

Measuring Algorithmic Monoculture in Credit Intermediation: Supply-Chain Concentration, Decision Convergence, and the Case for Reference-Portfolio Supervision

Deepak Ranjan Sahoo

dcdrsahoo@gmail.com

MIT Art, Design and Technology University

Vaishali Deepak Sahoo

Vishwakarma University

Research Article

Keywords: algorithmic monoculture, model risk, systemic risk, artificial intelligence, credit supply, supervisory benchmarking, foundation models, macroprudential policy

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10928394/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Measuring Algorithmic Monoculture in Credit Intermediation: Supply-Chain Concentration, Decision Convergence, and the Case for Reference-Portfolio Supervision

Abstract

As far as the diagnosis is concerned, supervisors are converging more quickly than they are on the measurement. The Financial Stability Board, the Advisory Scientific Committee of the European Systemic Risk Board, and the Financial Policy Committee of the Bank of England have all identified uniformity of models as a potential pathway for artificial intelligence to exacerbate systemic risk: a growing number of institutions now rely on a limited number of common foundation models, data vendors and cloud platforms. On what data, at what frequency, and what would be measured: none has said. This paper seeks to fill that gap by addressing the issue in three stages. First, it splits between two things that policy discussion combines: concentration in the supply of model components (which can be observed and is already more or less reported under the existing third-party risk setting) and convergence in the decisions made by institutions (which transmits shocks and is not yet observed). Neither implies the other. Second, it develops a signal-extraction model of credit supply in which shared model dependence raises the correlation of forecast errors while lowering their variance, and shows that aggregate credit volatility depends on the product of the two. Migration to a shared model is destabilising when the proportional increase in error correlation exceeds the proportional reduction in error variance; under plausible illustrative parameters, that threshold binds well before adoption is complete. Third, it proposes an estimator for the error-correlation parameter built on the supervisory benchmarking machinery that already exists under Article 78 of the Capital Requirements Directive, and is candid about where identification fails, chiefly on low-default portfolios, where realised outcomes are too scarce to pin down correlation. Simulation shows the raw residual-correlation statistic to be badly contaminated by the shared outcome term, and proposes a null-normalised monoculture index that recovers the parameter; its sampling error implies roughly eighty defaults are needed for usable precision, which retail portfolios supply and wholesale portfolios do not. The paper argues against capital add-ons and diversity mandates on present evidence, and for a narrower requirement that institutions maintain a genuinely uncorrelated fallback capable of carrying a defined share of volume. The empirical case for monoculture in banking today is weak; the case for building the measurement apparatus before it is not.

Keywords: algorithmic monoculture; model risk; systemic risk; artificial intelligence; credit supply; supervisory benchmarking; foundation models; macroprudential policy

1. Introduction

A supervisor who wants to know how exposed the banking system is to a common credit shock can look it up. Exposure concentrations are reported, large-exposure limits are binding, and the data are granular enough that the Greenwood, Landier and Thesmar (2015) style of vulnerability calculation can be run off supervisory returns. A supervisor who wants to know how exposed the same system is to a

common model shock has nowhere to look. Nobody collects the data, and it is not obvious what data would need to be collected.

That asymmetry is becoming awkward. The Financial Stability Board (2024) identified reliance on common models, datasets and infrastructure as a source of correlated behaviour across institutions, capable of amplifying market stress. The Advisory Scientific Committee of the European Systemic Risk Board went further, ranking model uniformity among five features of artificial intelligence that could meaningfully amplify the classic sources of systemic risk, alongside concentration and entry barriers, monitoring difficulty, overreliance, and speed (Cecchetti et al., 2025). The Bank of England's Financial Policy Committee has treated concentration in the supply of AI services as a channel warranting continued monitoring since its April 2025 assessment, and the Financial Stability Board's June 2026 consultation on sound practices returns to the same theme while conceding that authorities face real difficulty tracking AI use that is embedded inside procurement, vendor workflows and third-party tooling rather than registered as a model in a bank's inventory.

The diagnosis is now agreed upon. The measurement is not even considered. This paper focuses on that gap, and on a deliberately limited part of the problem - retail and small-business credit decisioning, where there are many decisions, they are similar across institutions, and where the final test of the decisions comes when defaults are observed. Trading and market-making raise related concerns and have attracted most of the herding literature, but they are harder to instrument because the counterfactual is unobservable and positions are confidential. Credit is the setting where a measurement framework can actually be built.

Three contributions follow. The first is conceptual. Policy documents use "model concentration" to refer to two distinct phenomena: concentration in who supplies model components, and convergence in what decisions come out. They then treat the first as evidence of the second. It is not. Two banks fine-tuning the same base model on different portfolios, with different feature sets, different cutoffs and different override cultures, can produce decisions whose errors are close to uncorrelated. Two banks running entirely different architectures on the same bureau file, optimising the same regulatory-capital objective against the same population, can produce decisions whose errors are nearly identical. Bommasani et al. (2022) found, to their own surprise, that shared foundation models did not reliably increase outcome homogenisation, and that the method of adaptation mattered more than the shared component. Any supervisory framework anchored on provider market shares will therefore misclassify in both directions.

The second contribution is analytical. Section 4 sets out a signal-extraction model in which each bank forms a credit assessment with error that has a common and an idiosyncratic component, and in which the choice of model technology moves both: adopting a shared frontier model lowers the variance of the error and raises the share of it that is common. Aggregate credit supply then fluctuates with the

product of error variance and error correlation, while the private return to model choice depends on variance alone. The wedge is a standard externality, and it yields a threshold condition that is easy to state and, in principle, estimable.

The third is practical. Section 5 argues that the estimator supervisors need is the cross-institution correlation of prediction residuals on a common outcome-labelled portfolio, and that the institutional machinery for collecting it already exists. Article 78 of the Capital Requirements Directive (Directive 2013/36/EU) requires competent authorities to assess the quality of internal approaches annually, and directs them to attend to approaches that produce significantly different own funds requirements for the same exposure and to approaches exhibiting particularly high or low diversity. The European Banking Authority's benchmarking exercises, built on hypothetical portfolios, have operationalised the first of those concerns for more than a decade. The second, low diversity, has never been operationalised at all, because for fifteen years the design problem was that internal models produced too much unexplained dispersion in risk weights, not too little. That framing is now obsolete, and the fix is largely a matter of re-pointing an existing instrument rather than building a new one.

