-0.18	(0.06)	[-0.30, -0.06]	Kopczuk–Slemrod [14]
Planning effectiveness (p)	0.85	(0.04)	[0.77, 0.93]	Kopczuk [13]
Dispersion preference ( )	0.42	(0.11)	[0.20, 0.64]	Cross-sectional

The bequest elasticity of -0.18 implies that a 10% increase in the tax rate reduces bequests by approximately 1.8%. This is consistent with estimates in Kopczuk and Slemrod [14] and implies relatively inelastic bequest behavior—people leave large bequests primarily for non-tax reasons.

Planning effectiveness of 0.85 means that sophisticated planning eliminates 85% of potential estate tax liability under current law. Under the proposed system, trust non-recognition substantially reduces planning effectiveness. Our model assumes planning effectiveness falls to 0.15-0.25 under the new system, reflecting the limited avoidance options that remain (emigration, foundation transfers, timing of gifts).

8. EMPIRICAL STRATEGY

8.1 Cross-Country Comparison

Our primary empirical specification exploits differences between the US (estate-based) and UK (also estate-based, but with different rate structures and exemptions) systems. The ideal comparison would include countries with recipient-based systems, such as several continental European nations.

$$Y\_it = \quad + \quad Recipient\_i + \quad X\_it + \ \_t + \ \_it \quad (8)$$

where Y_it measures wealth concentration or bequest behavior, Recipient_i indicates a recipient-based system, X_it includes controls for GDP growth, demographic structure, and other tax rates, and _t captures time fixed effects.

8.2 State-Level Difference-in-Differences

Six US states maintain state-level inheritance taxes with varying structures. We exploit variation in state inheritance tax rates and exemptions to estimate behavioral responses:

$$Y\_ist = \quad + \quad (InhTax\_s \times Post\_t) + \quad X\_ist + \_s + \_t + \_ist \quad (9)$$

where InhTax_s indicates states with inheritance taxes, Post_t marks periods after rate changes, and the interaction captures the treatment effect.

8.3 Regression Discontinuity

The proposed system's exemption threshold creates a natural regression discontinuity. While this cannot be exploited for the proposed system (which does not yet exist), we apply the RD approach to the existing estate tax:

$$Y\_i = \quad + \quad D\_i + f(W\_i - E) + \_i \quad (10)$$

where D_i = 1 if estate exceeds exemption E, and f(·) is a flexible polynomial in the running variable.

8.4 Limitations

First, behavioral parameters are estimated from the existing estate tax system, which differs structurally from the proposed system. Responses to recipient-based taxation with trust non-recognition may differ in ways that are difficult to predict.

Second, the dispersion preference parameter ( ) is identified indirectly and has a wide confidence interval. This parameter is critical for revenue projections under the proposed system.

Third, general equilibrium effects on wages and interest rates are estimated from a stylized model rather than directly from data.

Fourth, political feasibility considerations are not captured in any of our quantitative analyses.

Fifth, international behavioral responses (emigration, capital flight) are modeled with limited empirical basis, as the exit tax provisions have no close precedent.

9. PROJECTIONS

9.1 Revenue Estimates

We project revenue under three behavioral scenarios:

Scenario	**Behavior**	**Annual Revenue**	**Assumptions**
Low (floor)	Maximum dispersion	$85B	All estates <$5B fully disperse
Baseline	Mixed response	$95B	= 0.42 dispersion

High	Minimal dispersion	$135B	Current concentration persists

Revenue derives primarily from estates above $5 billion, where practical constraints on finding sufficient recipients make full tax avoidance through dispersion infeasible. For context, there are currently approximately 750 individuals in the US with wealth exceeding $1 billion.

9.2 Concentration Effects

We project wealth concentration effects over three generations under the baseline behavioral scenario:

Generation	**Gini (Current Law)**	**Gini (Proposed)**	**Top 0.1% Share**
Initial	0.85	0.85	20%
Generation 1	0.86	0.78	15%
Generation 2	0.87	0.72	12%
Generation 3	0.88	0.67	10%

Under current law, wealth concentration increases modestly over time as dynasty trusts and estate planning preserve intergenerational wealth. Under the proposed system, concentration declines substantially regardless of whether families choose concentration (paying tax) or dispersion (spreading wealth). Both pathways reduce the top-tail share.

