Correlated Intelligence:

AI Adoption and Systemic Risk

Dalin Sheng

Southwestern University of Finance and Economics

shengdl@swufe.edu.cn

July 2026

Abstract

Banks delegate decisions to AI models from a common vendor frontier that bundles accuracy with error correlation: the most accurate models share lineage and fail together. The cross-bank cost of that correlation remains unpriced even when lineage is observable, because a credit spread reprices the borrower, not the banks its choice endangers. Model diversity is inefficiently low; under pooled pricing, welfare losses grow quadratically with excess adoption. Frontier improvements can reduce welfare by tipping the market into monoculture. Disclosure makes a correlation surcharge feasible but does not substitute for it. (*JEL* D62, G21, G28, O33)

1 Introduction

On July 19, 2024, a single content update to one widely deployed security product disrupted an estimated 8.5 million Windows devices, grounding flights and halting hospital systems; direct losses to U.S. Fortune 500 firms other than Microsoft were estimated at $5.4 billion (Microsoft 2024; Parametrix 2024). No attack occurred and no individual deployment was negligent. The loss came from *correlation*: thousands of organizations had, one by one and each for good reasons, adopted the same technology, so a single flaw arrived everywhere at once. The question this paper studies is whether markets, left to themselves, price that kind of risk before it materializes.

Banks are now making the same choice at the core of their business. Credit scoring, fraud detection, risk management, and trading increasingly run on artificial-intelligence models licensed from a small set of frontier vendors. The technology has a property that makes the CrowdStrike episode a preview of what banks now face: *accuracy and correlation are bundled*. In a study of more than 350 large language models, Kim et al. (2025) document that, on one benchmark, two models agree on the erroneous answer in about sixty percent of cases in which both err and that larger and more accurate models exhibit *more* correlated errors, even across vendors with distinct architectures. A bank that wants the most accurate model may therefore be unable to buy its way out of the crowd without sacrificing first-order accuracy. Section 9.3 treats the step from benchmark tasks to bank decision data as a measurement problem in its own right.

Standard theory suggests two reasons not to worry. The first is market discipline: risks that creditors can observe are priced, so a bank that loads on a crowded technology should face a wider spread. The second is self-insurance: in Wagner (2011), institutions dislike failing together (joint liquidations are costly to them), so they endogenously choose to differ

from one another, and diversity supplies itself. Both arguments fail here, and they fail for the same reason: the *incidence* of the harm, not the information of the parties. When a bank joins the crowd, most of the incremental cost of a joint collapse falls on other banks' creditors and on institutions that finance resolution through deposit-insurance assessments; taxpayers enter only in tail states. The spread the bank pays can price what its choice does to its own creditors; no bilateral contract prices what it does to everyone else's. The missing margin is not observability but addressability: correlation has no natural counterparty.

This paper asks whether a banking system in which each bank privately optimizes its model choice delivers the efficient degree of *model diversity*, and what regulation restores it when it does not. The answer is organized around a single friction with an unusual property. When bank failures become correlated because banks share model lineage, the cost of that correlation is a cross-bank externality that *bilateral financial contracts cannot price even under full information*. A bank's credit spread can charge the bank for its own default risk and, if the bank's model choice is observable, for the incremental fire-sale exposure of its own creditors. But when bank i adopts the model that bank k already uses, the parties injured are k's creditors and the institutions outside the contracting pair that mutualize resolution costs. No spread charged *to* i prices that harm, and no spread charged *to* k disciplines i's choice: *pricing reprices the victim, not the deviator*. Correlation is therefore unpriced in equilibrium: first because model lineage is unobservable to financiers (vendors do not disclose training provenance), and, more fundamentally, because even mandatory disclosure attaches the price to the wrong margin.

The economy has two dates. Two (then N) limited-liability banks finance a unit investment with competitively priced uninsured debt and choose between two models: a frontier model H with low failure probability p_H and loading λ on a common "lineage factor," and

a legacy model L with higher failure probability p_L and idiosyncratic errors. The menu embodies the Kim et al. (2025) frontier: accuracy and correlation come together. Joint failures are more costly than isolated ones: simultaneous liquidations create a net resource loss φ per bank and an external social cost γ, so the excess social cost of a joint failure is $\Lambda = 2\varphi + \gamma$. The single statistic that organizes the entire paper is the *excess joint exposure* of model monoculture, $\kappa = p_H[\lambda(1 - p_H) - \Delta_a]$, the increase in the probability that two banks fail *together* when the second bank leaves the legacy model and joins the first on the frontier model.

Three results form the paper's spine. First, *under-diversification* (Propositions 1–3). Every pricing regime is summarized by one number: the internalization coefficient ℓ, the share of the joint-failure cost Λ that a bank's own objective registers per unit of κ. Private adoption follows the rule "adopt the frontier model iff $\Delta > \ell\kappa$," where Δ is the private value of accuracy, while the planner's rule replaces ℓ with Λ. Within the baseline economy, the relevant coefficients are

$$\underbrace{0}_{\text{pooled pricing}} \leq \underbrace{\varphi}_{\text{observable lineage or own full liability}} < \underbrace{\Lambda}_{\text{planner}}.$$

Under pooled pricing the frontier model is strictly dominant and monoculture is the unique equilibrium for *any* accuracy edge, however small. Mandatory lineage disclosure changes the internalized coefficient from zero to φ: the bank's own creditors reprice, but the cross-bank term $(\varphi+\gamma)\kappa$ (the harm to the *other* bank's creditors and to parties outside the contracting pair) survives full observability. In the Wagner (2011) benchmark, $\gamma = 0$ and hence $\varphi = \Lambda/2$: even when banks bear their own losses with unlimited liability and endogenously seek heterogeneity, each internalizes exactly half of that benchmark's joint-failure cost. Diversity is a

public good *within* the pair. None of this hinges on simple debt. A spot-contract invariance theorem shows that every bilateral arrangement that is competitively refinanced and breaks even profile by profile induces the same coefficient φ, regardless of its within-profile state contingency. The internalization coefficient is, in general, *the contracting coalition's stake*: an arrangement among both bank–financier pairs reaches 2φ, and the coefficient Λ requires including or pricing the outside bearers of γ.

Second, *scale* (Proposition 5 and Corollary 1). With N banks the same statistic generalizes: the marginal adopter's excess joint exposure is $X(m) = (m-1)\kappa - (N-m)p_L\Delta_a$, growing linearly in the herd m because each entrant endangers every incumbent, and the equilibrium is a threshold rule in $\ell X(m)$. Holding the pairwise cost coefficients fixed, the welfare consequence splits sharply by regime: under pooled pricing the loss from unpriced correlation grows with the *square* of the number of over-adopting banks, while a fixed partial-internalization coefficient $\ell \in (0, \Lambda)$ caps the number of excess adoptions at a constant that does not grow with N. When $\varphi > 0$, lineage disclosure separates the quadratic loss from a bounded residual; the corrective surcharge removes that residual from pure-strategy allocations.

Third, the *progress paradox* (Propositions 6–7). Because progress on the frontier raises the private prize of accuracy *and* erodes the priced deterrent (κ falls as failure probabilities fall), a first-order improvement in model accuracy can tip the economy from diversity into monoculture, and welfare drops discontinuously by $(\Lambda - \ell)\kappa$, several times the improvement's direct benefit in the parameter sweeps we report. The paradox is an equilibrium phenomenon, not a technological one: the planner's welfare is continuous and increasing in progress, and the drop shrinks to zero as $\ell \to \Lambda$. Its incidence is selective. Under pooled pricing there is nothing to tip: the economy is already in monoculture. When $\varphi > 0$, the paradox is

therefore a property of the *disclosure* regime. Repairing observability without repairing the price completes only half the mechanism and thereby creates the possibility that better models make the system worse. With N banks the single drop disaggregates into a staircase of losses, one per bank tipped, each equal to $(\Lambda/\ell - 1)\Delta$ evaluated at its trigger: *successive* upgrades do increasing damage as the herd swells.

The policy analysis (Section 7) draws these threads into a two-part prescription: disclose lineage, then price the exposure it reveals. An ex-ante correlation surcharge of $s = \Lambda - \ell$ per unit of audited joint-failure exposure aligns private and social pure-strategy thresholds, but it is implementable only if lineage is observable: an ex-ante charge cannot condition on a hidden model choice, and an ex-post charge on a failed bank is uncollectible under limited liability. Disclosure is therefore the *precondition* for the corrective instrument, not a substitute for it: disclosure converts a quadratic loss into a bounded residual, the surcharge removes the residual, and disclosure *without* the surcharge can create the progress paradox. What survives the surcharge is a selection problem: a welfare-dominated mixed equilibrium in the anti-coordination region, so full implementation also requires a coordination device (Proposition 8). The current U.S. interagency guidance, SR 26-2, asks each banking organization to manage models both individually and in aggregate, including dependencies from common assumptions, data, and methods (Board of Governors, OCC, and FDIC 2026). It does not create the cross-bank lineage registry or systemwide joint-exposure price studied here.

Vendor competition is the natural counterforce: a second frontier vendor with independent lineage should deliver diversity for free. Section 8 carries the mechanism into the multi-vendor market. The contemporaneous common-vendor model of Keloharju and Okat (2026) studies outsourcing scale economies and the strategic-attacker incentive created by a widely

used vendor; our vendor margin instead asks whether independently sold decision models inherit correlated honest errors from shared lineage. With two frontier vendors whose cross-vendor lineage correlation is λ_{12}, the disclosure-regime level and progress results survive with κ replaced by $\kappa_x = \lambda_{12} p_H(1-p_H) - p_H \Delta_a$ if and only if λ_{12} exceeds the threshold $\Delta_a/(1-p_H)$. The threshold makes the theory falsifiable: whether cross-vendor lineage correlation clears it is a measurement question (Section 9.3), and that branch of the mechanism disappears the day a frontier-accuracy model with sufficiently independent lineage is cheap to build. The pooled-pricing baseline follows a separate route: with unpriced correlation, any vendor-side scale economy, however small, makes single-vendor monoculture the unique robust equilibrium even when lineages are fully independent; that equilibrium is inefficient only if the scale benefit does not exceed the social correlation cost. Vendor competition disciplines correlation only to a floor: with V symmetric vendors the effective exposure converges to κ_x as adoption becomes balanced, so competition dilutes the vendor-specific component but cannot compete away the industry-wide lineage, the shared pretraining corpus. Antitrust and provenance regulation are complements, not substitutes, and they address different components of κ.

The mechanism also delivers cross-sectional and event-based implications (Section 9). Point-in-time bank–technology mappings would permit three tests: whether institutions sharing model lineage comove in the tails after controlling for asset-side similarity; whether rising vendor concentration forecasts joint distress without moving individual distress; and whether adoption waves follow frontier-widening upgrades most strongly where pre-upgrade accuracy gaps were smallest. The intermediate variable is model lineage, recorded in procurement records and vendor disclosures but not yet assembled in a historical cross-bank registry.

A final observation unifies the analysis (Proposition 10). The adoption game is an exact

potential game whose potential is *welfare computed with the wrong Pigouvian coefficient*: pure-strategy Nash equilibria are the coordinatewise maxima of W_ℓ, every global maximizer of W_ℓ is a pure-strategy equilibrium, and the planner maximizes W_Λ. Every proved distortion in the finite model is therefore a comparison of the same objective at two coefficients: the level result, the quadratic scaling, and the staircase. The continuous-frontier extension uses this structure to derive local symmetry diagnostics and verified examples of configuration failure. The market is a mis-parameterized planner, and the regulatory problem is the replacement of one scalar.

2 Related Literature

The paper joins two literatures. The banking literature on correlated exposures supplies the systemic stakes and models the correlated object as an asset portfolio; the algorithmic-monoculture literature supplies the technology and measures its cost in selection quality. Here the correlated object is a prediction technology and the outcome variable is joint bank failure, and that pairing produces the pricing impossibility and the corrective instrument.

Our mechanical starting point is the diversification–diversity literature: Wagner (2011) shows that joint-liquidation risk leads investors to hold endogenously heterogeneous portfolios, partially self-insuring the system, and Wagner (2010), Acharya and Yorulmazer (2007), and Acharya and Yorulmazer (2008) study correlated asset choices by banks anticipating collective bailouts; Ibragimov, Jaffee, and Walden (2011) show how diversification by intermediaries can create diversification disasters in the aggregate. Banks in this model share a prediction *technology*, and the correlation comes from a documented accuracy–correlation frontier rather than from an assumed menu of correlated assets. The market's self-correction

channel, the heart of Wagner (2011), then internalizes exactly half of the relevant cost and is defeated entirely by the pricing friction. The fire-sale component of our externality is the classic pecuniary externality under incomplete markets (Lorenzoni 2008; Dávila and Korinek 2018); our contribution is the margin on which it operates (model choice), the pricing impossibility (spreads reprice victims, not deviators), and the regulatory instrument that follows.

A rapidly growing literature studies AI and financial stability. Dou, Goldstein, and Ji (2025) show that reinforcement-learning traders can sustain collusive strategies without communication, including through *homogenized learning biases* when algorithms share a common foundation. Dou, Goldstein, and Ji (2026) study destabilizing AI planning. The homogeneity in that agenda coordinates strategic under-trading and *raises* the adopters' profits: it is a rent-extraction channel that harms market quality. Ours is the opposite margin: correlated forecast *errors* create joint tail losses for the adopting banks themselves; there is no collusion, and the harm runs through solvency. Danielsson, Macrae, and Uthemann (2022) articulate AI homogeneity as a stability concern, and Danielsson and Uthemann (2025) provide a global-games model of AI-assisted crises; Anand et al. (2026) study AI investors and run risk. Closest on the endogenous algorithmic-adoption margin is Meng and Chen (2026), who embed exogenous signal correlation in an adoption game and obtain a monoculture transition, tail-risk amplification, and prudential-policy implications in a Kyle trading environment; Qiu and Han (2026) study representation homogeneity and synchronized trading. Relative to these papers, our adoption margin is an endogenous accuracy–correlation menu, our outcome variable is bank solvency and joint default, and our incidence result derives a lineage-contingent surcharge from the part of the credit externality that no bilateral spread charges.

Closest on the banking and common-vendor margin is Keloharju and Okat (2026). They model outsourcing concentration as making a widely used technology vendor a more valuable target for a strategic cyber attacker; banks and vendors ignore losses borne by depositors and customers, while deposit insurance weakens private discipline. We share the over-concentration question and the prescription to price marginal external cost, but the primitive and incidence differ. Their commonality changes attack incentives on outsourced infrastructure. Here honest prediction errors are correlated because frontier accuracy is bundled with model lineage, and the bilateral credit spread reaches only the exposed borrower, so no spread charges the bank creating cross-bank exposure. That difference generates the progress-induced tipping and staircase results and makes the corrective charge a function of audited lineage-specific joint-default exposure rather than vendor concentration alone.