Two things this paper does not claim. It does not claim that banking currently exhibits algorithmic monoculture. The joint Bank of England and Financial Conduct Authority survey for 2024 found that while 75% of responding firms were using some form of AI, up from 58% in 2022, foundation models accounted for only 17% of use cases and just 2% of use cases were fully autonomous. The Committee has come to its own conclusion that by 2026 advanced AI has not yet been applied in a way that creates systemic risk. Nor is the paper against monoculture, once it exists, simply being evil. Kleinberg and Raghavan (2021) show that convergence on a common screening algorithm can lower decision quality even absent any shock, but the same convergence delivers real accuracy gains to each adopter, and those gains fall disproportionately on borrowers who are poorly served by conventional scoring (Berg et al., 2020; Fuster et al., 2022). A policy that mandates diversity imposes its cost on precisely those applicants. That trade-off is the reason Section 6 declines to recommend the two instruments most often floated — a capital add-on scaled to model commonality, and a hard diversity requirement — and recommends something narrower instead.

The timing argument is worth stating plainly, because it determines what kind of paper this is. Under the Digital Omnibus on AI (European Commission, 2025), endorsed by the European Parliament on 16 June 2026 and approved by the Council on 29 June 2026, the high-risk obligations applying to stand-alone Annex III systems, creditworthiness assessment among them, now apply from 2 December 2027 rather than 2 August 2026. Institutions are in a compliance runway during which they will build AI inventories, document data provenance, and formalise the boundary between provider and deployer responsibilities. Almost all of the raw material a monoculture metric would need is being assembled anyway, for other reasons, on a known schedule. If the reporting schema is specified now, the marginal cost of collecting it is small. If it is specified after the exposure has built, supervisors will have no pre-

adoption baseline against which to interpret anything they measure, which is the position they were in with securitisation in 2005 and with money market funds in 2007.

2. Antecedents: an old trade-off in new hardware

The observation that individually prudent risk management can be collectively destabilising is not new, and the current policy discussion has largely proceeded without reference to the literature that formalised it. Wagner (2010) showed that diversification at the level of the individual institution, which reduces that institution's probability of failure, makes institutions more similar to one another and can raise the probability of joint failure; the socially optimal degree of diversification is therefore lower than the privately optimal one. Wagner (2011) sharpened the point into a diversity–diversification trade-off in which full diversification is never socially optimal because it eliminates the heterogeneity that allows some institutions to survive and absorb the assets of others. Ibragimov, Jaffee and Walden (2011) reached a compatible conclusion from a different direction, showing that intermediaries' privately optimal diversification produces excessive systemic risk when returns have heavy tails. Allen, Babus and Carletti (2012) traced how asset commonality across institutions interacts with debt maturity to determine whether the system experiences contagion. Acharya and Yorulmazer (2007) supplied the incentive complement: banks may deliberately correlate their exposures because joint failure attracts more forbearing treatment than isolated failure.

Together these papers give a general result. Institutions making their own choice and in the case of a convex failure cost, private optimisation yields too much homogeneity. In none of that argument is there reliance on the technology by which similarity is created. For all intents and purposes, the shared foundation models are just a new, very efficient way of producing it. What is new is the mechanism's quickness and its transparency. Portfolio commonality only emerges over time and is observable in the exposure reporting. Model commonality can come in a procurement cycle, and it is not apparent in every bank's return that it files today. A bank that has changed from an internally estimated logistic model to an underwriting engine that is a vendor product based on a widely licensed logistic model has altered the correlation structure of its errors with all other banks utilizing the vendor, and has revealed nothing that would allow others to guess. If a correlated tightening does materialize, it is amplified by the channels through which such events happen in the crisis literature, such as the leverage and margin adjustment channels that Adrian & Shin (2010) and Brunnermeier & Pedersen (2009) describe. Shared model error is like a new source of the initial impulse, rather than a new transmission channel.

The prudential apparatus that ought to catch this was not designed to. Model risk management as codified in the Federal Reserve's SR 11-7 (2011) and, more recently, in the Prudential Regulation Authority's SS1/23 (2023) is organised around a single institution's model inventory: identification, validation, ownership, independent review, and a tiering of models by materiality to that institution. Every question it asks is well posed at firm level and none of them has an answer that depends on what

other firms are doing. A model can pass validation at all forty banks using it and still constitute a system-level vulnerability, and no step in the process would register the fact. This is not a deficiency in the guidance, which does what it set out to do; it is a statement about the level at which the guidance operates.

The computer science literature supplies the second antecedent. Kleinberg and Raghavan (2021) model firms screening applicants either through a shared algorithm or through independent noisy evaluation, and show that convergence on the algorithm can reduce the quality of decisions across the full set of firms even when the algorithm is more accurate for any single firm in isolation, and even under normal operations with no shock at all. Their mechanism runs through competition for the same candidates rather than through correlated failure, which makes the result stronger than the agricultural analogy it displaces: monoculture does not need a blight to be costly. Bommasani et al. (2022) expanded this definition to shared components (training data and pre-trained models) and determined that data sharing resulted in consistent outcomes among individuals and racial groups, but foundation-model sharing did not necessarily lead to the same results. The second finding contradicts the simple story, and should be given greater consideration than it has been in the financial policy debate. It implies that the transmission takes place in the way institutions do, rather than in the base model.

The third strand is the recent financial-stability work. Aldasoro et al. (2024) frame the financial system as an information-processing apparatus and assess how generative models and emerging agentic systems change it, including implications for prudential policy. Danielsson and Uthemann (2024) argue that AI adoption alters the character of crises rather than merely their probability, because engines optimised on historical data behave unpredictably in states that have no historical analogue. Foucault et al. (2025) collect the field's current state. Three sets of official-sector documents are consistent both in naming model uniformity as a vulnerability and in stopping short of proposing a metric: the Financial Stability Board (2024, 2026), Cecchetti et al. (2025) for the ESRB, and the Bank of England's Financial Stability in Focus (2025). The Financial Stability Board's 2026 consultation is explicit that authorities' visibility is limited by the difficulty of tracking AI embedded in third-party services and workflow tools rather than declared in a model inventory.

3. The dependency stack, and why provider shares mislead

Consider what a bank's credit decision now depends on. At the base sits compute and cloud infrastructure, where the effective number of providers globally is small and where the UK's critical third parties regime, established by the Financial Services and Markets Act 2023, and the EU's Digital Operational Resilience Act (Regulation (EU) 2022/2554) already impose obligations. Above that sits the pre-trained model, which for a growing share of use cases is licensed rather than built. On top of that, all of the institution's efforts to make a general model a decision; the polishing, extra data manipulation, bureau and alternative data feeds, vendor middleware, prompt and retrieval setting,

calibration to its own portfolio. The deployment policy is at the top: the threshold, the exceptions, the exception process, the exception committee. The questions of relevance for supervision are presented in Table 1: is the layer observed? Does the layer have a concentration that actually leads to decisions?

