Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Measuring agentic artificial intelligence intensity in banking: a validated disclosure-based index and identification strategy for estimating effects on bank performance and systemic risk

**Ramyashree M S**

ramyashree-scms@dsu.edu.in

Dayananda Sagar University

**Jayashree N**

Dayananda Sagar University

**Pavithra Salanke**

Dayananda Sagar University

**Method Article**

**Keywords:** Agentic artificial intelligence, Construct validation, Disclosure-based measurement, Bank performance, Systemic risk, Text-based indices

**Posted Date:** September 17th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10613252/v1

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

**Additional Declarations:**
No competing interests reported.

Table 1 to 4 are available in the Supplementary Files section.

# Measuring agentic artificial intelligence intensity in banking: a validated disclosure-based index and identification strategy for estimating effects on bank performance and systemic risk

**Ramyashree M S [a*], Jayashree N [b], Dr. Pavithra Salanke [c]**

[a*] Assistant Professor, School of Commerce & Management Studies (PG), Dayananda Sagar University, Bengaluru - 560068, India, ramyashreemsr@gmail.com, Ph: +91 9663002067. https://orcid.org/0009-0005-3111-1920. ramyashree-scms@dsu.edu.in,

[b] Assistant Professor, School of Commerce & Management Studies (PG), Dayananda Sagar University, Bengaluru - 560068, India, jayash21088@gmail.com, Ph: +91 9740516753, https://orcid.org/0000-0003-2850-8917, jayashreen@dsu.edu.in

[c] Associate Professor, School of Commerce and Management Studies, Dayananda Sagar University, Bengaluru- 560068, India, Pavithra.salanke@gmail.com, Ph: +91 9844964637, https://orcid.org/0000-0001-6106-3214, pavithra.s@dsu.edu.in

**Corresponding Author: Ramyashree M S,** Assistant Professor, School of Commerce & Management Studies (PG), Dayananda Sagar University, Bengaluru - 560068, India, ramyashreemsr@gmail.com, ramyashree-scms@dsu.edu.in, Ph: +91 9663002067. https://orcid.org/0009-0005-3111-1920.

## Abstract

Banks are moving from generative artificial intelligence (AI) systems that support human decision-makers towards agentic AI systems that plan, decide and execute financial transactions under delegated authority. Empirical study of this transition is currently blocked by a measurement problem: no bank-level instrument distinguishes autonomous execution from advisory or generic AI adoption, and existing disclosure-based technology indices report no reliability or validity evidence. This article develops and specifies such an instrument. We present the Agentic AI Intensity Index (AAII), a disclosure-based measure constructed from annual reports, management discussion and analysis sections, and earnings-call transcripts through a five-step protocol: theory-driven keyword-pool generation, two-round expert Delphi refinement, inter-rater reliability assessment, exploratory and confirmatory factor analysis, and large language model-assisted

classification with mandatory human adjudication. We specify acceptance thresholds at each step, an authority-weighted scoring rule that separates delegated execution from decision support, and a normalisation that isolates agentic intensity net of general AI-discussion volume. Construct validity is established through convergent, discriminant, criterion and known-groups tests against independent deployment indicators. We then set out the identification machinery required to estimate AAII effects on efficiency, risk detection, operational risk and market reaction, combining two-way fixed effects, event studies, staggered-adoption difference-in-differences, instrumental variables, system generalised method of moments, matching and synthetic control, together with explicit mitigations for AI-washing and strategic disclosure. The article reports the method rather than estimates and is intended as a directly executable protocol for researchers and supervisors.

**Keywords:** Agentic artificial intelligence; Construct validation; Disclosure-based measurement; Bank performance; Systemic risk; Text-based indices

## 1 Introduction

### 1.1 The measurement problem

Artificial intelligence (AI) in banking has passed through three broad stages: rule-based automation, generative copilots that draft and summarise for a human decision-maker, and, from approximately 2025, agentic systems that plan and execute multi-step financial workflows with limited human intervention [1, 2]. The third stage is not distinguished by the quality of the model, but by who has the authority. An agentic system does not make a recommendation for approval, but acts under a delegated mandate, and the actions can be irreversible once completed.

Supervisory bodies have taken note of this evolution. The International Monetary Fund describes specific authorisation, liquidity and compliance risks associated with autonomous payment execution and argues for containment mechanisms precisely because finalised agentic payments cannot easily be unwound [3]. The Bank for International Settlements records central banks' use of AI for its own purposes but recognises the supervisory challenge of evaluating systems whose internal logic is not directly observable [4]. The applied work has begun to model multi-agent credit-risk configurations with formal model-risk-management oversight [5].

What's missing is measurement. Empirical claims about agentic banking, that it reduces costs, that it increases operational risk, that markets price it, all assume a bank-level quantity that does not

now exist in usable form. The absence has a particular shape. It is worth stating exactly because it determines what this article adds.