The computer-science literature on algorithmic monoculture begins with Kleinberg and Raghavan (2021), who show that a more accurate shared ranking algorithm can lower selection quality, and who do so explicitly *without* correlated failures. That scoping choice is inherited by the subsequent matching and price-of-anarchy analyses (Peng and Garg 2024; Kleinberg, Sinanaj, and Tardos 2026), which find monoculture losses bounded by small constants. We take the complementary half of the agricultural metaphor: monoculture's defining risk is not lower average quality but simultaneous failure, and restoring the aggregate-shock margin overturns the near-optimality verdict. Bommasani et al. (2022) document outcome homogenization from shared foundations, and Kim et al. (2025) provide the cross-model error-correlation evidence on which our frontier axiom rests.

Our regulatory analysis connects to the economics of model risk regulation—Colliard (2019) on strategic selection of risk models under capital rules, and recent work on supervisory models (Leitner and Williams 2023; Parlatore and Philippon 2025)—which stud-

ies misreporting to the regulator rather than cross-bank correlation of honest errors. The interdependent-security tradition (Kunreuther and Heal 2003; Chen, Kataria, and Krishnan 2011) and the economics of cyber insurance recognize software monoculture as an aggregation risk but treat correlation as exogenous to adoption; we endogenize it. Finally, the configuration results connect to specialization in public-good games (Bramoullé and Kranton 2007): staying off the frontier is the local public good, and polarized equilibria are its specialized provision. On the empirical side, Kotidis and Schreft (2025) provide direct evidence that technology-mediated shocks propagate through the financial system via shared third-party providers.

3 Model

3.1 Environment

The economy has four ingredients: two banks, so that correlation across banks exists; debt financing, so that a market price could in principle carry the externality; a menu that bundles accuracy with lineage, which gives diversity an opportunity cost; and a convex cost of joint failure, which makes correlation matter socially. Together they isolate one margin: a bank's choice of prediction technology when accuracy is priced and correlation is not. Every pricing regime studied below differs in a single number: how much of the joint-failure cost a bank's objective registers.

There are two dates. At date 0, each of two risk-neutral banks $i \in \{1, 2\}$ (Section 5 generalizes to N) invests one unit, financed entirely by uninsured wholesale debt with face value R_i posted by a competitive risk-neutral financier; equity holders have limited liability. At date 1 the bank's loan book pays y if its decision model performed ("success") and must

be liquidated otherwise.

The model menu. Each bank adopts one prediction model $j \in \{H, L\}$. Model H (the frontier model) errs on the pivotal decision with probability p_H; model L (the legacy model) errs with probability $p_L > p_H$. Write $\Delta_a \equiv p_L - p_H > 0$ for the accuracy gap. Model H carries a loading $\lambda \in (0, 1]$ on a common *lineage factor*; model L's errors are idiosyncratic.

Assumption 1 (Accuracy–correlation bundling)**.** *The menu contains no model with failure probability p_H and zero lineage loading: frontier accuracy is available only jointly with the loading λ.*

Assumption 1 is the Kim et al. (2025) frontier: across more than 350 models, the most accurate models exhibit the most correlated errors, including across vendors, because they share training data, architecture, and distillation lineage. Escaping correlation costs first-order accuracy. Section 8 replaces this assumption with an explicit multi-vendor menu and derives exactly when it binds.

Failure technology. With probability λ a *common lineage event* occurs: both H-banks' failures are governed by the same draw $S \sim \text{Bern}(p_H)$—the shared blind spot triggered by the same input. With probability $1 - \lambda$, each bank on H fails according to an independent idiosyncratic draw $\eta_i \sim \text{Bern}(p_H)$. Banks on L always fail idiosyncratically, $\eta_i \sim \text{Bern}(p_L)$.

Lemma 1 (Joint failure probabilities)**.** *Marginal failure probabilities are p_j regardless of loadings; λ is exactly the failure correlation of two H-banks; and*

$$\pi_{HH} = p_H^2 + \lambda p_H(1 - p_H), \qquad \pi_{HL} = p_H p_L, \qquad \pi_{LL} = p_L^2.$$

The excess joint exposure *of monoculture is*

$$\kappa \;\equiv\; \pi_{HH} - \pi_{HL} \;=\; p_H\big[\lambda(1-p_H) - \Delta_a\big].$$

Assumption 2 (Frontier case)**.** $\lambda(1-p_H) > \Delta_a$*, so* $\kappa > 0$*.*

Assumption 2 holds whenever the documented correlation is large relative to the accuracy gap, the empirically relevant configuration at the frontier. It selects the case in which the frontier model raises joint failure risk even as it lowers each bank's own failure risk, so that the private and social rankings of the frontier can diverge.

Payoffs and the cost of joint failure. On success the loan book pays y; on an isolated failure it is liquidated for recovery $r < 1$; when both banks fail simultaneously, congested liquidation recovers only $r - \varphi$ per bank ($\varphi \geq 0$). We interpret φ as a net resource loss from hurried or congested liquidation, measured after any gains to outside buyers. The joint collapse also imposes an external cost $\gamma \geq 0$ borne outside the contracting pair: deposit-insurance resolution costs, which are mutualized across surviving banks through assessments, together with payment-disruption deadweight and the taxpayer backstop in tail scenarios. The total *excess* social cost of a joint failure, relative to two isolated failures, is

$$\Lambda \;\equiv\; 2\varphi + \gamma \;>\; 0.$$

The private NPV of model j under fair pricing is $v_j \equiv (1-p_j)y + p_j r - 1$, and the *private value of accuracy* is $\Delta \equiv v_H - v_L = \Delta_a\,(y-r) > 0$.

Assumption 3 (Regularity)**.** $0 < p_H < p_L < 1$*;* $0 \leq \varphi \leq r < 1 < y$*;* $y > (1+\varphi)/(1-p_L)$*; and* $v_L > \varphi p_L$*.*

The third condition guarantees $R_i < y$ under every candidate pricing (debt is risky but repayable from success cash flow); the last is participation.

Information and pricing regimes. The friction is hidden action: banks and financiers share the same knowledge of the menu and the failure technology. In the *baseline* regime, a bank's model choice (its vendor lineage) is neither observable nor contractible by financiers, who therefore price debt at their rational-expectations conjecture of equilibrium play; a deviating bank's spread does not move. In the *disclosure* regime, lineage is observable and the financier prices contingent on its own borrower's choice (and its correct conjecture of the other bank's choice). We also study a *full-liability (Wagner) benchmark* in which equity holders cover their financier's shortfalls (debt is riskless) and $\gamma = 0$, so all joint-failure costs fall inside the banking pair. Welfare is

$$W(j,k) \;=\; v_j + v_k - \Lambda\,\pi_{jk},$$

the sum of bank and financier values net of the external cost: prices are transfers, and the only deadweight beyond marginal failure risk (already in v_j) is Λ per joint failure. The planner benchmark used throughout is constrained: it assigns models from the same menu under the same failure technology, with no transfers and no additional instruments; it differs from the market only in internalizing the full coefficient Λ.

Timing and equilibrium. The sequence of events is:

$t = 0$: Financiers post financing schedules. The baseline offers one pooled term based on equilibrium conjectures; the disclosure regime offers terms contingent on audited lineage. Banks then choose models and finance, so a deviation is repriced only under disclosure.

$t = 1$: The lineage event and idiosyncratic draws realize; failures occur; assets are liquidated,

with the fire-sale discount when failures are joint; the external cost is incurred; claims are settled.

An *equilibrium* consists of model choices (j_1, j_2) and face values (R_1, R_2) such that (i) each bank's model choice maximizes its expected equity value given the other bank's choice and the pricing it faces; (ii) each financier earns zero expected profit given its information; and (iii) financiers' conjectures about model choices are correct. The regimes differ in one place: under disclosure R_i is a function of bank i's observed choice, while in the baseline it depends only on the financier's equilibrium conjecture. The model abstracts from risk aversion, deposit-insurance pricing, and asset-side portfolio choice; Section 9 develops the bailout variant.

3.2 The Internalization Coefficient

The paper's regimes are summarized by one number.

> *The internalization coefficient ℓ is the joint-failure cost a bank's objective registers per unit of excess joint exposure κ: the private adoption rule in every regime is "adopt H iff $\Delta > \ell\kappa$."*

The analysis establishes $\ell = 0$ under pooled pricing, $\ell = \varphi$ under disclosure or own full liability, and $\ell = \Lambda$ for the planner. In the Wagner benchmark $\gamma = 0$, so the same private coefficient φ equals one-half of that benchmark's $\Lambda = 2\varphi$. All welfare statements reduce to comparisons of thresholds $\ell\kappa$ against $\Lambda\kappa$.

4 Unpriced Correlation

4.1 The Efficient Benchmark

Two frictionless variants of the environment pin down what the main results are about. If the menu were unbundled (if a model with frontier accuracy and zero lineage loading existed), first best would be attainable without an accuracy sacrifice and would be selected whenever joint exposure is priced; under pooled pricing, however, the correlated and uncorrelated frontier models are payoff-equivalent, so the inefficient correlated profile can remain an equilibrium by indifference (Proposition 3(iii)). Alternatively, if banks internalized the full joint-failure cost, with the coefficient ℓ equal to Λ, adoption would follow the planner's rule: take the frontier model when the accuracy gain justifies the total excess joint exposure, $\Delta > \Lambda\kappa$; keep one bank apart otherwise; and never retreat from the frontier entirely. This is the standard market-discipline intuition made precise: fully priced risk is chosen efficiently. Everything that follows measures how far each contracting regime falls short of this benchmark and attributes the shortfall to a single object, the gap between ℓ and Λ. The sequence of results below is thus a controlled experiment on the internalization coefficient: the environment, the menu, and the planner never change; only who pays for correlation does.

4.2 Baseline: Monoculture for Any Accuracy Edge

Under pooled pricing at conjectured face value $R_i < y$, limited-liability equity collects only in its own success state, so bank i's payoff from model j is $(1 - p_j)(y - R_i)$, independent of the rival's model and of λ. The gain from adopting H is $\Delta_a(y - R_i) > 0$ for every $R_i < y$.

Proposition 1 (Unpriced correlation implies monoculture)**.** *In the baseline regime, H is strictly dominant and monoculture HH is the unique equilibrium (in pure or mixed strategies) for every $\Delta > 0$. The planner chooses diversity HL if and only if $\Delta < \Lambda\kappa$, and never chooses LL. Hence for every $\Delta \in (0, \Lambda\kappa)$ the market is in monoculture while diversity is efficient, with welfare loss*

$$\Omega \;=\; W_{HL} - W_{HH} \;=\; \Lambda\kappa - \Delta \;>\; 0.$$

The private adoption threshold is 0 *against the social threshold $\Lambda\kappa$: banks internalize a zero share of the joint-failure cost, and under-diversification obtains on the* entire *region where diversity is efficient.*

Three features of the result matter for what follows. First, the inefficiency arises over a range of parameters: the region $(0, \Lambda\kappa)$ is open and, under Assumption 2, nonempty precisely when accuracy gaps are small, frontier correlation is high, and systemic convexity is large—the configuration the evidence describes. Second, the planner's problem is two-dimensional: because $\pi_{HL} < \pi_{LL}$ (the accurate bank fails less often *with everyone*), moving one bank to the frontier raises accuracy *and* lowers joint exposure; full retreat from the frontier is never optimal. One diverse bank suffices—diversity in this economy means *separation*, not abstinence. Third, the welfare loss $\Lambda\kappa - \Delta$ is unbounded in the systemic-cost parameter while the equilibrium is unresponsive to it: under pooled pricing, no magnitude of social cost, however large, changes any bank's behavior.

The wedge decomposes exactly (Appendix, proof of Proposition 1) into three uninternalized terms:

$$\underbrace{\big[\varphi\kappa - \Delta_a(R^{HH} - r)\big]}_{\text{own financier}} + \underbrace{\varphi\kappa}_{\text{rival's financier}} + \underbrace{\gamma\kappa}_{\text{outside parties}}, \tag{1}$$

a classification that dictates the effect of every intervention below: observability transfers

exactly the first bracket into the bank's objective and no more.

4.3 Disclosure: Pricing Reprices the Victim

Suppose lineage is fully observable: each financier prices its own borrower's choice, $R_{j|k} = (1 - p_j r + \varphi\pi_{jk})/(1 - p_j)$, giving bank payoffs $u(j,k) = v_j - \varphi\pi_{jk}$. Fair pricing transmits to the bank exactly its *own* financier's expected fire-sale loss, and no more.

Proposition 2 (Disclosure is not enough)**.** *In the disclosure regime: (i) LL is never an equilibrium; (ii) if $\Delta > \varphi\kappa$ the unique equilibrium is HH; if $\Delta < \varphi\kappa$ the pure equilibria are $\{HL, LH\}$ (a symmetric mixed equilibrium exists and is welfare-dominated). (iii) For every $\Delta \in (\varphi\kappa,\ \Lambda\kappa)$ the equilibrium is HH while the planner prefers HL: the residual wedge per adoption is*

$$(\Lambda - \varphi)\kappa \ = \ (\varphi + \gamma)\kappa \ > \ 0,$$

the fire-sale harm to the rival's financier plus the resolution and disruption costs borne outside the pair. Contingent fair pricing restores efficiency only if $\varphi + \gamma = 0$—precisely when there is no systemic problem to price.

When $\varphi > 0$, disclosure activates a self-correction channel: for small accuracy edges the market itself delivers diversity, because joining the monoculture now carries a price $\varphi\kappa$. At the allowed boundary $\varphi = 0$, disclosure changes information but not adoption payoffs, so this channel is absent. When active, the channel remains structurally incomplete. If bank i adopts the frontier model, the probability that *the rival* fails simultaneously rises; the injured parties are the rival's creditors and the outside institutions that mutualize resolution losses. Bank i's own spread cannot price this harm under any information regime, because a bilateral contract charges its own borrower: *pricing reprices the victim, not the deviator.*

The impossibility is architectural, and it survives full transparency. This distinguishes the friction from standard hidden-action problems, where observability restores efficiency.