Table 1. The model dependency stack in credit decisioning

Layer	Content	Observability today	Transmission to decision errors
L1 Infrastructure	Compute, cloud hosting, inference serving	High — DORA register of information; UK CTP designations	Availability and outage risk; little effect on the content of decisions
L2 Base model	Pre-trained foundation model or licensed scoring engine	Low — no reporting obligation; often embedded in vendor products	Potentially high, but conditional on L3
L3 Adaptation	Fine-tuning data, bureau and alternative-data feeds, vendor middleware, calibration	Very low — partially captured in model inventories, rarely at component level	High; the empirical evidence points here
L4 Deployment	Cutoffs, override policy, appetite, human review	Moderate — internal governance records, ICAAP documentation	High and institution-specific; the main source of legitimate divergence

Notes: CTP denotes the UK critical third parties regime established under the Financial Services and Markets Act 2023. DORA denotes Regulation (EU) 2022/2554.

Two features of the table drive the argument. The layers where concentration is easiest to observe, L1 and to some extent L2, are the layers that transmit least directly to the content of credit decisions. An outage at a cloud provider is a serious operational event and belongs squarely inside the operational resilience perimeter, but it does not make two banks approve the same borrowers. The layers that determine whether errors are correlated, L3 and L4, are the layers on which supervisors currently have almost no data.

The second feature is that the layers do not compose multiplicatively. A vertically integrated provider may occupy L1, L2 and L3 simultaneously, in which case a stack-wide index built by combining layer-level Herfindahl measures will understate the true dependency. Conversely, a bank using a widely licensed base model with a proprietary adaptation pipeline and an idiosyncratic override culture may have low error correlation with its peers despite maximal apparent concentration at L2. Provider share is a poor proxy in both directions, and a supervisory regime built on it would generate false comfort for the second case and false alarm for the first.

There is a further complication that the operational resilience framing tends to obscure. Concentration risk as conventionally understood is about failure: a provider goes down, and everyone who depends

on it goes down together. The essence of the monoculture risk is that the provider performs exactly as intended, all institutions that use him adjust the assessment of borrower quality in the same direction at the same time, and the supply of credit follows suit. No incident occurs. Nothing is reported. It's procyclicality and the critical third parties regime does not assume this.

The distinction can be stated as a proposition.

P1. Concentration in the supply of model components is neither necessary nor sufficient for convergence in decision errors across institutions. That is, a supervisory framework that infers systemic model risk from provider market shares will misclassify institutions in both directions, and the misclassification is not conservative.

4. A model of shared dependence and aggregate credit supply

The formal argument is deliberately simple, because the point is a threshold rather than a full equilibrium.

4.1 Setup

Let there be N banks indexed $i = 1, \dots, N$, lending to a common pool of borrowers. Borrower quality in period t is summarised by a latent state θ , distributed normally with variance $\sigma^2\theta$. Bank i forms an assessment

$$s_i = \theta + \varepsilon_i \quad (1)$$

where the assessment error ε_i is normal with mean zero and variance $\sigma^2\varepsilon$, and $\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho$ for $i \neq j$. Decompose the error into a common and an idiosyncratic part,

$$\varepsilon_i = \sqrt{\rho} \cdot \eta + \sqrt{(1 - \rho)} \cdot v_i \quad (2)$$

If η and v_i are independent and have the same variance $\sigma^2\varepsilon$. model error" η is a general term for any error made by all banks, when they are using the shared technology: a mis-specified feature; a training distribution that no longer reflects the population; a systematic mis-calibration of a treatment for a borrower segment. The term v_i is the error associated with bank i , which is due to bank specific data, calibration and judgement.

The two parameters include the economics: the better the model, the more accurately it can predict the borrower quality; the higher the value of ρ the more strongly the banks are dependent on each other, the more similar models they use, the more similar their training sets are, the more similar their pipeline of borrowers, and the more they are pushed by regulatory or commercial pressures toward a common feature set. The critical structural assumption is that the two move together. Adopting a shared frontier technology lowers $\sigma^2\varepsilon$ and raises ρ . Institutions choose a point on a frontier $\rho(\sigma^2\varepsilon)$ with $d\rho/d\sigma^2\varepsilon < 0$, and they choose it one at a time, taking the technology as given.

Bank i extends credit according to a threshold rule on the posterior, approving when $E[\theta | s_i]$ exceeds an appetite parameter. Linearising around the operating point, bank-level credit supply is

$$L_i = a + b(\mu + \sqrt{\rho} \cdot \eta + \sqrt{(1 - \rho)} \cdot v_i - c) \quad (3)$$

where c is the cutoff, μ the conditional mean of borrower quality, and $b > 0$ the sensitivity of lending volume to assessed quality.

4.2 Aggregate fluctuation

Let L denote the cross-sectional average of L_i . Averaging and taking the variance conditional on the true state,

$$\text{Var}(L | \mu) = b^2 \sigma^2 \varepsilon [\rho + (1 - \rho)/N] \quad (4)$$

which converges, as N grows, to

$$\text{Var}(L | \mu) \rightarrow b^2 \sigma^2 \varepsilon \cdot \rho \quad (5)$$

The idiosyncratic component washes out in aggregate; the common component does not. Figure 1 plots the bracketed term in (4) against ρ for several values of N . The lesson of the figure is how little cross-sectional diversification is worth once ρ is even moderately positive: at $\rho = 0.5$, a system of a hundred banks has 99.5% of the aggregate variance that a system of five banks would have. Adding institutions does not help. Only reducing ρ does.

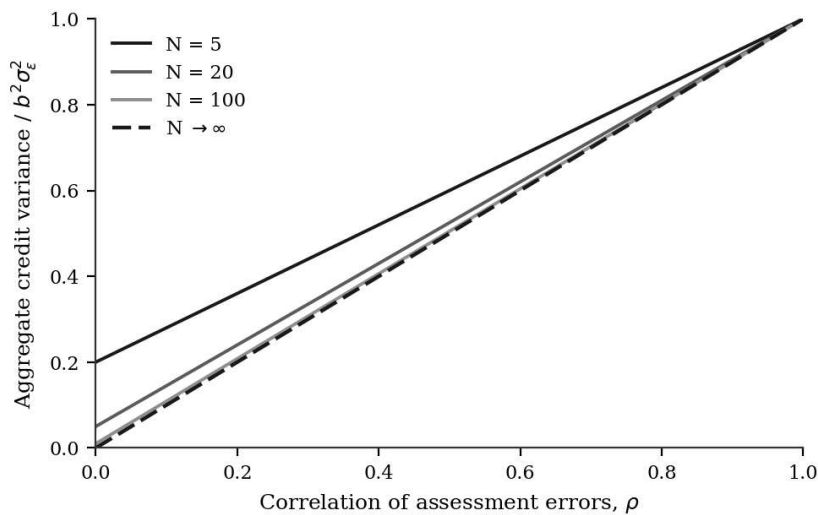


Figure 1. Cross-sectional diversification of assessment error

Notes: the vertical axis is $\text{Var}(L | \mu)$ scaled by $b^2 \sigma^2 \varepsilon$, from equation (4). The idiosyncratic term $(1 - \rho)/N$ vanishes in N , so all curves converge to the 45-degree line.