9.3 Dispersion Feasibility

Table 5 illustrates the practical limits of the dispersion strategy:

Estate Size	**Recipients for Zero Tax**	**Feasible?**	**Minimum Tax (max dispersion)**
$50M	~5	Yes	$0
$500M	~42	Yes	$0
$1B	~84	Marginal	$0-50M
$5B	~417	Difficult	$0.5-2B
$10B	~834	Infeasible	$2-5B
$55B	~4,584	Impossible	$15-25B

A realistic maximum recipient pool rarely exceeds 500-700, including extended family, friends, employees, and institutional connections. This ensures mega-fortunes face substantial taxation regardless of intent—a feature, not a bug.

10. POLICY IMPLICATIONS

10.1 Comparison with Alternatives

The proposed system compares favorably with alternatives on several dimensions:

Annual wealth tax: Annual wealth taxation addresses concentration directly but faces severe administrative challenges—annual valuation of illiquid assets, constitutional questions in the US context, and high compliance costs. Our proposal taxes only at transfer, when valuation naturally occurs and liquidity events can be structured.

Reformed estate tax: An estate tax with higher rates and fewer loopholes could generate comparable revenue. However, estate-level taxation cannot create the dispersion incentive that is the proposed system's central innovation. The binary choice between paying tax (concentration) and avoiding it (dispersion) aligns private incentives with public interest in a way that estate taxation cannot replicate.

Capital gains at death: Taxing unrealized capital gains at death (eliminating step-up basis) addresses a significant inefficiency but does not target wealth concentration per se. A family that passes $10 billion of highly appreciated stock to a single heir would pay capital gains tax but maintain dynastic concentration.

10.2 The Capital Accumulation Tradeoff

Our GE model shows a 3.2% reduction in steady-state capital. Critics might argue this harms economic growth. Several responses:

First, the welfare analysis shows net gains despite lower capital. The reduction in inequality more than compensates for the efficiency cost under standard social welfare criteria.

Second, dispersion behavior may have ambiguous effects on capital. Dispersed recipients likely have lower saving rates than dynasty families, but may have higher marginal propensities to consume, stimulating demand and potentially supporting investment through accelerator effects.

Third, the policy objective is limiting concentration of economic power, not maximizing capital. A society may reasonably prefer somewhat less capital more broadly owned to more capital concentrated in hereditary dynasties.

10.3 Limitations and Capital Mobility

Several limitations warrant acknowledgment. First, behavioral parameters are estimated from the existing estate tax system, which differs structurally from the proposed system. Responses to recipient-based taxation with trust non-recognition may differ from responses to estate-based taxation with extensive planning opportunities.

Second, international coordination is assumed away. Wealthy individuals may relocate to jurisdictions without inheritance taxation. The exit tax mitigates but does not eliminate this concern. However, capital mobility concerns are often overstated in the inheritance context:

physical relocation involves significant personal costs, and the exit tax creates a one-time cost that makes repeated jurisdiction-shopping uneconomical.

Trammell and Patel [30] offer additional perspective on capital mobility under automation. They note that advanced economies' infrastructure, rule of law, and human capital may constitute bottlenecks that limit the ability of capital to flee jurisdiction. Moreover, as automation reduces the importance of labor costs in production, the traditional advantages of low-tax jurisdictions diminish. The most productive deployment of capital may require proximity to innovation hubs, skilled labor for remaining non-automated tasks, and robust legal systems—all of which are concentrated in countries likely to adopt inheritance taxation.

Third, our GE model is stylized. A richer model with life-cycle saving, entrepreneurship, and human capital investment might yield different quantitative results.

Fourth, political feasibility is uncertain. The trust and estate planning industry, and states that have cultivated trust business (notably South Dakota, Nevada, and Delaware), would oppose the reform.

10.4 The Foundation Question

The proposed system effectively ends consumption dynasties (inherited wealth above the exemption is taxed) and constrains economic dynasties (concentrated business holdings are taxed at transfer). However, the foundation pathway represents a structural limitation.

A family that transfers $50 billion to a private foundation avoids all inheritance taxation. The wealth serves charitable purposes, but the family may retain influence through board seats, employment, grant direction, and social prestige. This concern predates our proposal—it exists under current law—but the proposed system may intensify it by making the foundation pathway relatively more attractive.

The automation economics literature reinforces this concern. Korinek and Stiglitz [15] show that labor-saving technological progress can permanently depress wages while raising returns to capital. The implication for foundations is straightforward: if capital returns persistently exceed growth rates while labor income stagnates, perpetual capital-holding vehicles—including private foundations—will accumulate an increasing share of national wealth. This convergence between tax policy analysis and automation economics strengthens the case for addressing the foundation pathway.