4.4 The Wagner Benchmark and the Internalization Ranking

Does the market solve the problem if banks bear *all* the costs? Set $\gamma = 0$ (all joint-failure costs fall inside the pair) and give shareholders unlimited liability; debt becomes riskless and observability is moot. Bank payoffs are $v_j - \varphi\pi_{jk}$: the Wagner (2011) channel operates, and for $\Delta < \varphi\kappa$ banks endogenously choose heterogeneity.

Proposition 3 (Internalization ranking)**.** *(i) In the full-liability benchmark, strict under-diversification survives on* $\Delta \in \big((\Lambda/2)\kappa,\ \Lambda\kappa\big)$*: banks internalize exactly half the joint-failure cost. When one bank departs the monoculture, the fall* κ *in joint-failure probability saves* both *banks* $\varphi\kappa$ *each, but the mover compares only its own saving with its own accuracy loss. Diversity is a reciprocal public good within the pair. (ii) In the baseline economy, pooled pricing has* $\ell = 0$*, disclosure and own full liability both have* $\ell = \varphi$*, and the planner has* $\ell = \Lambda = 2\varphi + \gamma$*. In the Wagner benchmark* $\gamma = 0$*, so* $\ell = \varphi = \Lambda/2$*. The private and planner adoption thresholds coincide for every admissible accuracy gain if and only if* $\ell = \Lambda$*; for a fixed* Δ*, their selected allocations can coincide outside the wedge region even when* $\ell < \Lambda$*. (iii) Neither limited liability nor hidden lineage is individually necessary for strict under-diversification: the irreducible source is the cross-bank joint-exposure externality, which survives full observability and full internal cost-bearing. If an uncorrelated model with failure probability* p_H *exists, first best is attainable without sacrificing accuracy and is selected when lineage is observable and* $\varphi > 0$*, or when banks break payoff ties in favor of lower joint exposure. When the relevant private coefficient is zero—under pooled pricing, or under disclosure at the boundary* $\varphi = 0$*—the correlated and uncorrelated frontier models*

have identical private payoffs, so inefficient correlated equilibria can survive by indifference. Thus bundling is necessary for the strict *under-diversification result, not for every set-valued inefficiency. Ex-ante contracting between the banks on model choice (a Coasean merger of the pair) restores efficiency when* $\gamma = 0$.

Part (iii) locates the mechanism. The pricing friction ($\ell = 0$) suffices for the extreme outcome, monoculture at any accuracy edge, but the deeper problem is that no bilateral liability rule or disclosure regime assigns to a bank the protection its diversity provides to others. The bundling assumption is what makes the exposure unavoidable: it is the empirical content of the frontier, and the vendor analysis of Section 8 derives it as an explicit theorem about cross-vendor lineage.

4.5 Spot-Contract Invariance: Coefficients Are Coalition Stakes

Propositions 2 and 3 used specific arrangements: competitively priced simple debt, and a liability rule. Richer spot financing is the natural candidate for doing better, through covenants conditioning on the rival's failure, warrants, insurance-like payoffs, or bonds posted against the joint event. The next result settles the question for the whole class of spot arrangements: financing in which every observable profile can be competitively refinanced. Exclusive contracts signed before adoption can commit to cross-profile transfers and require an explicit contracting game; they lie outside its scope.

Proposition 4 (Spot-contract invariance)**.** *Fix the disclosure environment. Consider any financing arrangement in which (i) contracts are* bilateral *(every payment involves bank i and its own financiers only) but may condition on the entire observable profile and on any realized events; and (ii) every profile can be competitively refinanced, so financiers break*

even profile by profile. Then: (i) every such arrangement induces adoption payoffs equal, up to profile-independent constants, to $v_{j_i} - \varphi\pi_{j_i j_k}$: the equilibrium correspondence, the residual wedge $(\varphi+\gamma)\kappa$, and the internalization coefficient $\ell = \varphi$ are invariant to the state-contingent spot contract space; (ii) enlarging the contracting coalition raises ℓ to the coalition's stake in the joint-failure event: an arrangement among both banks and both financiers attains $\ell = 2\varphi$, efficient if and only if $\gamma = 0$; attaining Λ when $\gamma > 0$ requires including or pricing the outside bearers of γ; (iii) hence, when $\varphi > 0$ and $\gamma > 0$, the ladder $0 < \varphi < 2\varphi < \Lambda$ is spanned by pooled pricing, competitively refinanced bilateral finance, two-pair multilateral contracts, and the systemwide planner.

For a bank and its creditors, the result says that no covenant, warrant, or joint-failure bond changes what the bank is charged for correlation: the spot menu is rich, and the coefficient it delivers is always φ. Two observations produce this. First, risk neutrality collapses every within-profile state-contingent scheme to its profile-contingent *expected* transfer. Second, profile-by-profile competition transfers the contracting pair's joint surplus to the bank, and that surplus at profile (j, k) is $v_j - \varphi\pi_{jk}$ plus constants: the pair's entire stake in a joint failure is the φ its own financier loses. What raises the coefficient is enlarging the coalition. Cross-pair payments (bank k compensating bank i for staying diverse) are exactly what bilaterality excludes and what a two-pair multilateral arrangement (a merger, a mutual liability pool, or a clearinghouse rule) supplies, reaching 2φ. The final increment γ lies outside the modeled two-pair coalition: in practice, deposit-insurance costs assessed on surviving banks and payment-disruption deadweight. Liability rules reallocate the stake *within* the pair without enlarging it, which is why the full-liability benchmark and bilateral spot contracting sit at the same rung when $\gamma = 0$.

5 Many Banks: The Price of Unpriced Correlation

The system-level stakes appear at N banks. Let m denote the number of banks on the frontier model ("the herd"). The lineage event is common to the herd: with probability λ all H-banks fail by the same draw, with probability $1-\lambda$ independently; pairwise joint probabilities are exactly those of Lemma 1. When M banks fail simultaneously, each failing bank's recovery is $r-\varphi(M-1)$ and each failing pair imposes γ: the excess social cost is $\Lambda\binom{M}{2}$, a convexity that charges for every *pair* of simultaneous failures. (Only pairwise joint probabilities enter expected costs, so results are invariant to higher-order details of the lineage construction.) Regularity generalizes as $\varphi(N-1) \leq r$ and $y > (1+\varphi(N-1))/(1-p_L)$.

Everything runs through the marginal generalization of κ:

$$X(m) \;\equiv\; (m-1)\,\kappa \;-\; (N-m)\,p_L\Delta_a,$$

the excess joint exposure created by the m-th adoption: the entrant forms $m-1$ new frontier pairs (excess exposure κ each) and dissolves $N-m$ legacy partnerships whose joint-failure probability falls by $p_L\Delta_a$ each. X is strictly increasing in m: *each entrant endangers every incumbent, so the externality compounds with herd size.* $X(1) < 0$: the first frontier adoption is always socially valuable. For boundary inequalities below, set $X(0) = -\infty$ and $X(N+1) = +\infty$.

Proposition 5 (Threshold equilibrium and under-diversification at N)**.** *(i) For $\ell > 0$, the pure-strategy equilibria are exactly the profiles whose herd size m satisfies $\ell X(m) \leq \Delta \leq \ell X(m+1)$; the equilibrium herd size m^* is generically unique (identities of herd members are indeterminate). Under pooled pricing, $m^* = N$: monoculture regardless of N. (ii) Welfare $W(m) = mv_H + (N-m)v_L - \Lambda Q(m)$ is discretely strictly concave with increments*

$\Delta - \Lambda X(m)$*; the planner's herd is* $m^{**} = \max\{m : \Delta \geq \Lambda X(m)\} \geq 1$. *(iii) For* $0 < \ell < \Lambda$, $m^* \geq m^{**}$*, and the welfare loss is the exact sum* $\sum_{m=m^{**}+1}^{m^*}[\Lambda X(m) - \Delta]$*, with per-adoption wedge* $(\Lambda - \ell)X(m)$ *growing linearly in the herd. At* $\ell = \Lambda$*, every pure-strategy equilibrium herd size is planner-optimal; under pooled pricing the same ordering and loss sum apply with* $m^* = N$.

For $N = 2$, $X(1) = -p_L\Delta_a$ and $X(2) = \kappa$. Thus, for $\ell > 0$, Proposition 5 reduces exactly to the two-bank allocation thresholds: generically, $m = 1$ if $\Delta < \ell\kappa$ and $m = 2$ if $\Delta > \ell\kappa$; under pooled pricing, $m = 2$. The many-bank environment introduces endogenous herd size and scale, not a different mechanism.

Proposition 5 carries the two-bank logic to N banks, and the threshold structure clarifies who fails to pay for what. The equilibrium is a congestion game on the herd: the double inequality pins the herd's *size* while leaving its membership indeterminate, since any bank is willing to be the diverse one provided enough others are not. The marginal entrant weighs the same private prize Δ against ℓ times a statistic that now grows with the crowd, because its lineage exposure multiplies across every incumbent. Under pooled pricing this growth is irrelevant: the coefficient is zero, and the herd absorbs the entire system at any accuracy edge, however large N or Λ. The scale of the system changes the social stakes without changing a single private decision; the next result measures what that gap costs.

Corollary 1 (Quadratic versus bounded losses)**.** *(i) Under pooled pricing, let* $K = N - m^{**}$ *be the number of excess adoptions. The welfare loss is* $\Lambda(\kappa + p_L\Delta_a)K^2/2 + O(K)$. *(ii) Holding the primitives* $(\Delta, \ell, \Lambda, \kappa, p_L\Delta_a)$ *fixed, any partial internalization* $0 < \ell < \Lambda$ *bounds the number of excess adoptions by* $\Delta(1/\ell - 1/\Lambda)/(\kappa + p_L\Delta_a) + 1$*, independently of* N*, and therefore bounds the welfare loss; at* $\ell = \Lambda$ *every pure-strategy equilibrium is efficient. When both solutions of the continuous relaxation are interior, its loss equals* $\Delta^2(\Lambda - \ell)^2/\big[2(\kappa +$

$p_L\Delta_a)\Lambda\ell^2\big]$.

Corollary 1 attaches magnitudes to the two policy instruments. In a two-bank economy, disclosure looked merely insufficient (Proposition 2). In a many-bank system with $\varphi > 0$ the statement splits: *mandatory lineage disclosure is what stands between a quadratic loss in excess adoption and a bounded residual; the corrective surcharge then removes the residual from pure-strategy allocations.* In a fixed-parameter instance with $p_H = 5.8\%$, $p_L = 6\%$, $\lambda = 0.9$, $y = 1.6$, and recovery $r = 0.765$ (the bank-level FDIC severity anchor), the pooled-pricing loss grows from 0.153 to 9.872 (in units of one bank's investment) as N goes from 4 to 28, while the disclosure-regime loss stays between 0.023 and 0.0245, with one excess adopter throughout. Figure 1 plots the two loss paths. The comparison holds the pairwise cost coefficients fixed over the admissible range $\varphi(N-1) \leq r$; market-depth scalings in which φ or Λ shrink with N slow both loss paths.

6 The Progress Paradox

How does the equilibrium respond to progress on the frontier? Parameterize gap-widening progress by δ: $p_H(\delta) = p_H^0 - \delta$ with p_L fixed, the empirically documented pattern of the frontier pulling away. (The tipping result below is specific to gap-widening progress: proportional improvement of both models never tips the equilibrium, and catch-up progress by the legacy model pushes *toward* diversity.) Progress moves both blades of the scissors: the private prize $\Delta(\delta)$ rises, and the priced deterrent $\ell\kappa(\delta)$ falls. While Assumption 2 holds,

$$\frac{d\kappa}{d\delta} = -\big[\lambda(1-p_H) - \Delta_a + p_H(1-\lambda)\big] < 0,$$

so no additional derivative condition is required.

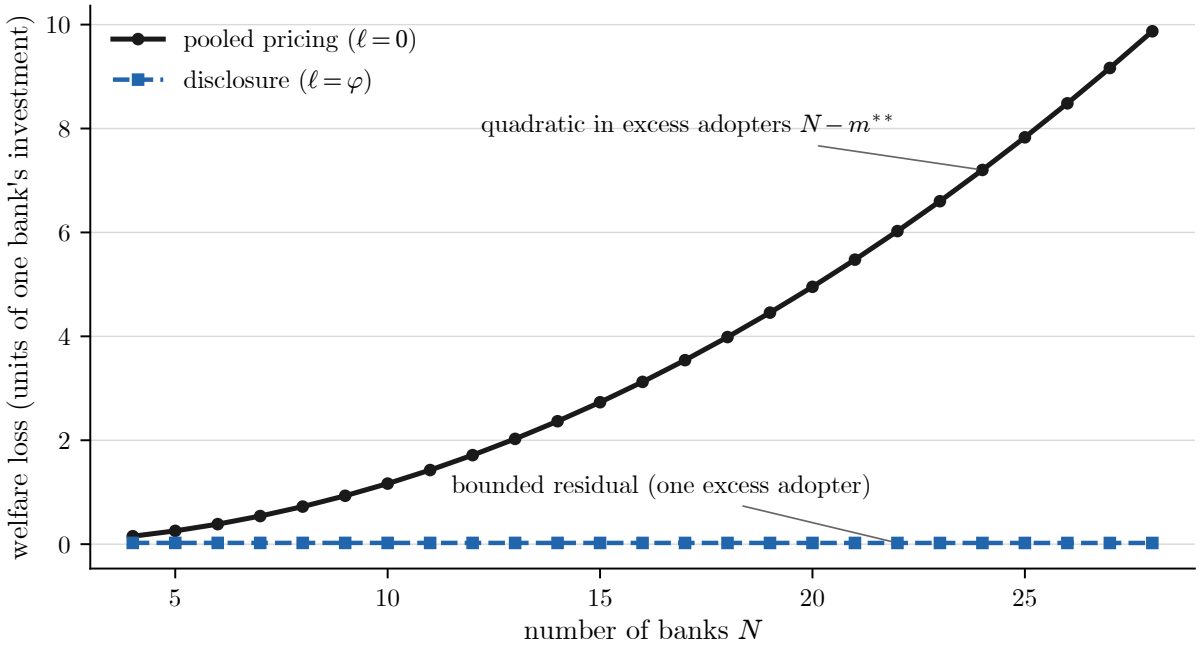

Figure 1: **Quadratic versus bounded losses (Corollary 1).** Welfare loss relative to the constrained optimum, in units of one bank's investment, as the number of banks grows over the admissible range, in the fixed-parameter instance of Section 5 ($p_H = 5.8\%$, $p_L = 6\%$, $\lambda = 0.9$, $y = 1.6$, $r = 0.765$, $\varphi = 0.018$, $\gamma = 0.5$). Under pooled pricing (solid), every bank joins the herd and the loss is quadratic in the number of excess adopters. Under disclosure (dashed), the priced deterrent leaves one excess adoption throughout this range. The gap between the curves is what lineage disclosure is worth here; the dashed line's distance from zero is what the correlation surcharge removes from pure-strategy allocations.