Define the systemic model error

$$S \equiv \rho \cdot \sigma^2 \varepsilon \quad (6)$$

as the variance of the common component of assessment error. Equation (5) says that aggregate credit fluctuation driven by model error is proportional to S , and to nothing else about the model technology. Meanwhile the private benefit of a better model, fewer bad loans approved and fewer good applicants rejected, depends on $\sigma^2\varepsilon$ alone, because a bank's own misallocation cost is a function of its own error variance regardless of whether other banks share it.

That gap is the whole argument. Write the social loss as

$$W = k_1\sigma^2\varepsilon + k_2\rho\sigma^2\varepsilon = \sigma^2\varepsilon(k_1 + k_2\rho) \quad (7)$$

with k_1 the per-institution misallocation cost and k_2 the marginal social cost of aggregate credit volatility. A bank minimising its own loss minimises the first term only. Its choice coincides with the social optimum if and only if $k_2 = 0$.

4.3 The adoption threshold

Suppose the system begins with heterogeneous internally built models, characterised by (σ^2_0, ρ_0) , and migrates to a shared technology characterised by (σ^2_1, ρ_1) with $\sigma^2_1 < \sigma^2_0$ and $\rho_1 > \rho_0$. Migration raises systemic model error when $S_1 > S_0$, that is when

$$\rho_1/\rho_0 > \sigma^2_0/\sigma^2_1 \quad (8)$$

P2. Migration to a shared model technology increases aggregate credit volatility whenever the proportional increase in the correlation of assessment errors exceeds the proportional reduction in their variance. That is, accuracy gains buy stability only up to the point where they are outrun by the commonality they create.

The condition is testable in principle and, more usefully, it is calibratable. Table 2 reports the break-even correlation multiple implied by a range of accuracy improvements, together with the break-even ρ_1 under an assumed baseline of $\rho_0 = 0.15$, a value chosen to reflect the correlation that already exists between independently built bank models by virtue of their shared reliance on the same credit bureau data.

Table 2. Break-even correlation under improved model accuracy (illustrative)

Reduction in error variance ($1 - \sigma^2_1/\sigma^2_0$)	Break-even ρ_1/ρ_0	Break-even ρ_1 at $\rho_0 = 0.15$
10%	1.11	0.17
20%	1.25	0.19
30%	1.43	0.21
40%	1.67	0.25
50%	2.00	0.30

Reduction in error variance ($1 - \sigma_1^2/\sigma_0^2$)	Break-even ρ_1/ρ_0	Break-even ρ_1 at $\rho_0 = 0.15$
60%	2.50	0.38

Notes: Entries are derived from condition (8) and are illustrative rather than estimated. The second column reports the ratio σ_0^2/σ_1^2 at which migration is neutral for systemic model error; the third applies that ratio to a baseline correlation of 0.15. Values of ρ_1 above the third column imply that migration raises aggregate credit volatility.

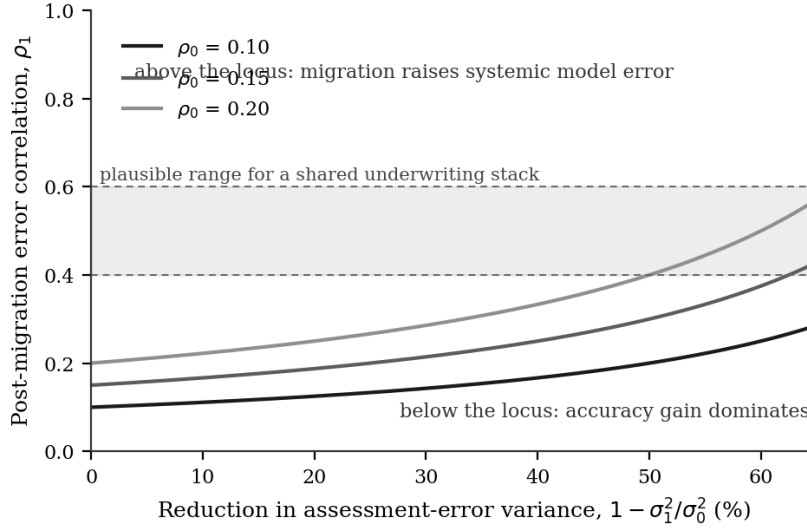


Figure 2. The break-even locus for migration to a shared model

Notes: each curve traces $\rho_1 = \rho_0/(1 - g)$, where g is the proportional reduction in assessment-error variance, from condition (8). Points above a curve satisfy $S_1 > S_0$. The shaded band marks the range of error correlations that shared training data alone generated in the experiments of Bommasani et al. (2022).

These figures are illustrative and are not estimates; no dataset currently exists from which ρ_1 could be recovered, which is the subject of Section 5. What the table shows is a matter of arithmetic rather than of assumption. Even an accuracy improvement of 50%, which would be a large gain by the standards of the machine-learning credit literature, leaves the break-even error correlation at 0.30. If a widely shared underwriting stack produces error correlations in the range that shared training data alone generated in the experiments of Bommasani et al. (2022), the threshold is crossed comfortably. The asymmetry is due to the fact that ρ is limited above by one, but the irreducible noise in borrower outcomes, $\sigma_2\epsilon$, is limited below by the noise of the borrower. There are three qualifications that are important and the last one will be most likely to be cited by a referee.

The first is that in equation (5) b is considered fixed. If the risk is the same at a greater distance down the quality distribution for better models of the bank, b will also change and the volatility comparison won't be tidy. The second is that the model doesn't have any pricing. In a complete specification, correlated tightening would also affect loan rates, which in turn would have a corresponding effect on demand, and so partially offset the effect of the quantity channel, which should be interpreted as an upper bound on the quantity channel in a complete specification. Substantive is the third qualification.

Equation (7) only attributes the volatility cost of correlation and does not attribute any benefits from correlation. If there is a shared model, but it is really better, it doesn't make the same types of mistakes in the same direction, it makes fewer mistakes, and if people are rewarded for the credit that is due, then fewer people default, which is stabilising.