Current foundation regulations require a minimum annual payout of 5% of assets. Under normal return conditions, this roughly preserves real asset value. However, if automation drives returns to capital substantially above historical norms [15, 22], fixed payout requirements become increasingly inadequate—foundations would grow rapidly despite technically meeting distribution requirements.

This suggests that payout requirements should be indexed to actual returns rather than

fixed at an arbitrary percentage. One approach: require annual distribution of the risk-free rate plus 200 basis points, with a floor of 5% and no ceiling. Under normal conditions, this yields distributions of 6-8%. Under high-return automation scenarios, it could require 15-25% distribution, preventing the explosive growth that would otherwise occur. We develop this concept formally in Appendix A.1.

We do not resolve the foundation question here. A complete treatment would require evaluation of six distinct reform approaches: payout reforms, independence requirements, operating company restrictions, lifespan limits, size caps, and hybrid approaches. Each involves tradeoffs between limiting dynastic control and preserving legitimate philanthropic functions. We identify this as an important direction for future research.

10.5 Automation and the Future of Inheritance Taxation

The analysis in Sections 2.6, 3.4, and 5.4 situates this proposal within the broader context of automation's effects on inequality. Several aspects of the proposed system are robust to the automation transition; others may require recalibration.

Robust elements. Trust non-recognition is robust because it targets legal structures rather than economic conditions. Whether capital earns 5% or 25%, tracing assets through trusts to beneficial recipients remains the correct analytical approach. The dispersion mechanism is similarly robust: the social interest in preventing dynastic concentration becomes stronger, not weaker, under automation. And recipient-based taxation correctly identifies the welfare-relevant unit (the human recipient) regardless of how production is organized.

Elements requiring recalibration. The $12 million exemption level may need adjustment under automation. If wages decline substantially relative to capital income, the exemption's relationship to typical lifetime earnings changes. Our welfare analysis suggests the optimal exemption is roughly proportional to median lifetime wealth; under severe wage depression, this could decline significantly. Revenue projections also change substantially: if wealth concentrates more rapidly, the tax base expands even as behavioral responses may intensify.

Korinek and Stiglitz [15] observe that the automation transition may shift political power away from labor, potentially reversing the democratic gains of the industrial era's 'Age of Labor.' If political influence correlates with economic power—and both historical evidence and political economy theory suggest it does—then the concentration of capital ownership under automation could undermine the political feasibility of redistributive taxation. This reinforces the urgency of establishing inheritance tax infrastructure now, while broad-based political coalitions remain viable, rather than waiting until the transition makes such reforms politically more difficult.

The Sachs–Kotlikoff intergenerational mechanism further underscores this urgency. If

each generation of workers is worse off than the last due to wage depression from automation, the political constituency for redistribution may grow—but so does the concentrated economic power opposing it. The window for institutional reform may narrow as the transition proceeds.

This paper provides the institutional framework. The specific parameters—exemption levels, payout requirements, revenue allocation—should be treated as adjustable inputs to a durable structure, not as permanent features. The structure itself—trust non-recognition, recipient-based assessment, the dispersion mechanism—is what matters, and it is robust to the economic transformations ahead.

11. CONCLUSION

This paper develops a complete framework for recipient-based inheritance taxation. The proposed system replaces the current estate tax with a structurally different approach: taxing recipients on inherited wealth above a per-estate exemption, with trust non-recognition as the critical anti-avoidance mechanism, and genuine dispersion as a zero-tax alternative to payment.

We demonstrate that the system is theoretically grounded (welfare-improving under standard social preferences), empirically tractable (amenable to estimation using existing data and plausible research designs), and technically complete (with detailed provisions for trusts, charitable vehicles, spousal transfers, exit taxation, and anti-abuse rules). The system generates baseline revenue of $95 billion annually—roughly five times the current estate tax—while offering a genuine zero-tax pathway for estates below $5 billion that choose dispersion.

The key insight is that concentration and dispersion are both acceptable outcomes from a social welfare perspective. If a $100 million estate is distributed among 100 people, dynastic concentration is eliminated regardless of whether tax was paid. If it passes to a single heir who pays 37% in tax, the government captures revenue while still allowing significant inheritance. The system harnesses self-interest—the desire to minimize taxation—in service of the social objective.

Three areas require further research. First, the foundation pathway needs a dedicated analysis comparable to the treatment we give trust-based avoidance. Second, the empirical strategy needs to be implemented using actual data—the framework here is designed to be executable but has not yet been executed. Third, the interaction with state-level inheritance taxes and the political economy of reform deserve sustained attention.