Alt text: Line chart of welfare loss against the number of banks from 4 to 28. The pooled-pricing loss rises convexly from about 0.15 to 9.87, while the disclosure loss remains approximately flat near 0.024.

Proposition 6 (The tipping paradox)**.** *Consider a regime with partial internalization $\ell \in (0, \Lambda)$, initial diversity ($\Delta^0 < \ell\kappa^0$), and the strict pure equilibria selected in the anti-coordination region. Then the tipping statistic $\Delta(\delta) - \ell\kappa(\delta)$ is strictly increasing, there is a unique progress size δ^* (the root of an explicit quadratic) at which the equilibrium jumps from diversity to monoculture, and: (i) welfare is strictly increasing in δ on either side of δ^* but drops discontinuously at δ^* by*

$$(\Lambda - \ell)\,\kappa(\delta^*) \;=\; \Big(\frac{\Lambda}{\ell} - 1\Big)\Delta(\delta^*) \;>\; 0;$$

(ii) the planner still strictly prefers diversity at δ^*: the market tips strictly before monoculture becomes efficient, and it tips because of *progress; (iii) welfare remains below its pre-tipping level on a nonempty open interval (δ^*, δ^{**}); the length of this recovery interval is parameter-dependent (it is several times the triggering improvement in the numerical instance and sweeps reported below); (iv) the paradox is an equilibrium phenomenon: the planner's welfare is continuous and strictly increasing in δ (it switches regimes exactly at indifference), and the drop $(\Lambda - \ell)\kappa(\delta^*)$ vanishes as $\ell \to \Lambda$.*

The incidence of the paradox across regimes carries its policy content. Under pooled pricing there is no tipping point: the economy is in monoculture for every δ, and progress raises welfare throughout, because welfare already sits on the lower branch. When $\varphi > 0$, the paradox can instead live in the *disclosure* regime: partial internalization sustains diversity that progress can destroy. At the boundary $\varphi = 0$, disclosure leaves adoption incentives identical to pooled pricing and cannot generate the tipping path. With positive φ, mandatory lineage disclosure without the accompanying surcharge can convert an economy where progress is always good but the level is always bad into one where the level is better

but progress is locally perverse. Disclosure raises welfare pointwise at every δ, but it can make the *trajectory* non-monotone. Completing the price aligns the pure-strategy switch with planner indifference and removes its welfare jump (Section 7). Figure 2 illustrates the mechanism in the verified instance.

The Monte Carlo sweeps use a parameter box disciplined by resolution data: legacy failure risks of one to eight percent, read as multi-year or stress-conditional; recoveries of 72–91.4%, the resolution-severity range; and fire-sale increments of two to fifteen percent (the Internet Appendix reports the design; Section 9.3 explains which inputs are anchored and which remain open). Across that box, the median tipping point has $p_H^*/p_L = 0.92$, and the median welfare drop is 2.1% of first-best welfare, with a ninetieth percentile of 5.9%. Scaled by banking assets instead of by first-best welfare, the same drops have a median of 0.6% and a maximum of 3.1%. A discrete 5–10% accuracy improvement is welfare-reducing on 4.5–7.0% of the anchored box and, conditional on harm, destroys a median 4.7–10.8 times the welfare gain the same improvement would deliver under efficient adoption. The harmful region is an intermediate band: at a 50% improvement its incidence falls to 0.3% and the median loss-to-gain ratio to 0.2. The paradox concerns *marginal* frontier progress. In applications, the progress path corresponds to frontier model releases (discrete, well-dated events in which the leading vendor's accuracy advantage widens), so the paradox attaches to the adoption waves that follow such releases (Prediction 3), and its incidence is highest where pre-release accuracy gaps were smallest.

With many banks the single cliff becomes a staircase.

Proposition 7 (The staircase)**.** *With N banks in a partial-internalization regime, gap-widening progress triggers a strictly increasing sequence of thresholds $\delta_1 < \delta_2 < \cdots$ at which the equilibrium herd grows by one bank; welfare rises strictly between triggers and drops by*

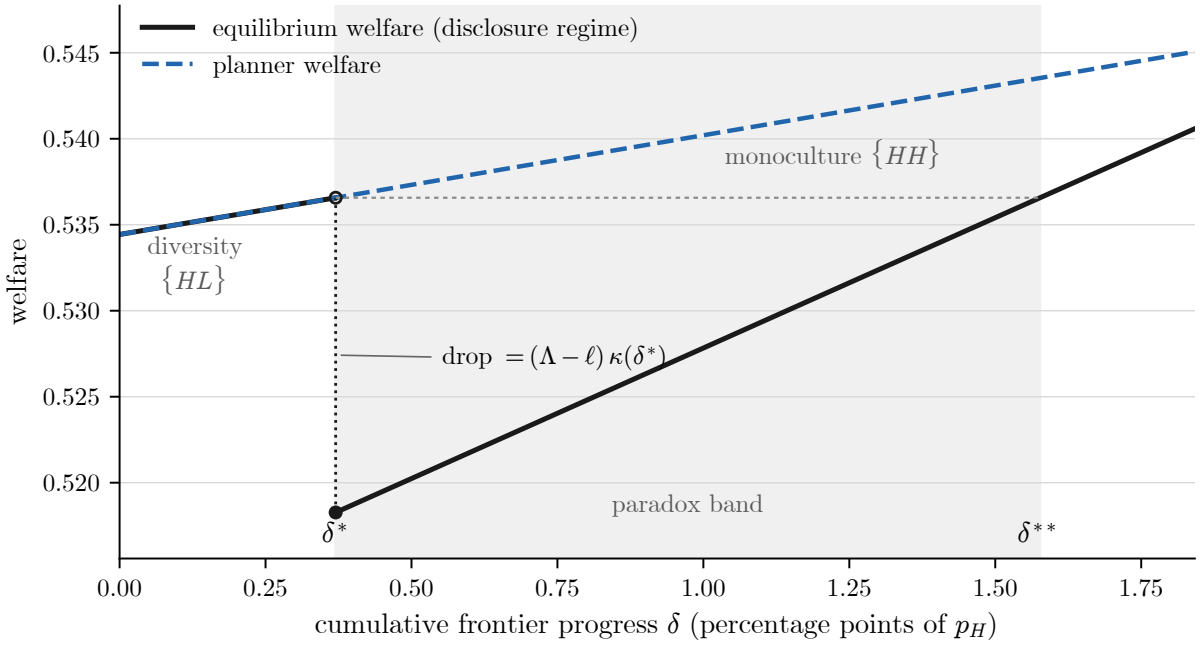

Figure 2: **The progress paradox (Proposition 6).** Equilibrium welfare in the disclosure regime (solid) and planner welfare (dashed) along gap-widening frontier progress, measured on the horizontal axis in percentage points of p_H, in the verified anchored instance ($p_H^0 = 5.8\%$, $p_L = 6\%$, $\lambda = 0.6$, $y = 1.3$, $r = 0.765$, $\varphi = 0.10$, $\gamma = 0.50$). Welfare rises within each regime; at δ^* a 0.37 percentage-point (6.4%) reduction in p_H tips the market from diversity into monoculture and welfare drops by $(\Lambda - \ell)\kappa(\delta^*)$, roughly 3.4% of its pre-tip level. The shaded band is the region in which welfare remains below that level: recovery requires cumulative progress δ^{**} about 4.3 times the improvement that caused the damage. The planner's path coincides with the equilibrium path before δ^* and is continuous and strictly increasing throughout: its tipping point lies to the right of the plotted range, and its switch occurs at indifference. The entire drop is therefore an equilibrium phenomenon.

Alt text: Line chart of equilibrium and planner welfare against cumulative frontier progress in percentage points. The two paths coincide until the tipping point δ^*, where equilibrium welfare jumps downward and remains below its pre-tip level through δ^{**}, while planner welfare increases smoothly throughout.

$(\Lambda/\ell - 1)\,\Delta(\delta_m)$ *at each. The drops* grow *along the staircase, and their sum is the total discrete adoption loss; the corresponding continuum model has the same marginal wedge. The planner's staircase has zero drops.*

The two-bank discontinuity is thus the discrete-menu image of a steep negative welfare gradient: with many banks the loss disaggregates into per-upgrade drops that *grow* as the herd swells, because each successive upgrade tips a marginal bank into an ever-larger crowd.

7 Policy: Disclosure, Then a Price

The analysis delivers a two-instrument prescription in which sequencing is essential.

Proposition 8 (The correlation surcharge and its precondition)**.** *(i) An ex-ante surcharge of* $s = \Lambda - \ell$ *per unit of joint-failure exposure, conditioned on audited model lineage, aligns every pure-strategy adoption threshold with the planner's in the finite* H/L *model. Consequently, every pure-strategy equilibrium of that model is planner-optimal, and the pure-strategy underdiversification and welfare jumps in Propositions 1–7 disappear. The surcharge implements the planner in pure strategies only: in the two-bank anti-coordination region* $\Delta < \Lambda\kappa$*, the two efficient asymmetric pure equilibria coexist with a welfare-dominated symmetric mixed equilibrium, and ruling out that mixed equilibrium requires a coordination or equilibrium-selection device. (ii) The instrument is implementable only if lineage is observable: under hidden model choice an ex-ante charge cannot condition on the choice, and an ex-post charge on a failed bank is uncollectible under limited liability. (iii) Disclosure alone changes* ℓ *from* 0 *to* φ*; the change is strict when* $\varphi > 0$ *but remains insufficient (Proposition 2). Disclosure is the precondition for, not a substitute for, the surcharge.*

Three remarks map the instrument into the existing supervisory architecture. First,

the object to be disclosed and audited is *lineage*—training-data provenance, base-model identity, distillation ancestry—not model internals: the statistic the regulator needs is which banks share a failure factor, i.e., the inputs to κ. The instrument accordingly inherits the quality of the audit: misreported provenance recreates the hidden-choice friction inside the disclosure regime, which makes lineage *verification* the natural object of supervisory technology. Second, one implementation is a capital add-on increasing in audited joint exposure (in the N-bank model, a charge per unit of $X(m)$). The model derives the required Pigouvian wedge; because capital choice is outside the model, converting that wedge into risk-weighted capital is an implementation question. SR 26-2 already tells each banking organization to assess aggregate dependencies among its own models and to validate vendor products (Board of Governors, OCC, and FDIC 2026); the missing object here is cross-bank exposure to shared lineage, which no single institution can measure or price from its own inventory. Third, the sequencing matters quantitatively: disclosure converts the quadratic pooled-pricing loss into a bounded residual (Corollary 1), the surcharge removes the residual, and disclosure without the surcharge can create the progress paradox (Proposition 6). A coordination rule completes the package when the regulator requires implementation in every equilibrium, mixed as well as pure.

8 Vendor Competition: How Much Diversity Can the Market Buy?

Vendor entry is the market's own candidate remedy for correlation: if a second frontier vendor offers equal accuracy with independent lineage, banks can have diversity for free, and the planner's dilemma dissolves. This section formalizes that channel and identifies the

conditions under which it delivers diversity.

Extend the menu to $\{H_1, H_2, L\}$: two frontier vendors with equal accuracy p_H and a *nested lineage* structure. With probability λ_{12} an industry-wide lineage event hits all frontier banks (shared pretraining corpora, common architectures, cross-vendor distillation); with probability $\lambda - \lambda_{12}$ a vendor-level event hits each vendor separately; with probability $1 - \lambda$ errors are idiosyncratic. Pairwise, banks on the same vendor have failure correlation λ; banks on different frontier vendors, $\lambda_{12} \in [0, \lambda]$. Define the within- and cross-vendor excess exposures $\kappa_w = \lambda\hat{c} - p_H\Delta_a$ and $\kappa_x = \lambda_{12}\hat{c} - p_H\Delta_a$, where $\hat{c} = p_H(1 - p_H)$.

Proposition 9 (Vendor competition: substitution, threshold, and immunity)**.** *(i)* (Substitution.) *Suppose $\varphi > 0$. In the two-bank disclosure regime with $\lambda_{12} < \lambda$, the banks never share a vendor in equilibrium; vendor anti-coordination replaces tier anti-coordination, and the remaining margin, frontier versus legacy, is governed by the same thresholds with κ_x in place of κ: the market splits across vendors iff $\Delta > \varphi\kappa_x$, the planner iff $\Delta > \Lambda\kappa_x$. The disclosure-regime under-diversification and progress results survive with $\kappa \to \kappa_x$ if and only if*

$$\kappa_x > 0 \iff \lambda_{12} > \frac{\Delta_a}{1 - p_H},$$

i.e., cross-vendor lineage correlation exceeds (approximately) the accuracy gap. Below the threshold, vendor diversity is free and first best obtains. At the allowed boundary $\varphi = 0$, switching vendors is payoff-neutral: a lower-joint-exposure tie-break selects the cross-vendor split, but same-vendor frontier profiles also remain equilibria. The selected split is first-best only when $\kappa_x \leq 0$ or $\Delta \geq \Lambda\kappa_x$. (ii) (Immunity of the baseline.) *Under pooled pricing, banks are vendor-indifferent, and any vendor-side scale economy $\varepsilon > 0$ per same-vendor peer makes single-vendor monoculture the unique robust equilibrium* even at $\lambda_{12} = 0$. *The pooled-pricing*

results require no assumption on cross-vendor correlation. (iii) (Dilution to the floor.) *With V symmetric vendors and a balanced herd for which $m-1$ is divisible by V, the marginal entrant's statistic is $X_V(m) = (m-1)\kappa_{\text{eff}}(V) - (N-m)p_L\Delta_a$ with*

$$\kappa_{\text{eff}}(V) \;=\; \kappa_x + \frac{\kappa_w - \kappa_x}{V} \;\xrightarrow[V\to\infty]{}\; \kappa_x :$$

vendor competition eliminates the vendor-specific component of systemic exposure and cannot dilute the industry-wide lineage floor. For an indivisible herd, replace $(m-1)/V$ by $\lfloor (m-1)/V \rfloor$ in the same-vendor pair count; the rounding error is bounded and the limit is unchanged.