If the reduction in $\sigma^2\varepsilon$ is large enough, realised loss volatility can fall even as the correlation of assessment errors rises. Whether the net effect is stabilising is an empirical question that this framework poses but does not answer, and honest reporting of it should not be deferred to a limitations section. The framework's claim is narrower than the policy discussion's: it says the sign of the effect depends on two parameters, one of which nobody currently measures.

4.4 Comparative statics along the technology frontier

Treating ρ and $\sigma^2\varepsilon$ as free parameters, as Section 4.3 does for expositional convenience, understates the problem, because the two are tied together by the technology on offer. Let the frontier take the constant-elasticity form

$$\sigma^2\varepsilon(\rho) = \sigma^2_0 (\rho_0/\rho)^\gamma, \quad \gamma \in (0, 1) \quad (9)$$

so that a one per cent increase in shared dependence buys a γ per cent reduction in assessment-error variance. The restriction $\gamma < 1$ says that accuracy improves less than proportionately with commonality, which is the empirically relevant case. The first shared component, a common credit bureau file, buys a great deal. The fifth, a common prompt template, buys almost nothing.

Substituting (9) into the social loss (7) and minimising over ρ yields an interior optimum

$$\rho^* = [\gamma/(1 - \gamma)] \cdot (k_1/k_2) \quad (10)$$

The private objective (7), with $k_2 = 0$, is strictly decreasing over the whole range, whereas the private objective is strictly decreasing in the whole range for the case of $k_2 > 0$. The answer will be a corner solution, then: to use whatever the most popular technology available is.

P3. Under a constant-elasticity frontier with $\gamma < 1$, the socially optimal degree of shared model dependence is interior and given by (10), while the privately optimal degree is the maximum available. That is, some commonality is efficient; the disagreement between private and social choice concerns how much, not whether.

Figure 3 plots both objectives for $\gamma = 0.30$. The social optimum is at $\rho^* = 0.43$ when $k_1 = k_2$, whereas a tenfold increase in the weight attached to the social cost of misallocation of systemic resources brings it to 0.21 and a tenfold decrease brings it out to 0.86, which is beyond the range of technologies plotted. The relevance for policy is the linear form of the comparative static of k_1/k_2 . Whether the authority admits it or not, an opinion as to the degree of tolerability of model commonality is, arguably, an opinion

about the relative cost of misallocated credit versus volatile credit, and this mechanism gives no indication of the ratio. It is a political decision in parameter's attire.

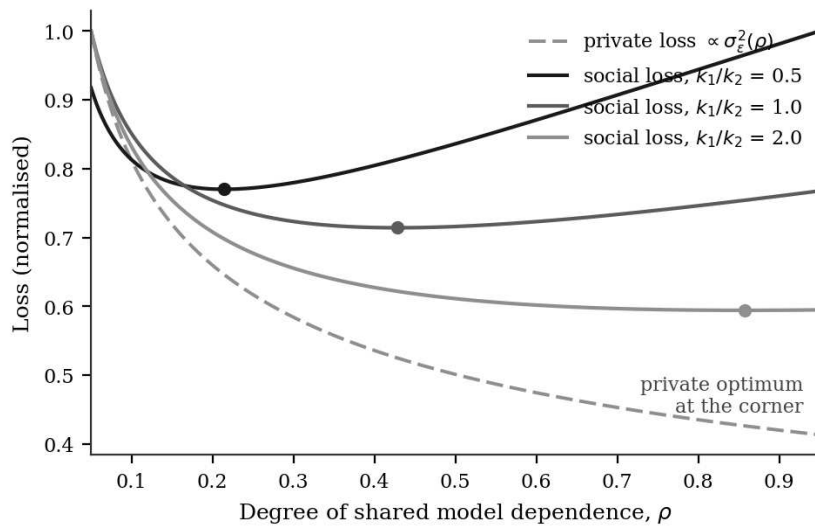


Figure 3. Private and social loss along the technology frontier

Notes: $\gamma = 0.30$, $\rho_0 = 0.15$, k_2 normalised to one; each curve scaled by its own maximum. Markers show ρ^* from equation (10). The private objective is monotonically decreasing, so the privately optimal technology is the most shared one available.

The form of the constant elasticity is only a convenience and has no empirical justification. The value of γ has never been estimated, so if it turns out that the true frontier is different, then the optimum within the interior in (10) may not exist. What remains after any reasonable specification is weaker but still useful: If the accuracy improves less than proportionately to the commonality, and the cost of the system increases proportionately, then the social optimum is interior and so is not the private optimum.

4.5 A simulated comparison

Whether condition (8) binds at parameter values anyone would actually encounter is a quantitative question, and the arithmetic of Table 2 answers it only under assumptions the table imposes. A short simulation makes the comparison concrete. Thirty banks assess a common borrower pool over 160 periods. Under the bespoke technology, $\rho = 0.10$ and $\sigma_\epsilon = 1.00$. Under the shared technology, $\rho = 0.55$, with σ_ϵ implied by the frontier (9) at $\gamma = 0.30$, which gives $\sigma_\epsilon = 0.60$ and therefore a 64% reduction in assessment-error variance. That is a deliberately generous accuracy gain, larger than any improvement reported in the machine-learning credit literature, and it is granted to the shared technology so that the comparison is not rigged.

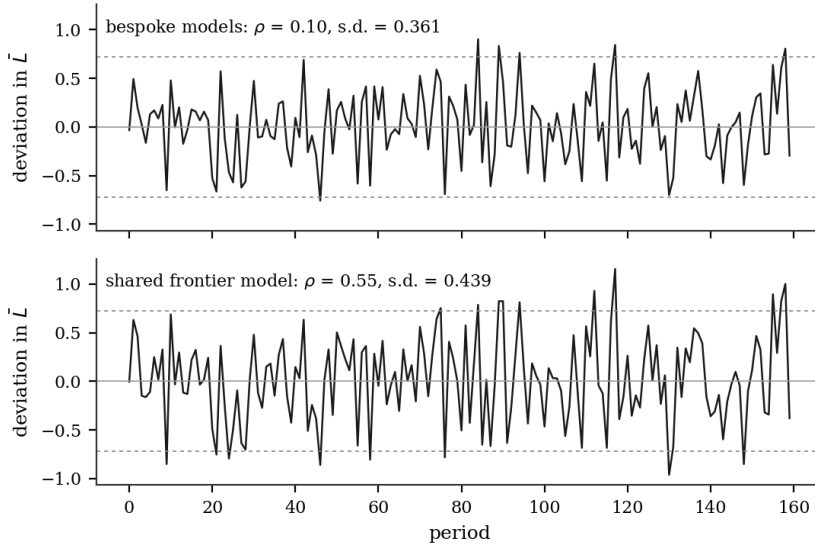


Figure 4. Simulated aggregate credit supply under bespoke and shared technologies

Notes: $N = 30$, $T = 160$, common random numbers across panels. Dashed lines mark $\pm$ two standard deviations of the bespoke series. Series are deviations of average assessed quality from its conditional mean, in units of σ_ε under the bespoke technology.