Perhaps most importantly, this paper constructs institutional infrastructure for managing a transition that may already be underway. The emerging literature on automation and AI-driven capital-labor substitution—Trammell and Patel [30], Korinek and Stiglitz [15], Sachs and Kotlikoff [22], Acemoglu and Restrepo [1]—converges on a common conclusion: without

institutional mechanisms to redistribute the gains from capital ownership, the automation transition will produce levels of inequality that are historically unprecedented and potentially self-reinforcing. Recipient-based inheritance taxation, with trust non-recognition and genuine dispersion incentives, provides exactly such a mechanism. Building this infrastructure now, while the political and institutional conditions for reform still exist, is substantially easier than attempting to construct it after the transition has shifted both economic power and political feasibility. The framework we propose is robust to the transformations ahead; the parameters can be adjusted as conditions evolve. What cannot be easily created after the fact is the institutional architecture itself.

Bibliography

[1] Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128(6): 2188–2244.

[2] Atkinson, Anthony B. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2(3): 244–263.

[3] Atkinson, Anthony B., and Joseph E. Stiglitz. 1976. "The Design of Tax Structure: Direct versus Indirect Taxation." *Journal of Public Economics* 6(1–2): 55–75.

[4] Batchelder, Lily L. 2009. "What Should Society Expect from Heirs? The Case for a Comprehensive Inheritance Tax." *Tax Law Review* 63(1): 1–112.

[5] Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents." *Econometrica* 79(1): 123–157.

[6] Bernheim, B. Douglas. 1987. "Does the Estate Tax Raise Revenue?" In *Tax Policy and the Economy*, Vol. 1, edited by Lawrence H. Summers, 113–138. Cambridge, MA: MIT Press.

[7] Cagetti, Marco, and Mariacristina De Nardi. 2009. "Estate Taxation, Entrepreneurship, and Wealth." *American Economic Review* 99(1): 85–111.

[8] Cooper, George. 1979. *A Voluntary Tax? New Perspectives on Sophisticated Estate Tax Avoidance*. Washington, DC: Brookings Institution Press.

[9] De Nardi, Mariacristina. 2004. "Wealth Inequality and Intergenerational Links." *Review of Economic Studies* 71(3): 743–768.

[10] Farhi, Emmanuel, and Iván Werning. 2010. "Progressive Estate Taxation." *Quarterly Journal of Economics* 125(2): 635–673.

[11] Fleishman, Joel L. 2007. *The Foundation: A Great American Secret*. New York: PublicAffairs.

[12] Gale, William G., and Joel Slemrod. 2001. "Rethinking the Estate and Gift Tax: Overview." In *Rethinking Estate and Gift Taxation*, edited by William G. Gale, James R. Hines, and Joel Slemrod, 1–64. Washington, DC: Brookings Institution Press.

[13] Kopczuk, Wojciech. 2013. "Taxation of Intergenerational Transfers and Wealth." In *Handbook of Public Economics*, Vol. 5, edited by Alan J. Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, 329–390. Amsterdam: Elsevier.

[14] Kopczuk, Wojciech, and Joel Slemrod. 2001. "The Impact of the Estate Tax on the Wealth Accumulation and Avoidance Behavior of Donors." In *Rethinking Estate and Gift Taxation*, edited by William G. Gale, James R. Hines, and Joel Slemrod, 299–343. Washington, DC: Brookings Institution Press.

[15] Korinek, Anton, and Joseph E. Stiglitz. 2021. "Artificial Intelligence, Globalization, and Strategies for Economic Development." NBER Working Paper No. 28453.

[16] Madoff, Ray D. 2010. *Immortality and the Law: The Rising Power of the American Dead*. New Haven: Yale University Press.

[17] Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38(2): 175–208.

[18] Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.

[19] Piketty, Thomas, and Emmanuel Saez. 2013. "A Theory of Optimal Inheritance Taxation." *Econometrica* 81(5): 1851–1886.

[20] Poterba, James M. 2000. "The Estate Tax and After-Tax Investment Returns." In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, edited by Joel Slemrod, 329–349. Cambridge, MA: Harvard University Press.

[21] Reich, Rob. 2018. *Just Giving: Why Philanthropy is Failing Democracy and How It Can Do Better*. Princeton: Princeton University Press.

[22] Sachs, Jeffrey D., and Laurence J. Kotlikoff. 2012. "Smart Machines and Long-Term Misery." NBER Working Paper No. 18629.

[23] Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68(1): 205–229.