Part (i) turns the paper's load-bearing assumption into a falsifiable threshold: the cross-vendor branch of the mechanism operates when lineage correlation across vendors outruns the accuracy gap. Frontier accuracy gaps can be small while cross-provider co-errors are substantial on benchmarks, the configuration in which the threshold binds; whether it binds for bank decision models is the measurement question of Section 9.3. The threshold also bounds the mechanism's scope in time: it stops applying the day a frontier-accuracy model with independent lineage becomes cheap to build. Part (i) also shows what market-supplied diversity delivers: in the region $(\varphi\kappa_x, \Lambda\kappa_x)$ the market's equilibrium is the *vendor split*, which looks like diversity, while the planner prefers keeping a bank on the legacy model, because both frontier vendors share the industry lineage. The market buys the cheap diversity and skips the deep one.

Part (ii) establishes immunity of the baseline: monoculture has two routes, technological correlation or unpriced correlation plus any scale economy, and the second requires nothing of λ_{12}. Part (iii) delivers the policy frontier: antitrust and procurement-splitting address

$\kappa_w - \kappa_x$, the vendor-specific component; lineage disclosure and the correlation surcharge address κ_x, the shared-corpus floor that no amount of vendor competition can reach. The two instrument families are complements operating on different components of the same statistic.

9 Empirical Content and Extensions

This section has two roles. Sections 9.1–9.2 extend the theory: a continuous frontier makes the two-model menu an endogenous equilibrium configuration, and anticipated collective rescues add a correlation subsidy to the pricing friction. Sections 9.3–9.4 connect the mechanism to measurement. Public sources establish institutional scope and anchor observable quantities; testing the mechanism itself requires point-in-time bank–technology mappings, so the section concludes with model-implied predictions and the variation that would identify them.

9.1 The Market as a Mis-Parameterized Planner

A single observation organizes every result and extends the model to a continuous frontier.

Proposition 10 (Potential structure)**.** *Let banks choose loadings $\rho_i \in [0, \bar{\rho}]$ on the lineage factor along a frontier $p(\rho)$ with $p' < 0$, with pairwise failure covariance $c\rho_i\rho_k$, and regime payoffs $u_i = v(\rho_i) - \ell \sum_{k \neq i} \pi_{ik}$. The adoption game is an exact potential game with potential $W_\ell = \sum_i v(\rho_i) - \ell \sum_{i<k} \pi_{ik}$: a pure-strategy profile is a Nash equilibrium in pure strategies if and only if it is a coordinatewise maximizer of W_ℓ, every interior pure-strategy equilibrium is a critical point, and every global potential maximizer is a pure-strategy equilibrium.*

The market uses the right potential with the wrong number. Beyond unification, the

formulation yields one analytic structural result and a set of disciplined numerical examples (Internet Appendix). With a linear frontier, payoffs are multilinear in own loadings; away from indifference, equilibrium best responses are at the endpoints, and a vertex equilibrium and vertex optimum always exist. The discrete H/L menu of Section 3 is therefore a generic configuration, and its thresholds map exactly onto $X(m)$. With curvature, the ratio of own concavity to pairwise coupling gives a local symmetry test. Verified examples display smooth over-loading, joint polarization, and *configuration failure*, in which the potential-selected market equilibrium is symmetric while the planner prefers a few specialists deep at the frontier and a diverse periphery: the market can get the *configuration* of diversity wrong, not only its amount.

9.2 Bailout Expectations Subsidize Correlation

Adding a Farhi and Tirole (2012) bailout—equity holders receive $B > 0$ in the joint-failure state only (rescues are collective)—makes the private pull toward the frontier model strictly increasing in $B\kappa$: collective moral hazard subsidizes correlation, herding banks into the state in which they are rescued. The internalization coefficient becomes negative before it becomes zero; all formulas extend with ℓ replaced by $\ell - B$-type terms. The too-many-to-fail literature (Acharya and Yorulmazer 2007) obtains correlated choices from anticipated collective rescues of asset portfolios; here the correlated object is the prediction technology, and the corrective instrument conditions on model lineage.

9.3 Measurement and Descriptive Evidence

The available evidence disciplines the scope and quantitative interpretation of the theory. We report three layers: how concentrated bank decision technologies already are; which

Table 1: Vendor concentration in bank decision technologies

Decision domain	Concentration or common layer	Evidence
Core banking systems	Fiserv, Jack Henry, and FIS served roughly 42%, 21%, and 9% of surveyed U.S. banks; the top three exceeded 70% combined	Kansas City Fed (Alcazar et al. 2024)
Credit scoring	FICO scores are used by roughly 90% of top U.S. lenders; GSE loan delivery required Classic FICO for decades, before FHFA opened loan-level model choice in 2026	Vendor-reported share; FHFA policy (FICO 2026; FHFA 2026)
Card fraud detection	FICO reported that Falcon screened about 85% of U.S. card transactions and served 17 of the 20 largest global issuers	Vendor statements, 2006 and 2011 (FICO 2006, 2011)
Mortgage underwriting	Fannie Mae's DU and Freddie Mac's LPA accounted for more than 90% of mortgages purchased by the Enterprises	FHFA OIG (FHFA OIG 2019)
AI/ML models, system-wide	75% of U.K. financial firms use AI; third-party implementations rose from 17% to 33% of use cases; the top three providers account for 73% of cloud, 44% of models, and 33% of data	Bank of England and FCA (Bank of England and FCA 2024)
Technology service providers	Interagency supervision explicitly treats large shared service providers as systemically relevant; one provider cyberattack propagated through dependent banks	FFIEC (FFIEC 2011); Kotidis and Schreft (2025)

Units and vintages are stated in each row. The FICO credit-score and Falcon shares are vendor-reported; the remaining quantities come from official or peer-reviewed sources. Source files and coding details accompany the replication package.

resolution data anchor the failure and recovery parameters; and what the event evidence can and cannot identify about the correlation parameters.

Concentration. Table 1 collects concentration facts across the decision domains the model describes. Two readings matter for the theory. First, each high-stakes domain exhibits either a dominant common decision layer or a concentrated supplier structure close to the model's monoculture configurations. The newest layer, third-party AI models, is concentrating fastest: the third-party share of use cases almost doubled in two years, and the

top three model providers already account for 44% of models in the Bank of England and FCA census. Second, credit scoring provides a regime experiment aligned with Section 8: a regulator-set eligibility rule produced a single-model culture in conforming-mortgage delivery for decades, and the 2026 FHFA decision to admit model choice is a vendor-entry policy experiment whose consequences for correlated underwriting errors the model's comparative statics speak to directly.

Parameter anchors. Resolution data anchor the failure-technology parameters. Author calculations from FDIC BankFind give pooled annual failure incidence of 0.091% in the quiet 2014–2022 period, 1.04% across the 2008–2013 resolution wave (peaking at 1.96% in 2010), and 0.339% pooled over 2002–2025 (FDIC 2026). The model's failure probabilities of one to eight percent are therefore to be read as multi-year or stress-conditional risks, not one-year unconditional rates. Resolution severity anchors recovery: BankFind's current estimated losses average 23.5% of failed-bank assets at the bank level (median 22.3%; 8.6% asset-weighted, reflecting lower severities at the largest failures), consistent with the 28% reported by Granja, Matvos, and Seru (2017), whose finding that losses are higher when potential acquirers are themselves distressed is the resolution-market counterpart of the congestion technology $r-\varphi(M-1)$. We accordingly read $1-r$ as resolution severity near one quarter and treat φ as the congestion increment, disciplined by the severity elevation observed in crowded resolution periods. For the external component γ, the current BankFind `COST` estimate for 2002–2025 failures is $107.7 billion ($72.3 billion through 2022). These costs are mutualized across *surviving* banks through assessments, including the special assessment that followed the 2023 failures (FDIC 2026, 2025). Their first incidence component is therefore itself a cross-bank externality of the form the model prices: in normal resolutions γ measures the assessment burden mutualized across surviving banks, with the taxpayer backstop reaching

only tail scenarios.

What the event evidence identifies. The July 2024 CrowdStrike outage documents that a single vendor's update can impair many financial firms at once: among U.S. SEC registrants in financial industries, twenty-one issuers addressed the event in filings and six disclosed direct operational impact, with five further first-party operational-impact disclosures from non-U.S. financial institutions or market infrastructures; Kotidis and Schreft (2025) document the same propagation structure for a technology service provider cyberattack. This evidence establishes one claim: common technology shocks to multiple financial institutions are real and arrive through shared vendors. It cannot estimate λ or λ_{12}: an event list has no exposure denominator (which institutions ran the affected product) and no repeated-event frequency, so it cannot identify joint-failure probabilities. The correlation parameters remain the measurement frontier. The implementable design is the one Prediction 1 describes: residualize institution-level tail outcomes on institution and time effects and balance-sheet controls, then compare the residual products of institution pairs conditional on common technology. Commercial bank-to-vendor mappings identify a *same-vendor* differential, an adjacent-layer test of the mechanism. Identifying the cross-vendor parameter λ_{12} requires a registry field that links distinct vendors to a common distillation or model lineage, with unrelated lineages as the reference group. Section 7's lineage registry is the infrastructure that would let a supervisor run both contrasts as a matter of course. Our quantitative statements therefore hold the anchored parameters fixed and report results across the empirically open correlation range.

9.4 Testable Predictions

The model's distinctive content is cross-sectional and event-based; each prediction maps a formal result into observables and identifying variation.

Prediction 1. *Tail-risk comovement (joint downgrade, joint distress probabilities), controlling for asset-side similarity, is ordered across three pair types: highest for institutions on the same vendor's model, next for institutions on different vendors that share a distillation lineage, lowest for institutions on unrelated lineages. The three levels measure λ, λ_{12}, and zero, so the gap between the middle and bottom groups identifies the industry-wide lineage component that vendor competition cannot dilute (Proposition 9(iii)). If common shocks to model inputs are more frequent or severe under stress, each differential widens in stress periods.*

Prediction 2. *Within a decision domain, vendor HHI measures the probability that two randomly drawn adopters share a vendor. Conditional on model accuracy and portfolio composition, increases in HHI therefore forecast increases in systemic comovement among adopting institutions without a corresponding increase in their marginal failure probabilities—the model's signature separating correlation from level.*

Prediction 3. *Model upgrades that widen the frontier's accuracy lead are followed by measurable herding into the upgraded vendor by institutions previously on diverse stacks (the staircase), with the largest adoption waves, and the largest increases in joint exposure, occurring when pre-upgrade accuracy gaps were smallest.*

Prediction 4. *In jurisdictions or asset classes where model lineage is observable to creditors (e.g., disclosed core banking or risk systems), spreads load on the borrower's own expected loss from a joint event but not on the loss its model choice imposes on other banks' creditors or*

on institutions outside the pair that mutualize resolution costs. The spread response therefore understates the social concentration margin by $(\varphi + \gamma)$ *per unit of excess joint exposure.*

Prediction 1 has a direct antecedent: Kotidis and Schreft (2025) trace the propagation of an operational shock through banks sharing a third-party service provider, our mechanism's empirical shadow in an adjacent technology layer.

Together, the extensions establish the mechanism's reach and its observable content. The two-point menu is the generic equilibrium configuration of a continuous frontier; vendor competition relocates the externality to the industry's shared lineage without removing it; bailout expectations compound the pricing friction; and point-in-time bank–technology mappings would expose the mechanism through pair-level tail comovement, concentration-driven joint risk without level effects, and upgrade-triggered herding.

10 Conclusion

This paper asked whether banks choosing prediction technologies from a common vendor frontier deliver the efficient degree of model diversity. They do not: when accuracy and error correlation are bundled, private pricing captures only the part of correlation borne inside the contracting pair. The failure is architectural rather than purely informational, and it is invariant to the state contingency of competitively refinanced bilateral contracts: the deviator's counterparty is not the other bank's creditor or the outside institutions bearing resolution costs, so full transparency leaves the cross-bank share of joint-failure costs unpriced. The consequences scale badly (losses quadratic in the number of excess adopters under fixed pairwise cost coefficients) and they respond perversely to technological improvement: progress can tip diverse systems into monoculture when the frontier pulls ahead. The

regulatory implication has a sequencing structure: lineage disclosure is the precondition that makes a correlation surcharge implementable; when $\varphi > 0$, disclosure bounds the loss, and the surcharge—the replacement of one mis-set scalar by its social value—aligns pure-strategy adoption thresholds with the planner's. Vendor competition helps down to the industry's shared lineage and no further.

The mechanism may extend beyond banks and beyond artificial intelligence. When many decision-makers source a critical input from a common technological frontier whose quality dimension is priced and whose commonality dimension is not (cloud infrastructure, security software, data vendors, core banking systems), private adoption can over-concentrate and improvement can occasionally harm. The model isolates the adoption margin and holds the vendor side fixed; the natural next steps attach to the same mechanism. A dynamic version in which switching costs make tipping irreversible would convert the progress paradox into hysteresis, and a vendor-side game in which lineage itself is chosen (train independently or distill cheaply) would make the industry's shared-lineage floor an equilibrium object shaped by the same unpriced externality. The banking application is the one where the stakes are systemic and the supervisory architecture to carry the instrument already exists; it is missing only the correlation dimension this paper prices.