The result is in Figure 4. Aggregate credit volatility rises from 0.361 to 0.439 in the illustrated draw, a ratio of 1.22, and to 1.25 averaged over 400 replications, against 1.25 predicted analytically by equation (4) at $N = 30$. The 64% accuracy improvement does not come close to offsetting the rise in correlation, which is what condition (8) says it cannot: $\rho_1/\rho_0 = 5.5$ against $\sigma_0^2/\sigma_1^2 = 2.8$.

The tail effect is the larger one. Periods in which aggregate credit deviates by more than two bespoke-system standard deviations rise from 7.0 to 17.3 per 160, an increase of about a factor of two and a half against a volatility increase of a quarter. This is Gaussian arithmetic rather than a separate finding, and it should not be presented as one. It is nonetheless the arithmetic that matters for a supervisor, whose loss is convex in the size of a credit swing rather than linear in its variance, and it is the reason a change in ρ that looks modest in a variance ratio can look alarming in an exceedance count.

5. Measurement

5.1 What the right object is

The parameter that matters is ρ , the correlation of assessment errors. It is not the correlation of decisions. Two banks that both reject the same applicant because the applicant is genuinely uncreditworthy have agreed correctly, and their agreement carries no systemic information at all. Agreement is informative about monoculture only to the extent that it survives conditioning on the outcome, which is to say only when banks are wrong together.

This has an uncomfortable consequence. Estimating ρ requires realised outcomes, so any estimate is available only with the lag of the performance window, typically twelve to twenty-four months in

unsecured retail and considerably longer in mortgages. A supervisor who wants a contemporaneous read must work with a proxy, and every available proxy conflates skill with commonality.

The estimator itself is unremarkable once the object is defined. For a panel of borrowers with observed outcome y , let p_i denote bank i 's predicted default probability and define the residual $r_i = y - p_i$. The sample statistic is

$$R = [2 / (N(N - 1))] \cdot \sum_{i < j} \text{Corr}(r_i, r_j) \quad (11)$$

that is, R is the average pairwise residual correlation across institutions. Under the model of Section 4, it is a consistent estimator of ρ . The corresponding within-institution statistic, computed across model generations rather than across banks, gives a useful benchmark for how much of the measured commonality is attributable to the shared bureau data that all models ingest rather than to shared model components.

5.2 The reference-portfolio exercise

The data requirement, several banks scoring the same borrowers with outcomes observed, looks prohibitive until one notices that supervisors already run something very close to it. Article 78 of the Capital Requirements Directive obliges competent authorities to assess the quality of internal approaches at least annually, and directs them to pay particular attention both to approaches producing significantly different own funds requirements for the same exposure and to approaches exhibiting particularly high or low diversity. The European Banking Authority's benchmarking exercises implement this through hypothetical portfolios: a common set of exposures is distributed, institutions return the risk parameters their models produce, and the resulting dispersion is analysed and attributed.

The apparatus is therefore built. What has never been done is to run it in the other direction. For fifteen years the concern animating benchmarking was unwarranted *variability*, the suspicion that banks were using internal model discretion to reduce risk weights and that comparable exposures were attracting incomparable capital. Convergence was the objective. The Article 78(3) reference to particularly low diversity has sat in the text without an operational programme behind it, because low diversity was not a problem anyone had.

Repurposing the exercise requires four changes, none of them structural. The reference portfolio must be outcome-labelled, which means using anonymised historical exposures with realised performance rather than purely hypothetical ones, and accepting the disclosure and data-protection engineering that implies. Reporting must extend below the risk-parameter level to the model-component level, so that measured residual correlation can be attributed to a shared base model, a shared data vendor, or a shared adaptation pipeline. The reporting population must extend beyond institutions using internal ratings-based approaches, since the banks most likely to buy a common vendor stack are mid-sized standardised-approach institutions with the least capacity to build in house. And the analytical output

must invert: the flag is a dispersion statistic that is too low relative to what independent models of the same individual accuracy would generate, not too high.

Table 3 sets out the indicator family this would produce.

Table 3. Candidate indicators of model commonality

Indicator	Construction	Source	Frequency	Identification requirement
Layered dependency index	Exposure-weighted Herfindahl by stack layer, adjusted for vertical integration	Extended DORA register of information; AI inventory	Annual	None; structural
Base-model share	Share of credit decisions by volume traceable to each pre-trained model	AI inventory, extended	Annual	Requires vendors to disclose base-model provenance
Monoculture index M	Pairwise residual correlation on the reference portfolio, normalised against an independent-model null	Reference-portfolio exercise	Annual, lagged	Requires realised outcomes; ≈ 80 defaults for s.e. of 0.05
Decision agreement rate	Pairwise share of reference borrowers receiving identical approve/reject decisions at each bank's own cutoff	Reference-portfolio exercise	Annual	Contemporaneous but conflates skill with commonality
Fallback capacity	Share of monthly volume the institution could decision through a non-correlated channel within 30 days	Firm attestation, supervisory review	Annual	Verification is the binding constraint

5.3 A correction, and how much data the index needs

Equation (11) is not a consistent estimator of ρ , and the reason only becomes visible in simulation. Residuals across institutions share the realised outcome y . Expanding the covariance of two banks' residuals leaves a term in $\text{Var}(y)$ that is present whether or not the two models have anything in common, and in a low-default portfolio that term dominates everything else. The consequence is severe. In a simulated portfolio with a 5% default rate and eight banks whose models are correlated at $\rho = 0.55$, the average pairwise residual correlation is 0.95. Under the null of fully independent models of the same individual accuracy it is 0.89. The raw statistic is estimated very precisely and is almost entirely uninformative, moving by six points across the whole range of the parameter it is meant to measure.

The correction is to normalise against that null rather than against zero. Define

$$M = (R - R_{\perp}) / (1 - R_{\perp}) \quad (12)$$

where $R_{\perp}$ is the average pairwise residual correlation obtained when each institution's model is replaced by an independently drawn model of the same individual accuracy. $R_{\perp}$ is not observed but it is computable, either by simulation from institutions' reported calibration curves and discriminatory power, or by a permutation in which the idiosyncratic component of each model's error is redrawn while its accuracy is held fixed. The resulting index is zero when models are no more alike than independent models of equal skill, and one when they are identical. Call it the monoculture index. In the experiment just described it averages 0.51 against a true ρ of 0.55, attenuated but usable, and panel (a) of Figure 5 shows the three quantities side by side.