First, no available instrument separates autonomous execution from AI adoption in general. Existing constructs treat AI or digitalisation as an undifferentiated whole [6, 7, 8]. A bank that has deployed a customer-service chatbot and a bank that has delegated payment release to an autonomous agent score similarly on such measures, although the risk and efficiency implications differ categorically.

Second, disclosure-based technology indices report essentially no psychometric evidence in this literature. Term-frequency indices constructed from annual reports [6, 8] are based on keyword lists, the derivation, inter-coder agreement and dimensional structure of which are not documented, leaving open the question of whether an independent research team applying the stated procedure would recover the same scores. However, in adjacent literatures in economics, text-based measures are often explicitly validated against external benchmarks [9, 10, 11]. Bank technology measurement has not adopted this standard.

Third, the few studies that do address AI and bank outcomes rely on fixed effects alone for identification, leaving the obvious reverse-causality channel — better-capitalised banks can afford more AI — unaddressed.

### 1.2 Contribution and article type

This is a Methodology article. It presents a new measurement instrument and the identification apparatus required to use it, rather than empirical estimates obtained from applying it. Three contributions follow.

**A validated construct.** We define agentic AI intensity as the extent to which a bank has delegated consequential financial decision *and execution* authority to autonomous systems, and we operationalise it through the AAII. The AAII is different from previous disclosure-based indices in that it is defined along with a full validation protocol: Delphi-based item refinement. Inter-rater reliability with specified acceptance levels. Exploratory & confirmatory factor analysis. 4-part construct validity assessment (convergent, discriminant, criterion, known-groups). Section 6 describes this protocol in its entirety since it is the methodological contribution, not a supporting detail.

**Scoring derived from theory, not atheoretical scoring:**

The dimensions of the index as well as their weights are based on Dynamic Capabilities Theory [12] and the Technology–Organisation–Environment (TOE) framework [13], and not just on keyword frequency. This is operationally relevant: it defines what disclosure language is evidence of autonomy, and it justifies the authority weighting in Section 5.3.

**An identification strategy matched to the construct.** We specify a design combining two-way fixed effects, event studies, staggered-adoption difference-in-differences, instrumental variables, system generalised method of moments (GMM), propensity-score matching and synthetic control. These are deliberately overlapping because each rests on a different identifying assumption, and none is individually decisive in this setting.

The fragmentation this article addresses is not only academic. Practitioner surveys report large aggregate cost savings from agentic AI, but these figures are rarely benchmarked against the same institution's operational-risk disclosures or fair-lending record within a single framework. A bank's investor materials may highlight a reduction in onboarding time from agentic know-your-customer automation while its risk-committee disclosures separately record a rise in automated-decisioning complaints, with no analysis linking the two. A single, continuously measured, validated construct is what makes joint testing of those claims possible.

### 1.3 Structure

Section 2 develops the theoretical foundations and the conceptual model. Section 3 reviews the 2020–2026 literature across ten thematic areas and states the resulting measurement gap. Section 4 develops four hypotheses and one consolidated moderation proposition. Section 5 specifies the construction of the AAII. Section 6 specifies the validation protocol, including acceptance thresholds. Section 7 describes sample design and variables. Section 8 sets out the identification strategy. Section 9 addresses disclosure-specific measurement threats. The expected results and their interpretation are given in Section 10. Section 11 limitations, Section 12 implications and conclusion. Appendix A includes the full definitions of variables, Appendix B the pool of seed keywords, and Appendix C an excerpt from the coding manual used for reliability assessment.

## 2 Theoretical foundations

### 2.1 Theory of Dynamic Capabilities

According to Dynamic Capabilities Theory, sustained advantage is derived from a firm's capacity to sense opportunities, take them by reconfiguring resources and transform organisational structures [12]. Here, the AAII is not just a proxy for technology adoption. It proxies the bank's ability to sense the agentic opportunity, to seize it through capital and talent reallocation, and to reconfigure risk-governance functions commensurate with the delegated authority it has created.

There are two implications for the specification. First, efficiency gains are seen as conditional on organisational transformation capacity, not mechanical; hence the governance quality and AI maturity are moderators rather than controls. Second, the theory explains why two banks with the same AAII scores can perform very differently. The technology signal the index captures is necessary but not sufficient. The organisational capacity to transform around that signal is what makes the difference in performance in terms of autonomy.

This also motivates treating AI-enabled adaptive response as a dynamic capability rather than a technology attribute [14] and corresponds to the framing of AI as both a productivity technology and a priced risk factor [15] – a duality that must be evaluated jointly rather than sequentially.

### 2.2 The Technology-Organisation-Environment Framework

The TOE framework sees adoption as a process that is shaped by technological, organisational, and environmental context [13]. The control specification is directly motivated: technology capital expenditure, institution size and digital-channel share capture the technological and organisational dimensions. More importantly, it encourages us to consider the regulatory environment as a moderator rather than a control, as the same level of AAII may have different implications under a stringent regime than under a permissive regime.