A Proofs

Proof of Lemma 1

With probability λ both H-banks fail iff $S = 1$ (probability p_H); with probability $1 - \lambda$ they fail independently (probability p_H^2). Hence $\pi_{HH} = \lambda p_H + (1 - \lambda)p_H^2 = p_H^2 + \lambda p_H(1 - p_H)$, and the marginal is $\lambda p_H + (1 - \lambda)p_H = p_H$. The correlation of the two failure indicators is

$(\pi_{HH} - p_H^2)/[p_H(1-p_H)] = \lambda$. At $\lambda = 1$, $\pi_{HH} = p_H$, the comonotone Fréchet bound. π_{HL} and π_{LL} are products of independent marginals. Subtracting, $\kappa = p_H^2 + \lambda p_H(1-p_H) - p_H p_L = p_H[\lambda(1-p_H) - \Delta_a]$. □

Proof of Proposition 1

Equilibrium. Fix any $R_i < y$ (Assumption 3 guarantees this for every candidate conjecture since $R_i \leq (1+\varphi)/(1-p_L) < y$). Bank i's expected equity payoff from model j is $(1-p_j)(y-R_i)$, independent of j_{-i} and of λ; the gain from H is $\Delta_a(y-R_i) > 0$, so H strictly dominates, ruling out asymmetric and mixed equilibria; financier break-even at the conjecture HH gives $R^{HH} = (1-p_H r + \varphi\pi_{HH})/(1-p_H) < y$, confirming the conjecture. *Planner.* $W_{HL} - W_{HH} = \Lambda(\pi_{HH} - \pi_{HL}) - (v_H - v_L) = \Lambda\kappa - \Delta$, positive iff $\Delta < \Lambda\kappa$; and $W_{HL} - W_{LL} = \Delta + \Lambda p_L \Delta_a > 0$ always since $\pi_{HL} - \pi_{LL} = -p_L\Delta_a < 0$, so LL is never optimal. *Wedge decomposition.* Consider one bank's deviation $H \to L$ at fixed R^{HH}. Its own financier's expected profit changes by $-\Delta_a(R^{HH} - r) + \varphi\kappa$ (worse marginal risk, lower fire-sale exposure); the other financier gains $\varphi\kappa$; parties outside the contracting pair gain $\gamma\kappa$. Summing, $G_{\text{priv}} - G_{\text{soc}} = \Lambda\kappa - \Delta_a(R^{HH} - r)$, an exact identity equal to the display (1); every term is an effect on a party outside the bank's objective. □

Proof of Proposition 2

With contingent fair pricing, $u(j,k) = (1-p_j)(y - R_{j|k}) = v_j - \varphi\pi_{jk}$. Against L: $u(H,L) - u(L,L) = \Delta + \varphi p_L \Delta_a > 0$, so LL is never an equilibrium. Against H: $u(H,H) - u(L,H) = \Delta - \varphi\kappa$, giving the threshold characterization; below the threshold the game is anti-coordination with strict pure equilibria $\{HL, LH\}$ and the standard interior mixed equilibrium, which places probability on both HH and LL and is therefore welfare-dominated.

For $\Delta \in (\varphi\kappa, \Lambda\kappa)$, comparing thresholds yields part (iii); the residual wedge equals (1) minus its first bracket, i.e., $(\varphi + \gamma)\kappa$, which is zero only if $\varphi + \gamma = 0$. □

Proof of Proposition 3

(i) With unlimited liability and $\gamma = 0$, debt is riskless ($R = 1$) and bank payoffs are $v_j - \varphi\pi_{jk}$, identical in form to the disclosure regime; because this benchmark has $\Lambda = 2\varphi$, the private threshold is $\varphi\kappa = (\Lambda/2)\kappa$ against the social $\Lambda\kappa$. Under-diversification survives on the stated band; the mover's calculus counts one of the two $\varphi\kappa$ savings its departure creates. (ii) Pooled pricing removes the profile-dependent charge and gives $\ell = 0$; contingent fair pricing or own full liability gives $\ell = \varphi$; and welfare uses $\ell = \Lambda = 2\varphi + \gamma$. Setting $\gamma = 0$ yields the Wagner half-internalization statement. Equality of threshold rules for all Δ requires and is implied by $\ell = \Lambda$; coincidence at a particular Δ can occur on either side of both cutoffs. (iii) Sufficiency of the pricing friction is Proposition 1; survival under observability is Proposition 2; survival under full liability is part (i). If a model $(p_H, 0)$ exists, it weakly lowers every joint exposure at the same accuracy. It is strictly preferred whenever the resulting exposure is priced, but under pooled pricing it gives exactly the same equity payoff as correlated H, so the correlated profile remains an equilibrium by indifference. A merged owner of both banks maximizes joint surplus with coefficient $2\varphi = \Lambda$ when $\gamma = 0$, implementing the planner's rule; Proposition 4(ii) is the general coalition statement. □

Proof of Proposition 4

Part (i), Step 1 (only expected transfers matter). Fix an arrangement and a profile (j, k). Under limited liability the bank's equity receives cash only in its own success state, and failure recoveries accrue to financiers; risk neutrality implies that every party's objective depends on

the state-contingent payment schedule only through the profile-contingent expected transfer. The adoption game therefore depends on the arrangement only through the map from profiles to expected transfers.

Step 2 (competition pins the map). At any observable profile, competing financiers offer terms with zero expected profit at that profile; a bank facing more expensive terms refinances, and cheaper terms attract no supplier. Hence the financier's break-even condition holds profile by profile: expected net payments from the bank equal $1 - p_j r + \varphi\pi_{jk}$, and the bank's payoff is $(1 - p_j)y - \big(1 - p_j r + \varphi\pi_{jk}\big) = v_j - \varphi\pi_{jk}$. The induced best-response correspondence therefore coincides with that of the disclosure payoffs, and the wedge computation of Proposition 2 applies unchanged. This step would not follow for an exclusive ex-ante contract that cross-subsidizes profiles; such contracts are outside the proposition.

Part (ii). A coalition of both banks and both financiers has joint surplus $v_{j_1} + v_{j_2} - 2\varphi\pi_{j_1 j_2}$ plus constants; competitively supplied coalition-wide arrangements implement its maximizer, the threshold rule with $\ell = 2\varphi$: diversity iff $\Delta < 2\varphi\kappa$. Efficiency requires the threshold $\Lambda\kappa$, and $2\varphi = \Lambda$ iff $\gamma = 0$. The outside parties' stake $\gamma\pi_{jk}$ appears in no arrangement limited to the two bank–financier pairs, so such an arrangement has $\ell \leq \Lambda - \gamma$.

Part (iii). Assemble the rungs: $\ell = 0$ (pooled pricing, Proposition 1); $\ell = \varphi$ (competitively refinanced bilateral arrangements, part (i)); $\ell = 2\varphi$ (industry-wide, part (ii)); $\ell = \Lambda$ (planner). The inequalities are strict at the first and final private rungs when $\varphi > 0$ and $\gamma > 0$. □

Proof of Proposition 5

The proof has three steps: welfare increments, equilibrium characterization, and the comparison.

Step 1 (welfare increments). By linearity of expectation, expected excess cost is $\Lambda Q(m)$, where

$$Q(m) = \mathbb{E}\binom{M}{2} = \binom{m}{2}\pi_{HH} + m(N-m)\pi_{HL} + \binom{N-m}{2}\pi_{LL}.$$

Direct computation gives

$$Q(m+1) - Q(m) = m\kappa - (N-m-1)p_L\Delta_a = X(m+1),$$

hence $W(m+1) - W(m) = \Delta - \Lambda X(m+1)$. This increment is strictly decreasing in m because the slope of X is $\kappa + p_L\Delta_a > 0$, establishing discrete strict concavity and the stated m^{**}. Finally, $m^{**} \geq 1$ because $X(1) = -(N-1)p_L\Delta_a < 0$.

Step 2 (equilibrium). A bank on L facing a herd of size m gains $\Delta - \ell X(m+1)$ from joining: its financier's charge, under coefficient ℓ, changes by exactly ℓ times its own pairwise exposure change, which telescopes to $X(m+1)$. A herd member gains $-[\Delta - \ell X(m)]$ from leaving. No profitable deviation is exactly the displayed double inequality, with the stated boundary conventions, and monotone X makes the intervals $[\ell X(m), \ell X(m+1)]$ partition the line: the herd size is generically unique. Under pooled pricing deviation gains are $\pm\Delta_a(y - R_i)$, independent of others' choices: $m^* = N$.

Step 3 (comparison and loss). For $0 < \ell < \Lambda$, the private cutoff lies weakly to the right of the planner's because X is increasing and every adoption with $X(m) \leq 0$ is desirable under both coefficients, while for $X(m) > 0$ the smaller coefficient lowers the private threshold. Hence $m^* \geq m^{**}$. At $\ell = \Lambda$, the equilibrium inequalities are the one-step optimality conditions for the discretely strictly concave welfare function, so every pure-strategy equilibrium herd size is a global planner optimum. At $\ell = 0$, $m^* = N \geq m^{**}$. Whenever $m^* \geq m^{**}$, the

loss formula telescopes:

$$W(m^{**}) - W(m^*) = -\sum_{m=m^{**}+1}^{m^*} [W(m) - W(m-1)] = \sum_{m=m^{**}+1}^{m^*} [\Lambda X(m) - \Delta].$$

□

Proof of Corollary 1

(i) The loss $\sum_{m=m^{**}+1}^{N}[\Lambda X(m) - \Delta]$ is an arithmetic sum with increment $\Lambda(\kappa + p_L\Delta_a)$, first term nonnegative at $m^{**}+1$; writing $K = N - m^{**}$ yields $\frac{\Lambda(\kappa+p_L\Delta_a)}{2}K^2 + O(K)$. (ii) For $0 < \ell < \Lambda$, $\Delta \geq \ell X(m^*)$ and $\Delta < \Lambda X(m^{**}+1)$ imply $X(m^*) - X(m^{**}+1) < \Delta(1/\ell - 1/\Lambda)$; dividing by the slope bounds $m^* - m^{**}$. With fixed primitives, every term in the finite loss sum is then bounded as well. At $\ell = \Lambda$, Proposition 5(iii) gives zero pure-strategy loss. In the continuous relaxation of herd size, $Q''(m) = A \equiv \kappa + p_L\Delta_a$, so the private and social maximizers differ by $\Delta(1/\ell - 1/\Lambda)/A$. Since social welfare has curvature $-\Lambda A$, evaluating its quadratic loss gives $\Delta^2(\Lambda - \ell)^2/(2A\Lambda\ell^2)$. □

Proof of Proposition 6

The proof has four steps: uniqueness of the tipping point, within-regime slopes, the jump, and planner continuity.

Step 1 (unique tipping point). Write $\Phi(\delta) = \Delta(\delta) - \ell\kappa(\delta)$. Along $p_H(\delta) = p_H^0 - \delta$: $\Delta'(\delta) = y - r > 0$ and

$$\kappa'(\delta) = -\big[\lambda(1 - p_H) - \Delta_a + p_H(1 - \lambda)\big] < 0$$

whenever Assumption 2 holds. Hence $\Phi'(\delta) > y - r > 0$. Initial diversity gives $\Phi(0) < 0$; before κ reaches zero, $\Phi = \Delta > 0$, so continuity gives a unique δ^* solving the quadratic $\Delta(\delta) = \ell\kappa(\delta)$.

Step 2 (within-regime slopes). $dW_{HH}/d\delta = 2(y - r) + \Lambda[\lambda + 2p_H(1 - \lambda)] > 0$ and $dW_{HL}/d\delta = (y - r) + \Lambda p_L > 0$: welfare rises strictly on either side of δ^*, so the tipping jump is the only source of a decline.

Step 3 (the jump). At δ^*, $W_{HL} - W_{HH} = \Lambda\kappa - \Delta = (\Lambda - \ell)\kappa(\delta^*)$ using the tipping condition, equivalently $(\Lambda/\ell - 1)\Delta(\delta^*)$; both are positive for $\ell < \Lambda$. The planner's preference for HL at δ^* follows from $\Delta(\delta^*) = \ell\kappa(\delta^*) < \Lambda\kappa(\delta^*)$. For part (iii), W_{HH} is strictly increasing after the jump and, at $p_H = 0$, exceeds the pre-tipping value by $(p_H(\delta^*) + p_L)(y - r) + \Lambda p_H(\delta^*)p_L > 0$. Hence there is a unique recovery point $\delta^{**} > \delta^*$ solving $W_{HH}(\delta) = W_{HL}(\delta^{*-})$. The ratio $(\delta^{**} - \delta^*)/\delta^*$ is not signed by the theorem and is reported only for the stated numerical designs.

Step 4 (planner continuity). The planner switches where $W_{HL} = W_{HH}$, i.e., where the jump is identically zero; continuity and strict monotonicity of planner welfare follow from Step 2, and the jump $(\Lambda - \ell)\kappa(\delta^*)$ vanishes as $\ell \to \Lambda$. □

Proof of Proposition 7

Apply the argument of Proposition 6 to each threshold statistic $\Phi_m(\delta) = \Delta(\delta) - \ell X(m; \delta)$: each is strictly increasing, and they are ordered in m because X is increasing in m, giving ordered triggers δ_m. At δ_m the marginal adoption's social value is $\Delta - \Lambda X = \Delta(1 - \Lambda/\ell)$ by the trigger condition, hence the per-step drop; the planner's trigger satisfies $\Delta = \Lambda X$, where the step change in welfare is zero. Between triggers, welfare inherits the positive within-regime slopes. Summing the drops gives the total discrete loss generated at adoption thresholds;

the continuum envelope $dW/dm|_{m^*} = \Delta(1 - \Lambda/\ell)$ is its limiting marginal analogue. □

Proof of Proposition 8

(i) In the finite H/L model, the surcharge changes the private gain from the m-th adoption to $\Delta - (\ell + s)X(m) = \Delta - \Lambda X(m)$, the planner's increment, in every regime and at every m. Discrete strict concavity of welfare then makes every pure-strategy equilibrium herd size a planner optimum, and the pure-strategy tipping condition coincides with planner indifference, where the allocation switch has zero welfare jump. This alignment does not eliminate mixed equilibria. In the two-bank anti-coordination region $0 < \Delta < \Lambda\kappa$, if the rival adopts H with probability q, the expected payoff gain from H rather than L is

$$\Delta + \Lambda p_L \Delta_a - q\Lambda(\kappa + p_L \Delta_a).$$

It vanishes at

$$q^* = \frac{\Delta + \Lambda p_L \Delta_a}{\Lambda(\kappa + p_L \Delta_a)} \in (0, 1),$$

so a strictly interior symmetric mixed equilibrium exists throughout the region, not merely at a knife edge. It assigns positive probability to HH and LL, both strictly worse than HL or LH for the planner, and is therefore welfare-dominated. A coordination or selection device is needed to implement the asymmetric optimum in every equilibrium. (ii) Under hidden choice, an ex-ante charge cannot vary with j_i, adding a constant to both sides of every comparison; an ex-post per-joint-failure levy on a failed bank exceeds the zero equity value available under limited liability. (iii) is Proposition 2(iii). □