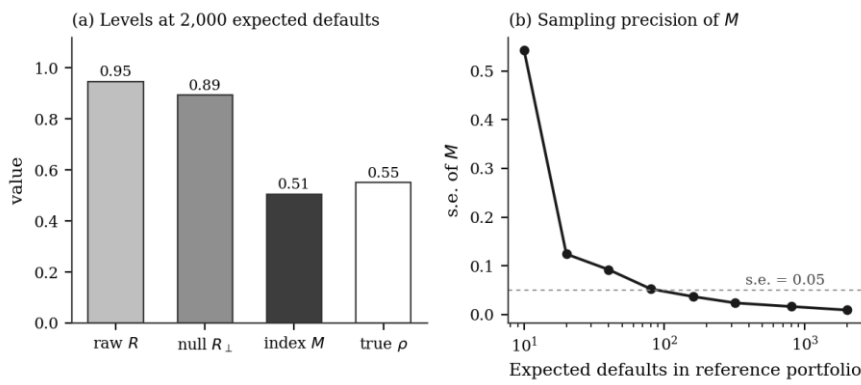


Figure 5. Bias in the raw statistic and sampling precision of the monoculture index

Notes: eight institutions, true $\rho = 0.55$, logistic outcome model, 200 replications per grid point. Panel (a) uses a portfolio of 40,000 exposures at a 5% default rate. Panel (b) varies portfolio size and default rate jointly; the horizontal axis is their product.

Panel (b) gives the sampling standard error of M against the expected number of defaults in the reference portfolio, which is the quantity that governs precision rather than the number of exposures. A standard error of 0.05 requires roughly 80 expected defaults; 0.02 requires around 400. Below twenty defaults the standard error exceeds 0.12, and at ten it reaches 0.54, which is larger than the parameter being estimated. The practical translation is unflattering and worth stating in the units a supervisor would use. A credit-card portfolio of 40,000 accounts at a 5% annual default rate yields 2,000 defaults and a standard error of 0.009. A large-corporate book of 500 names at 40 basis points yields two defaults and no usable estimate at all.

5.4 Where this fails

Three failure modes are worth stating before anyone builds on this.

The first is low-default portfolios, and Section 5.3 has already quantified it. Residual correlation is identified by outcome variation, and in wholesale, sovereign and large-corporate portfolios there are too few defaults to estimate anything. This is not a technicality that better econometrics will solve. The measure is weakest precisely where individual exposures are largest, and a supervisor relying on it

would have good visibility of unsecured retail lending and near-blindness in the segments most likely to generate a systemic event. Bounds may be constructible from rating migrations rather than defaults; the paper does not attempt it.

The second is Goodhart's law, in an unusually sharp form. A dispersion-based metric can be satisfied by adding noise. An institution instructed to demonstrate that its decisions diverge from its peers' can randomise its marginal cases, which raises measured dispersion, lowers accuracy, harms borrowers at the cutoff, and does nothing whatever to reduce the correlated failure the metric was designed to detect, because the common error component η is untouched by independent noise added on top of it. Any dispersion requirement must therefore be paired with an accuracy floor, and the pairing is what makes hard diversity mandates difficult to design well.

The third is that the reference portfolio is not the live portfolio. Banks lend to different populations; a common set of exposures measures the models, not the exposures, and a bank whose actual book is unrepresentative of the reference set may show correlation on the benchmark and something else in practice. This is a known limitation of the existing benchmarking exercises and it is inherited wholesale.

P4. The correlation of assessment errors is point-identified only on outcome-labelled portfolios with sufficient default variation, and is therefore least measurable in the exposure classes where the consequences of correlated error are largest.

6. Policy

6.1 Two instruments to resist

The instrument most often suggested in conversation, if rarely in print, is a capital add-on scaled to measured model commonality: a Pillar 2 adjustment rising in the monoculture index M or in a dependency index. The logic is Pigouvian and the analogy to systemic risk buffers is obvious. It should nonetheless be resisted for now, on two grounds. Calibration requires a parameter that Section 5 has just shown to be unmeasurable in half the relevant portfolio space, and an add-on calibrated on retail data and applied system-wide would be arbitrary in a way that supervisors would struggle to defend. More fundamentally, the add-on falls on model accuracy. Since ρ and $\sigma^2\epsilon$ move together along the technology frontier, an institution facing a charge in ρ can reduce it by adopting a worse model, and there is no reason to think the resulting reallocation improves anything.

The second instrument to resist is a diversity mandate: a requirement that institutions above some threshold not use the same base model, or that no base model exceed some share of decisions system-wide. The theoretical case is respectable and follows directly from Wagner (2011). Weak practical case. The second-best model has to be assigned to somebody, the assignment mechanism is unattractive either way, and the cost on the borrower and not the institutions. It hits, furthermore, unevenly. Previous work

on machine learning in credit finds that it can reap disproportionate benefits from better models for credit applicants who are less well-served by conventional scoring; for example, Berg et al. (2020) argue that the variables of the digital footprint provide significant predictive power for credit applicants with thin or no credit files, and Fuster et al. (2022) show that machine-learning underwriting does not benefit all borrower groups equally. The distributional effect of a diversity mandate is a choice to exclude less credit-worthy borrowers from financial markets. This still might be the correct swap. At a minimum it should be made explicitly.

6.2 What to do instead

There are three measures that survived the objections, ranked in order of preference. The first thing to report is near to free and very inexpensive. The register of information already provided under the Digital Operational Resilience Act includes information on ICT third-party arrangements – but not which pre-trained model is under a purchased underwriting engine, as the vendors are not required to say. At the very least, supervisors would have the structural layer of Table 3 at little marginal cost, as it would be sufficient to add base-model and material data-source provenance to the requirement of the register for any system used in the credit decisioning process, but the requirement would not rest on the deployer but on the provider of the system. The time frame during which the schema can be set prior to the freezes of inventories is approximately the next 12 months.

The second is a "reference-portfolio" exercise in Section 5.2. It costs more, takes legal time to work with data protection and the output comes with a delay. It's also the sole one of these proposals that takes account of what matters. It would be a sensible sequence to have it as a voluntary exercise, with results published only in aggregated form, at a small group of large retail lenders, to provide a baseline prior to its introduction. There is no denying the importance of obtaining a baseline prior to when foundation models become a significant component in the credit decision making process and no way to recoup that baseline later.