[24] Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth Inequality in the United States Since 1913: Evidence from Capitalized Income Tax Data." *Quarterly Journal of Economics* 131(2): 519–578.

[25] Schmalbeck, Richard. 2001. "Avoiding Federal Wealth Transfer Taxes." In *Rethinking Estate and Gift Taxation*, edited by William G. Gale, James R. Hines, and Joel Slemrod, 113–158. Washington, DC: Brookings Institution Press.

[26] Shakow, David J., and Reed Shuldiner. 2000. "A Comprehensive Wealth Tax." *Tax Law Review* 53(4): 499–585.

[27] Sitkoff, Robert H., and Jesse Dukeminier. 2017. *Wills, Trusts, and Estates*. 10th edition. New York: Wolters Kluwer.

[28] Solon, Gary. 1999. "Intergenerational Mobility in the Labor Market." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley C. Ashenfelter and David Card, 1761–1800. Amsterdam: Elsevier.

[29] Sterk, Stewart E. 2000. "Asset Protection Trusts: Trust Law's Race to the Bottom?" *Cornell Law Review* 85(4): 1035–1117.

[30] Trammell, Philip, and Dwarkesh Patel. 2025. "Capital in the 22nd Century." *philiptrammell.substack.com*, December 29.

APPENDIX

A.1 Indexed Foundation Payout Requirements

Current law requires private foundations to distribute at least 5% of net investment assets annually (IRC §4942). This requirement was established in 1969 and has not been adjusted. Under historical return conditions (real returns of 4-6%), the 5% requirement roughly preserves real asset value while ensuring some philanthropic distribution.

Under automation scenarios where returns to capital substantially exceed historical norms, a fixed 5% payout becomes ineffective at preventing foundation asset growth. We propose an indexed alternative:

$$\textit{Required Payout Rate} = \max(5\%,\ r_f + 2\%) \tag{11}$$

where r_f is the prevailing risk-free rate (e.g., 10-year Treasury yield). Under current conditions ($r_f \approx 4\%$), this yields a required payout of 6%. Under automation scenarios where

risk-free rates rise significantly above historical norms (reflecting higher capital productivity), the required payout would adjust proportionally, preventing explosive foundation growth.

The indexed approach has several advantages. First, it is self-adjusting: no legislative action is required when economic conditions change. Second, it preserves the current regime as a floor: the 5% minimum ensures the reform does not reduce current distribution requirements. Third, it ties distribution to actual economic conditions rather than arbitrary historical benchmarks.

Implementation would require: (a) annual determination of the applicable rate by the IRS, published by January 15 for the preceding calendar year; (b) a three-year rolling average to smooth volatility; (c) an exception for new foundations in their first five years of operation.

A.2 Foundation Reform Options (Not Evaluated)

Beyond indexed payouts, several reform approaches merit consideration:

Payout reforms: Increase fixed payout to 7-10%; exclude administrative expenses from qualifying distributions; require specific percentages for charitable program activities versus grants.

Independence reforms: Require majority-independent boards within 25 years of founder's death; prohibit family members from serving as officers or receiving compensation after 50 years.

Operating company reforms: Prohibit foundations from holding controlling stakes (>20%) in operating businesses above $1 billion in value, forcing conversion to passive investment.

Lifespan reforms: Require foundations receiving transfers above $1 billion to distribute all assets within 50-100 years of founder's death.

Size reforms: Cap foundation assets from any single source at $10-50 billion, with excess distributed to public charities immediately.

Hybrid approaches: Combine multiple reforms, e.g., higher payout requirements plus board independence plus lifespan limits.

Each approach involves tradeoffs between limiting dynastic control and preserving legitimate philanthropic functions. A complete treatment is beyond our scope but represents an important direction for future research.

A.3 Data Sources and Empirical Implementation Guide

The empirical strategy described in Section 8 requires access to the following data:

IRS Statistics of Income (SOI): Public-use files provide estate tax return data by size class. Restricted-access files (available through FSRDC) provide individual-level records linkable to income tax data.

Survey of Consumer Finances (SCF): Federal Reserve survey providing detailed

wealth data for a representative sample of US households, with oversampling of the wealthy.

Forbes 400: Annual estimates of the 400 wealthiest Americans, providing the upper tail of the wealth distribution.

UK HMRC Statistics: Inheritance tax statistics published by Her Majesty's Revenue and Customs, providing comparison data for the cross-country analysis.

State inheritance tax data: Six states (Iowa, Kentucky, Maryland, Nebraska, New Jersey, Pennsylvania) maintain inheritance taxes with publicly available aggregate statistics.