Proof of Proposition 9

(i) When $\varphi > 0$, a bank sharing a vendor with k others in the disclosure regime strictly gains $\varphi(\lambda - \lambda_{12})\hat{c} > 0$ per shared peer by switching to the other frontier vendor (equal accuracy, lower joint exposure), so no equilibrium has a shared vendor when $\lambda_{12} < \lambda$; with vendor-sharing eliminated, the frontier–legacy margin compares $v_H - \varphi\pi_{12}$-type payoffs against $v_L - \varphi\pi_{HL}$-type payoffs, and $\pi_{12} - \pi_{HL} = \kappa_x$ delivers thresholds $\varphi\kappa_x$ (market) and $\Lambda\kappa_x$ (planner) exactly as in Propositions 2–3; $\kappa_x > 0 \iff \lambda_{12}\hat{c} > p_H\Delta_a \iff \lambda_{12} > \Delta_a/(1-p_H)$. If $\kappa_x \leq 0$ the vendor split maximizes both private payoffs and welfare (it weakly lowers every pairwise probability at no accuracy cost). At $\varphi = 0$ the switching gain is zero, so all frontier-vendor assignments are privately payoff-equivalent; a lower-exposure tie-break selects the split, but same-vendor profiles remain equilibria. The split is a planner optimum iff $\Delta \geq \Lambda\kappa_x$, a condition that holds automatically when $\kappa_x \leq 0$. (ii) Under pooled pricing, payoffs are vendor-invariant absent ε; with $\varepsilon > 0$ per same-vendor peer, any profile with split frontier banks admits a strictly profitable move to the larger vendor, and single-vendor profiles admit none: the unique robust equilibrium up to relabeling. (iii) If $m-1$ is divisible by V, an entrant joining the balanced herd forms $(m-1)/V$ same-vendor pairs (excess κ_w) and $(m-1)(V-1)/V$ cross-vendor pairs (excess κ_x), giving the displayed formula. Otherwise it joins a smallest vendor and forms $\lfloor (m-1)/V \rfloor$ same-vendor pairs; replacing the average by this integer count changes exposure by at most $\kappa_w - \kappa_x$ and leaves the limit unchanged. Balancedness for the planner follows from minimizing $\sum_v \binom{m_v}{2}$ given the herd; for a market with a positive joint-exposure coefficient it follows from each entrant preferring the smallest same-vendor exposure, while at coefficient zero it is an admissible selection among vendor-indifferent profiles. $\square$

Proof of Proposition 10

$u_i(\rho_i, \rho_{-i}) - u_i(\rho_i', \rho_{-i}) = v(\rho_i) - v(\rho_i') - \ell \sum_{k \neq i} [\pi(\rho_i, \rho_k) - \pi(\rho_i', \rho_k)] = W_\ell(\rho_i, \rho_{-i}) - W_\ell(\rho_i', \rho_{-i})$, since the terms of W_ℓ not involving ρ_i cancel: the game admits W_ℓ as an exact potential (Monderer and Shapley 1996). Thus a pure-strategy deviation is profitable exactly when it raises W_ℓ, so pure-strategy Nash equilibria are precisely its coordinatewise maxima. Interior pure-strategy equilibria satisfy the first-order criticality conditions, but joint criticality alone is not the definition of equilibrium. The box is compact and W_ℓ continuous, so a global maximizer exists and admits no profitable unilateral deviation. □

References

Acharya, V. V., and T. Yorulmazer. 2007. Too many to fail—An analysis of time-inconsistency in bank closure policies. *Journal of Financial Intermediation* 16:1–31.

Acharya, V. V., and T. Yorulmazer. 2008. Information contagion and bank herding. *Journal of Money, Credit and Banking* 40:215–31.

Alcazar, J., S. Baird, E. Cronenweth, F. Hayashi, and K. Isaacson. 2024. Market structure of core banking services providers. Payments System Research Briefing, Federal Reserve Bank of Kansas City.

Anand, K., S. Kazinnik, A. Leonello, and E. Panetti. 2026. Ex machina: Financial stability in the age of artificial intelligence. Working Paper 3225, European Central Bank.

Bank of England and Financial Conduct Authority. 2024. Artificial intelligence in UK financial services—2024 survey.

Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation. 2026. Revised guidance on model risk management. Supervisory Letter SR 26-2, April 17.

Bommasani, R., K. A. Creel, A. Kumar, D. Jurafsky, and P. Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems* 35.

Bramoullé, Y., and R. Kranton. 2007. Public goods in networks. *Journal of Economic Theory* 135:478–94.

Chen, P.-Y., G. Kataria, and R. Krishnan. 2011. Correlated failures, diversification, and information security risk management. *MIS Quarterly* 35:397–422.

Colliard, J.-E. 2019. Strategic selection of risk models and bank capital regulation. *Management Science* 65:2591–606.

Danielsson, J., R. Macrae, and A. Uthemann. 2022. Artificial intelligence and systemic risk. *Journal of Banking and Finance* 140:106290.

Danielsson, J., and A. Uthemann. 2025. Artificial intelligence and financial crises. *Journal of Financial Stability* 80:101453.

Dávila, E., and A. Korinek. 2018. Pecuniary externalities in economies with financial frictions. *Review of Economic Studies* 85:352–95.

Dou, W. W., I. Goldstein, and Y. Ji. 2025. AI-powered trading, algorithmic collusion, and price efficiency. Working Paper 34054, National Bureau of Economic Research.

Dou, W. W., I. Goldstein, and Y. Ji. 2026. Financial market fragility in the era of AI planning. Working Paper, University of Pennsylvania and Hong Kong University of Science and Technology.

Farhi, E., and J. Tirole. 2012. Collective moral hazard, maturity mismatch, and systemic bailouts. *American Economic Review* 102:60–93.

Federal Deposit Insurance Corporation. 2025. Interim final rule on special assessment collection. Financial Institution Letter FIL-58-2025, December 16.

Federal Deposit Insurance Corporation. 2026. BankFind Suite: Failures and assistance and annual data summaries. Data retrieved July 23, 2026.

Federal Financial Institutions Examination Council. 2011. Annual report, section on the Multi-Regional Data Processing Servicer program.

Federal Housing Finance Agency. 2026. Homebuying advances into new era of credit score competition. News release, April 22.

Federal Housing Finance Agency Office of Inspector General. 2019. An overview of Enterprise debt-to-income ratios. White Paper WPR-2019-002.

FICO. 2006. Fair Isaac expands availability of leading decision management solutions in Europe with new SiNSYS alliance. Corporate news release.

FICO. 2011. Bayern Card-Services protects credit cards with FICO fraud solution. Corporate news release.

FICO. 2026. FICO Score credit insights report: Average FICO Score dips to 714. Corporate news release, March 24.

Granja, J., G. Matvos, and A. Seru. 2017. Selling failed banks. *Journal of Finance* 72:1723–84.

Ibragimov, R., D. Jaffee, and J. Walden. 2011. Diversification disasters. *Journal of Financial Economics* 99:333–48.

Keloharju, M., and D. Okat. 2026. Common vendors and systemic cyber risk in banking. Working Paper, SSRN 6923918.

Kim, E., A. Garg, K. Peng, and N. Garg. 2025. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*.

Kleinberg, J., and M. Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118:e2018340118.

Kleinberg, J., A. Sinanaj, and É. Tardos. 2026. Price of anarchy of algorithmic monoculture. Working Paper, Cornell University.

Kotidis, A., and S. L. Schreft. 2025. The propagation of cyberattacks through the financial system: Evidence from an actual event. *Journal of Finance* 80:3313–58.

Kunreuther, H., and G. Heal. 2003. Interdependent security. *Journal of Risk and Uncertainty* 26:231–49.

Leitner, Y., and B. Williams. 2023. Model secrecy and stress tests. *Journal of Finance* 78:1055–95.

Lorenzoni, G. 2008. Inefficient credit booms. *Review of Economic Studies* 75:809–33.

Meng, S., and X. Chen. 2026. Artificial intelligence and systemic risk: A unified model of performative prediction, algorithmic herding, and cognitive dependency in financial markets. Working Paper, arXiv:2604.03272.

Microsoft. 2024. Helping our customers through the CrowdStrike outage. Official Microsoft Blog, July 20.

Monderer, D., and L. S. Shapley. 1996. Potential games. *Games and Economic Behavior* 14:124–43.

Parametrix. 2024. CrowdStrike's impact on the Fortune 500. Impact analysis, July 24.

Parlatore, C., and T. Philippon. 2025. Designing stress scenarios. *Journal of Finance* 80:833–73.

Peng, K., and N. Garg. 2024. Monoculture in matching markets. In *Advances in Neural Information Processing Systems* 37.

Qiu, Y., and Q. Han. 2026. Representation homogeneity and systemic instability in AI-dominated financial markets: A structural approach. Working Paper, arXiv:2604.22818.

Wagner, W. 2010. Diversification at financial institutions and systemic crises. *Journal of Financial Intermediation* 19:373–86.

Wagner, W. 2011. Systemic liquidation risk and the diversity–diversification trade-off. *Journal of Finance* 66:1141–75.

Internet Appendix to

"Correlated Intelligence: AI Adoption and Systemic Risk"

This Internet Appendix contains material referenced in the main text: the continuous-frontier analysis, including local symmetry diagnostics and verified configuration examples summarized in Section 9.1 (Appendix IA.A); the correlated-progress variant of the progress paradox (Appendix IA.B); the design of the Monte Carlo sweeps behind the magnitudes reported in Section 6 (Appendix IA.C); and a description of the numerical verification suite (Appendix IA.D). Throughout, "Proposition n" with an unprefixed number refers to the main text.

IA.A The Continuous Frontier: Potential Structure and Local Polarization Diagnostics

IA.A.1 Environment

N banks each choose a loading $\rho_i \in [0, \bar{\rho}]$ on the common lineage factor. The frontier is $p(\rho) = p_0 - \beta\rho + \frac{g}{2}\rho^2$ with $\beta - g\bar{\rho} > 0$ and $0 < p(\bar{\rho}) \leq p(\rho) \leq p_0 < 1$, so accuracy is strictly increasing in loading and $g \geq 0$ indexes diminishing accuracy returns. The pairwise failure covariance is the primitive $c\,\rho_i\rho_k$, so $\pi_{ik} = p_i p_k + c\rho_i\rho_k$; feasibility requires $c\bar{\rho}^2 \leq p(\bar{\rho})(1-p_0)$. This bound is sufficient for a coherent N-variate Bernoulli law, not only pairwise Fréchet

bounds. Suppose first that $c > 0$ and $\bar{\rho} > 0$, and choose

$$t \in \left[\frac{\sqrt{c}\,\bar{\rho}}{1 - p_0}, \frac{p(\bar{\rho})}{\sqrt{c}\,\bar{\rho}} \right], \qquad q = \frac{t^2}{1 + t^2},$$

let $Z \sim \text{Bernoulli}(q)$, and, conditional on Z, draw failures independently with probabilities $p_i + \sqrt{c}\,\rho_i(Z - q)/\sqrt{q(1-q)}$. The displayed interval is nonempty by feasibility; its two endpoints keep these conditional probabilities in $[0, 1]$, and the resulting covariance is exactly $c\rho_i\rho_k$. If $c = 0$ or $\bar{\rho} = 0$, independent Bernoulli failures with marginals p_i implement the covariance structure directly, without the displayed normalization. Values are $v(\rho) = (1 - p(\rho))y + p(\rho)r - 1$; regime payoffs $u_i = v(\rho_i) - \ell \sum_{k \neq i} \pi_{ik}$ with the internalization coefficient ℓ as in the main text; welfare uses Λ.

IA.A.2 First- and Second-Order Structure

Proposition IA.1 (FOC statistic and SOC)**.** *(i) Private and planner first-order conditions share one statistic: $v'(\rho_i) = w\,Y_i$ with $Y_i = c\,S_{-i} - (\beta - g\rho_i)P_{-i}$, where $S_{-i} = \sum_{k \neq i} \rho_k$, $P_{-i} = \sum_{k \neq i} p_k$, and $w = \ell$ (market) or Λ (planner); the wedge is $(\Lambda - \ell)Y_i$, the local version of the statistic $X(m)$ of Proposition 5. (ii) $\partial^2 u_i / \partial \rho_i^2 = -g\,[(y - r) + \ell P_{-i}] < 0$ for every $g > 0$: the second-order condition holds automatically under any strictly diminishing-returns frontier. Cross-partials are $-\ell[(\beta - g\rho_i)(\beta - g\rho_k) + c] \leq 0$, strictly negative when $\ell > 0$: loadings are weak strategic substitutes, and strictly so under positive internalization. (iii) Along a stable interior symmetric branch with $Y > 0$, $d\rho^*/dw = Y/\Phi_\rho < 0$. Hence, when the market and planner solutions lie on the same such branch, the market ($\ell < \Lambda$) over-loads: $\rho^*(\ell) > \rho^*(\Lambda)$.*

The statistic Y nets a correlation charge cS_{-i} against an accuracy credit $(\beta - g\rho_i)P_{-i}$:

higher loading also lowers own failure probability, which reduces joint exposure with every counterparty. Interior analysis requires the charge to dominate ($Y > 0$); otherwise loading is corner-dominant.

IA.A.3 The Affine Frontier: Endogenous Two-Tier Structure

Proposition IA.2 (Affine frontier)**.** *Let $g = 0$. Then (i) each bank's payoff is linear in its own loading, so an endpoint best response always exists and, away from exact indifference, every best response is at an endpoint of $[0, \bar{\rho}]$; (ii) W_w is multilinear, its symmetric-point Hessian is $-w(\beta^2 + c)(J - I)$, concave along the symmetric direction and convex along every asymmetric direction, and at least one maximum over the box is attained at a vertex; (iii) consequently, a two-tier potential-maximizing equilibrium and a two-tier planner optimum exist, and generically all best responses are two-tier. With the corner mapping $p_H = p(\bar{\rho})$, $p_L = p_0$, $\Delta_a = \beta\bar{\rho}$, $\kappa = c\bar{\rho}^2 - p_H\Delta_a$, the two-tier thresholds satisfy the exact nesting identity*

$$X(n) \;=\; \bar{\rho}\left[(c + \beta^2)(n-1)\bar{\rho} - \beta(N-1)p_0\right],$$

so the model collapses exactly to the N-bank model of Section 5.

Away from exact indifference, therefore, the two-point menu of the baseline model is a generic equilibrium configuration of the affine frontier: diversity endogenously means *separation*, not moderation.

IA.A.4 Curvature Diagnostics and Configuration Examples

For $w > 0$, at an interior symmetric critical point $\tilde{\rho}$ of W_w, write $\tilde{p} = p(\tilde{\rho})$, $\tilde{\beta} = \beta - g\tilde{\rho}$. The asymmetric-direction Hessian eigenvalue is $d - o = -g[(y - r) + w(N-1)\tilde{p}] + w(\tilde{\beta}^2 + c)$, so

this point is a strict local maximum in every direction if and only if

$$\theta_{\text{pot}}(w) \;:=\; \frac{g\,[(y-r)+w(N-1)\tilde{p}\,]}{w\,(\tilde{\beta}^2+c)} \;>\; 1,$$

and a saddle if and only if $\theta_{\text{pot}}(w) < 1$. At a fixed evaluation point, $\partial\theta_{\text{pot}}/\partial w = -g(y-r)/[w^2(\tilde{\beta}^2+c)] < 0$: higher internalization makes that point more polarization-prone. The derivative holds the evaluation point fixed, and the market and planner critical points differ, so Table IA.1 compares their global maximizers directly.