Third and more limited (narrower) than a diversity mandate: a fallback capacity requirement. If it is an institution that relies heavily on a shared external model for credit decisioning, it would need to be able to do so for a certain percentage of monthly volume, e.g., 25% over a defined period, and prove it, with a channel that does not rely on the shared external model. The channel could be a challenger device channel that is internal to the system, another vendor, or a human underwriting with lower throughputs. This directly addresses the availability failure mode, and partially the correlated-error mode without imposing a worse model on anyone for making normal-course decisions, and without the distributional cost of a diversity mandate. The weakness of this is that a fall-back that has not been verified is not worth much and to be able to verify it properly is to periodically send live volume through it. The regime would have to be forthright on that cost, instead of contenting itself with attestation.

6.3 The interaction nobody planned

There's an undesirable correlation between the regulation of AI and its centralization, and it's noteworthy because it works in a way that is difficult to turn off.

Conformity assessment, technical documentation, data governance evidence and post market monitoring and its audit trail are mostly fixed costs. They are not proportional to the number of decisions that a model can make. For a large bank, they are spread over millions of decisions, and for a mid-sized regional bank or building society, they cannot and their rational reaction is to purchase a system that can be spread over their customers by the system provider. This is further supported by the distribution of responsibilities on the value chain under the AI Act (Regulation (EU) 2024/1689), where the burden of conformity work is lighter for a deployer who relies on a provider that complies with the Regulation than for one who does it on its own. The direction of the effect is not ambiguous: raising the fixed cost of supplying a compliant high-risk model raises equilibrium concentration in the supply of such models.

P5. Where compliance costs, tightening high-risk AI requirements weakly increases concentration in model supply, and therefore weakly increases the structural precondition for decision convergence.

This is not an argument against the requirements, which address real harms in individual decisions (inaccuracy, opacity, discrimination) that no amount of systemic diversity would cure. It is an argument that the systemic and the conduct objectives are in tension, that the tension has not been acknowledged in either the AI Act's impact assessment or the prudential commentary on it, and that the deferral to December 2027 has bought exactly the time in which the interaction could be studied. Whether the effect is large enough to matter is unknown. It could be checked directly: the share of standardised-approach institutions using third-party credit-decisioning systems, tracked before and after the obligations bite, would settle it.

7. Conclusion

The claim that many institutions relying on the same models will behave alike in stress is now standard in official-sector publications, and it is almost certainly true. What has not been determined is the extent of similarity from an unknown baseline or in which segments, or why, because the amount has never been carefully quantified to calculate. It is defined in this paper as the correlation of assessment errors across institutions, demonstrated that the variance of the aggregate credit fluctuates with the product of that correlation and the variance of the individual models, and derived the condition for which the system becomes more volatile if the individual models are improved.

It's easier to get it than you think. Correlation can multiply several times over; accuracy cannot improve without bound, because borrower outcomes contain irreducible noise. That asymmetry is the paper's central analytical point, and it holds regardless of how the parameters are calibrated.

The measurement proposal is more modest than it may appear, because it repurposes rather than builds. Supervisory benchmarking under Article 78 already distributes common portfolios and analyses cross-institution dispersion; the change required is to label the portfolio with outcomes, collect at component level, extend the population beyond internal-model banks, and read the dispersion statistic in the opposite direction from the one it was designed for. The mandate to attend to particularly low diversity has been in the Directive all along.

Whether any of this is urgent depends on a fact the paper cannot establish. Foundation models accounted for 17% of AI use cases in UK financial services as recently as the 2024 joint survey, and only 2% of use cases were fully autonomous; on that evidence, monoculture in credit decisioning is prospective rather than present. The argument for acting now rests entirely on the impossibility of constructing a pre-adoption baseline after adoption. Supervisors who wait until the exposure is visible will be able to measure its level and never its increase, which is the quantity that would tell them whether to be worried.

References

- Acharya, V. V., & Yorulmazer, T. (2007). Too many to fail — An analysis of time-inconsistency in bank closure policies. *Journal of Financial Intermediation*, 16(1), 1–31.
- Adrian, T., & Shin, H. S. (2010). Liquidity and leverage. *Journal of Financial Intermediation*, 19(3), 418–437.
- Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). *Intelligent financial system: How AI is transforming finance*. BIS Working Papers No. 1194. Bank for International Settlements.
- Allen, F., Babus, A., & Carletti, E. (2012). Asset commonality, debt maturity and systemic risk. *Journal of Financial Economics*, 104(3), 519–534.
- Bank of England. (2025). *Financial Stability in Focus: Artificial intelligence in the financial system*. April.
- Bank of England & Financial Conduct Authority. (2024). *Artificial intelligence in UK financial services — 2024*. November.
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897.
- Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management*. SR Letter 11-7 / OCC Bulletin 2011-12.
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35.
- Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6), 2201–2238.

- Cecchetti, S. G., Lumsdaine, R. L., Peltonen, T., & Sánchez Serrano, A. (2025). *Artificial intelligence and systemic risk*. Reports of the Advisory Scientific Committee No. 16. European Systemic Risk Board.
- Danielsson, J. (2025). Artificial intelligence and stability. *VoxEU*, 6 February.
- Danielsson, J., & Uthemann, A. (2024). *Artificial intelligence and financial crises*. Working paper.
- Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms (Capital Requirements Directive), Article 78.
- European Banking Authority. (2024). *Handbook on supervisory benchmarking*.
- European Commission. (2025). *Proposal for a Regulation amending Regulation (EU) 2024/1689 (Digital Omnibus on Artificial Intelligence)*. 19 November.
- Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. November.
- Financial Stability Board. (2026). *Sound practices for financial institutions' responsible adoption of artificial intelligence*. Consultation report, June.
- Foucault, T., Gambacorta, L., Jiang, W., & Vives, X. (2025). *Artificial intelligence in finance*. The Future of Banking 7. CEPR Press.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.
- Greenwood, R., Landier, A., & Thesmar, D. (2015). Vulnerable banks. *Journal of Financial Economics*, 115(3), 471–485.
- Ibragimov, R., Jaffee, D., & Walden, J. (2011). Diversification disasters. *Journal of Financial Economics*, 99(2), 333–348.
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118.
- Prudential Regulation Authority. (2023). *SSI/23 — Model risk management principles for banks*. Bank of England.
- Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA).
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Wagner, W. (2010). Diversification at financial institutions and systemic crises. *Journal of Financial Intermediation*, 19(3), 373–386.
- Wagner, W. (2011). Systemic liquidation risk and the diversity–diversification trade-off. *Journal of Finance*, 66(4), 1141–1175.