Proposition IA.3 (Local symmetry diagnostic)**.** *For $w > 0$, at an interior symmetric critical point of W_w, the symmetric-direction Hessian eigenvalue is strictly negative. The $N-1$ asymmetric eigenvalues are negative if $\theta_{\text{pot}}(w) > 1$ and positive if $\theta_{\text{pot}}(w) < 1$. Thus the point is respectively a strict local maximum or a saddle of the potential. At a common fixed evaluation point, $\theta_{\text{pot}}(w)$ is strictly decreasing in w.*

The diagnostic points to three configurations: smooth symmetric over-loading; a symmetric market paired with an asymmetric planner (configuration failure); and polarization of both objectives. Table IA.1 reports one machine-verified instance of each pattern ($N = 5$); in each, the best profile within a symmetric/two-tier search was locally refined, and the market profile was confirmed by fine-grid unilateral-deviation checks.

Remark IA.1 (Selection, uniqueness, and scope)**.** *Existence of a potential-maximizing equilibrium follows because W_ℓ attains its maximum on the compact box and a global maximizer admits no profitable unilateral deviation. We use the best symmetric/two-tier candidate as a transparent numerical selection and verify that the reported market profile is a Nash equilibrium. Two global questions remain open. First, the ordering $\theta_{\text{pot}}(\Lambda) < \theta_{\text{pot}}(\ell)$ across their* respective *evaluation points is proven only at a common evaluation point, though it holds*

Table IA.1: Verified instances of three configuration patterns

Pattern	$\theta_{\text{pot}}(\ell)$	$\theta_{\text{pot}}(\Lambda)$	Market outcome	Planner outcome	Welfare loss
(a) smooth herding	21.3	2.64	symmetric at 0.719	symmetric at 0.344	0.059
(b) configuration failure	3.49	0.44	all 5 at 0.662	1 at 0.831, 4 at 0	0.153
(c) endogenous polarization	0.53	0.09	2 at 0.775, 3 at 0	1 at 0.775, 4 at 0	0.021

"Welfare loss" compares the reported planner candidate with the market profile under W_Λ, in units of one bank's investment. Pattern (b)'s candidate places its specialist *deeper* at the frontier (0.831) than the market's common loading (0.662): fewer, deeper frontier banks plus a diverse periphery.

in every reported instance. Second, two-tier optimality for $g > 0$ remains open because the search ranges only over symmetric and two-tier profiles; for $g = 0$, multilinearity forces a vertex solution. We conjecture that every local maximum of W_w has at most two distinct loadings; the results above are proved without it.

Remark IA.2 (Relation to specialization in public-good games)**.** *Polarized equilibria under strategic substitutes are the specialization phenomenon of Bramoullé and Kranton (2007) on the complete graph: staying off the frontier is the local public good, and pattern (c) is its specialized provision. The differences are the covariance externality with an accuracy-bundled price (their spillovers are linear in effort), and that our planner/market comparison is over the* coefficient *of one objective rather than the network.*

IA.A.5 Proofs for Appendix IA.A

Proposition IA.1. Direct differentiation of u_i and W; the cross-partial and own second derivative are immediate; (iii) is the implicit function theorem applied locally to $\Phi(\rho, w) = v'(\rho) - wY(\rho)$ with $\Phi_\rho < 0$ at a stable symmetric point and $\Phi_w = -Y < 0$ when $Y > 0$.

Integrating the sign comparison requires that both solutions remain on this stable branch. □

Proposition IA.2. With $g = 0$, p is affine so every term of u_i and W_w is multilinear in the profile. A multilinear function on a box has a vertex maximizer, and linearity of u_i in ρ_i makes an endpoint a best response; both endpoints and interior actions can also be best responses only on the exact-indifference hyperplane. A vertex maximizer of W_w is a pure-strategy Nash equilibrium by the potential property. The Hessian of W_w at any point has zero diagonal and off-diagonal $-w(p'^2+c) = -w(\beta^2+c)$, giving eigenvalues $-w(\beta^2+c)(N-1)$ (symmetric direction) and $+w(\beta^2+c)$ (asymmetric directions, multiplicity $N-1$). The nesting identity is algebra: expand $X(n) = (n-1)\kappa - (N-n)p_L\Delta_a$ under the corner mapping. □

Proposition IA.3. The Hessian of W_w at a symmetric point has diagonal $d = -g[(y-r)+w(N-1)\tilde{p}]$ and off-diagonal $o = -w(\tilde{\beta}^2+c)$. Its symmetric eigenvalue is $d+(N-1)o < 0$; each of the $N-1$ asymmetric eigenvalues is $d-o$, and $d-o < 0 \iff \theta_{\text{pot}}(w) > 1$. Differentiating θ_{pot} with respect to w while holding the evaluation point fixed gives $-g(y-r)/[w^2(\tilde{\beta}^2+c)] < 0$. □

IA.B Correlated Progress: The Technological Variant of the Paradox

The main text's Proposition 6 studies progress that improves accuracy with the correlation structure fixed—the *strategic* paradox, in which progress destroys the diversity equilibrium. The empirically motivated variant lets progress carry correlation with it: larger, more accurate models are more correlated (Kim et al. 2025), so let $\lambda(\delta) = \lambda^0 + m_\lambda\delta$ along $p_H(\delta) = p_H^0 - \delta$, on any interval for which $0 < p_H(\delta) < p_L$ and $0 \le \lambda(\delta) \le 1$.

Proposition IA.4 (Smooth decline under pooled pricing)**.** *Along this feasible path, under pooled pricing the equilibrium is monoculture for every δ, every step of progress is adopted (adoption comparisons are λ-blind), and equilibrium welfare satisfies*

$$\frac{dW}{d\delta} < 0 \quad \Longleftrightarrow \quad m_\lambda \;>\; \bar{m}(\delta) \;=\; \frac{2(y-r) + \Lambda\big[2p_H(\delta) + \lambda(\delta)(1-2p_H(\delta))\big]}{\Lambda\, p_H(\delta)\,(1-p_H(\delta))}.$$

Under disclosure with $\varphi > 0$, rising λ inflates the priced threshold $\varphi\kappa$ while diverse profiles are λ-blind. Conditional on an interior diversity-to-monoculture transition, correlation-increasing progress therefore delays tipping relative to an accuracy-only path; the correlation channel affects welfare smoothly only after monoculture is selected.

Proof. Write $p = p_H(\delta)$ and $\lambda = \lambda(\delta)$. Under pooled pricing, strict dominance of the frontier model gives the HH equilibrium throughout the feasible path. Its welfare is

$$W_{HH} = 2\big[(1-p)y + pr - 1\big] - \Lambda\big[p^2 + \lambda p(1-p)\big].$$

Because $p'(\delta) = -1$ and $\lambda'(\delta) = m_\lambda$,

$$\frac{d}{d\delta}\big[p^2 + \lambda p(1-p)\big] = -2p + m_\lambda p(1-p) - \lambda(1-2p).$$

It follows that

$$\frac{dW_{HH}}{d\delta} = 2(y-r) + \Lambda\big[2p + \lambda(1-2p)\big] - \Lambda m_\lambda p(1-p),$$

which is negative exactly under the displayed threshold. Under disclosure, diverse welfare is

$W_{HL} = v_H + v_L - \Lambda p_H p_L$ and hence contains no λ. Moreover,

$$\frac{\partial \kappa}{\partial \lambda} = p_H(1 - p_H) > 0.$$

Thus, when $\varphi > 0$, a higher λ raises the private HH-versus-HL cutoff $\varphi\kappa$ at each fixed accuracy level and delays any interior tipping point relative to the path holding λ fixed. Once HH is selected, the derivative above governs the smooth correlation effect. □

In a verified instance ($p_H^0 = 0.095$, $\lambda^0 = 0.5$, $m_\lambda = 50$, $\bar{m} \in [33.3, 42.0]$), welfare declines monotonically along the entire path under pooled pricing, while disclosure sustains diversity until $\delta = 0.0027$ with strictly higher welfare at that point (1.659 versus 1.594). The two variants are complements: Proposition 6 is the strategic paradox, Proposition IA.4 the technological one.

Remark IA.3 (The accuracy–correlation coupling is an empirical input)**.** *How often the technological variant binds depends on the coupling between accuracy gains and correlation gains, a technological primitive that the model takes as given. Under the specification $d\lambda \propto \delta\, p_H$ (correlation gains vanishing as failure probabilities shrink), the variant binds on under one percent of the Monte Carlo box of Appendix IA.C; under the equally defensible $d\lambda \propto \delta$, it binds on* 47.9% *of the wide box and* 62.9% *of the anchored box. The threshold $\bar{m}$ is therefore the result; the coupling is an empirical input, and the cross-model evidence in Kim et al. (2025) is the natural source for pinning it down.*

IA.C Monte Carlo Design

The magnitudes reported in Section 6 come from sweeps over two parameter boxes, with all draws filtered by the model's admissibility conditions (feasibility $y > (1 + \varphi)/(1 - p_L)$;

participation; Assumption 2 where stated). The *anchored* box uses the FDIC discipline reported in Section 9.3: $p_L \in (1\%, 8\%)$ is interpreted as a multi-year or stress-conditional risk, not an annual unconditional failure rate; $r \in (0.72, 0.914)$ spans the 28% severity estimate in Granja, Matvos, and Seru (2017) and the 8.6% asset-weighted BankFind severity, with $r = 0.765$ corresponding to the 23.5% bank-level mean. The ranges for λ, φ, and γ remain design ranges because the available evidence does not identify them. The *wide* box repeats the sweep on substantially looser ranges.

Table IA.2: Monte Carlo sweep boxes

Parameter	Anchored box	Wide box
Legacy failure probability p_L	(0.01, 0.08)	(0.005, 0.25)
Frontier correlation λ	(0.3, 0.9)	(0.05, 1)
Success cash flow y	(1.05, 1.5)	(1.02, 2.5)
Recovery r	(0.72, 0.914)	(0.1, 0.8)
Fire-sale discount φ	(0.02, 0.15)	$(0, r)$
External cost γ	(0.1, 1)	(0, 2)
Frontier failure probability p_H	$U(0, p_L)$	

Each box uses 20,000 draws; draws violating admissibility are discarded and redrawn. The anchored box ties p_L and r to the evidence reported in Section 9.3; the ranges for λ, φ, and γ are design ranges.

Headline outputs (anchored box; wide box in parentheses): an interior tipping point exists on the progress path of *every* admissible draw—an analytic fact (the tipping condition is a sign-changing quadratic) that the sweep confirms; tipping occurs at small accuracy edges, p_H^*/p_L median 0.92 (0.96); the welfare drop at tipping has median 2.1% (2.8%) of first-best welfare, with ninetieth percentile 5.9% (13.0%). A discrete 5–10% accuracy improvement is welfare-reducing on 4.5–7.0% (2.8–3.4%) of the box under disclosure and on exactly zero draws under pooled pricing (there is nothing to tip); when harmful, the improvement destroys a median 4.7–10.8 (2.6–5.1) times the welfare gain it would deliver under efficient adoption. At a 50% improvement the harmful fraction falls to 0.3% in the anchored box (essentially

zero in the wide box) and the median loss-to-gain ratio to about 0.2: the paradox concerns marginal progress. Ratios to first-best inflate mechanically in thin-NPV corners of the wide box (near the feasibility boundary); absolute losses per unit of banking assets in the anchored box have median 0.6% and maximum about 3.1%. All headline numbers were reproduced by an independent implementation with a fresh random seed.

IA.D The Numerical Verification Suite

The encoded algebraic identities and finite-model equilibrium characterizations are checked symbolically and by brute force: equilibrium sets are computed by exhaustive deviation checks and compared against threshold characterizations on random parameter draws, while expected costs from exact enumeration of the failure technology are compared against pairwise formulas. These computations verify the stated identities and reproduce the reported examples; the universal and global claims are established by the proofs. Table IA.3 lists the scripts included in the replication package.

A separate empirical scaffold implements the measurement design of Section 9.3. It takes point-in-time bank–vendor and bank-outcome panels, residualizes the institution-level outcome on institution and period fixed effects and balance-sheet controls, and estimates pair-level residual covariance with dyadic institution and period clustering. The input contract separates a vendor identifier from a cross-vendor lineage identifier: absent the latter, the code suppresses the lineage decomposition and never relabels a same-vendor coefficient as λ_{12}. A synthetic acceptance test recovers the imposed ordering (same vendor $>$ different vendors with common lineage $>$ unrelated lineages), verifies identification from vendor switchers under pair fixed effects, and confirms that current snapshots cannot be backfilled

Table IA.3: The numerical verification suite

Script	Verifies	Result
`check_model.py`	Two-bank identities; equilibria; planner; wedge	all identities = 0; 4,000 draws, 0 failures
`*_referee_*.py` (3)	Independent rederivations, fresh code and seeds	$2 \times 20{,}000$ draws, 0 failures
`check_t2.py`	Tipping statistic, jump, planner continuity	identities = 0; 229 tipping cases, 0 failures
`t2_adversarial.py`	Fails-region claims; vendor entry; selection	all confirmed; reversals verified constructively
`t2_sweep.py`	Monte Carlo of Appendix IA.C	replicated with fresh seeds
`check_N.py`	$X(m)$; threshold equilibria; Corollary 1; staircase	identities = 0; 400 economies, 0 failures
`check_frontier.py`	Appendix IA.A: potential, local diagnostic, examples	identities = 0; three patterns reproduced
`check_vendors.py`	Section 8: nested lineage; κ_x; immunity; κ_{eff}	identities = 0; 150 economies, 0 failures
`check_invariance.py`	Proposition 4: profile-fair spot contracts; coalition stakes	identities = 0
`adversarial_audit.py`	Derivative, unbundling, and recovery counterexamples	corrected identities and counterexamples reproduced
`make_figures.py`	Figures 1–2 from exact model formulas	reproducible

"Identities = 0" means the displayed encoded identities were reduced to zero symbolically; enumeration checks compare finite equilibrium sets, planner solutions, and expected costs computed from primitives against the paper's characterizations.

into historical panels by default. These tests cover implementation and failure guards; testing the empirical prediction itself requires point-in-time bank–vendor panels.

References

Bramoullé, Y., and R. Kranton. 2007. Public goods in networks. *Journal of Economic Theory* 135:478–94.

Kim, E., A. Garg, K. Peng, and N. Garg. 2025. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning.*