TOE also justifies the inclusion of fintech competitive pressure and exposure to central bank digital currency (CBDC) as environmental controls. Both plausibly influence a bank's incentive to seek agentic autonomy independent of its internal capacity, and their omission would place environmental variation on the AAII coefficient.

### 2.3 Why you need both theories

No theory alone would justify the moderated specification developed in Section 4. Dynamic Capabilities accounts for variation through internal organizational capacity, TOE accounts for variation through external environmental context. The common application is why H1 to H3 are considered conditional relationships, not unconditional average effects. A specification that includes only one of the two would either treat regulatory regime as a nuisance parameter or treat governance quality as an exogenous background, and in both cases would misattribute conditional variation to the main effect.

### 2.4 Conceptual model

Figure 1 presents the resulting framework. AAII, derived from the two theories, is hypothesised to affect four outcome domains — efficiency, risk detection, operational risk and market reaction — each conditioned by five organisational and environmental moderators.

Reading the figure from left to right, the two theoretical anchors feed into the central AAII construct, expressing the claim that AAII is a theory-derived measurement rather than an atheoretical keyword count. The four outcome boxes correspond to H1–H4. The moderator band is the consolidated H5 proposition. It is positioned as conditioning the AAII-to-outcome paths rather than AAII itself, because the moderators are hypothesised to govern how agentic intensity converts into outcomes, not to determine the level of agentic intensity a bank chooses to pursue.

**Fig. 1** Conceptual framework linking Agentic AI Intensity (AAII) to four outcome domains. Dashed-border elements are the theoretical anchors, Dynamic Capabilities Theory and the Technology–Organisation–Environment framework. The heavy-border element is the focal construct. Solid arrows denote hypothesised direct effects H1–H4. Dashed arrows denote the moderating effects of the H5 moderator set, shown in the dotted-border band, which conditions the H1–H3 paths. Border and line style rather than colour distinguish element types

## 3 Related work and the measurement gap

This section synthesises the literature from 2020-2026 across ten themes. Related themes are brought together to keep the review analytical rather than enumerative. Each sub-section is concluded with the points of agreement, the points of disagreement and the specific implication for the design that follows.

### 3.1 Transition to assistance to authority

Policy sources concur that banks are moving from advisory AI to systems with governed execution authority. The BIS describes central banks' own use of AI while highlighting the supervisory challenge [4]; the IMF describes authorization, liquidity, and compliance risks specific to autonomous execution and advocates containment mechanisms in light of the irreversibility of finalized agentic payments [3]. Both present the problem as organizational and legal as well as technical: who is responsible if an autonomous agent executes an erroneous or fraudulent transaction and how quickly can a human supervisor intervene once execution authority has been delegated.

Agreement: The transition is real, it is happening. Disagreement: whether the current governance arrangements are sufficient. Implication: any design based on AAII requires a moderator of governance quality. A uniform treatment effect is not defensible.

### 3.2 Big data, machine learning and credit intermediation

Big-data variables improve consumer credit assessment [16]. Hybrid models combining logistic and tree methods outperform both purely machine learning and linear models, while still being interpretable [17]. Machine-learning underwriting improves aggregate accuracy but exacerbates disparities in access [18], and this finding is echoed in analyzes of lending discrimination in the fintech era [19]. These results carry over to emerging market and peer-to-peer settings, where ensemble methods regularly outperform simpler classifiers at some cost to transparency for the affected borrower [20, 21].

Agreement: Machine learning improves overall accuracy but often at the cost of disparate impact. Unresolved: none of these studies consider the autonomous execution of the resulting decision. Implication: H2 has to be coupled with a fairness sub-hypothesis, H2a, and tested together.

### 3.3 Systemic and operational risk from AI

The speed and correlated decision-making of artificial intelligence can amplify systemic risk under stress [22], and formal models show how artificial intelligence can accelerate crisis transitions once institutions move from profit-maximising to survival-maximising behaviour [23]. Another similar argument is that governance has not caught up with deployment, with ethical and risk-

management frameworks in fintech being reactive rather than proactive [24]. No side denies the productivity gains of AI. Both sides deny that these gains can be assessed in isolation from the correlated-behaviour and governance-lag risks that arise from near-simultaneous adoption across many institutions facing similar conditions.

There is consensus that AI is a real risk. Disagreement: mechanism, correlated models vs. lag of governance. Implication: H3 is separate from, and possibly conflicting with H1, and the two must be estimated in one framework.

### 3.4 Robo-advisors and algorithmic asset management

Personalised robo-advising has been modelled as an interactive learning between platform and client [25]. Access to purely automated advice appears to be driven by account-minimum frictions rather than investor characteristics [26]. A more skeptical strand highlights that in some configurations, automated rebalancing harms performance [27], a precursor of the autonomy-risk trade-off which this article investigates. The global performance and fee structures of robo-managers are heterogeneous [28], and a systematic review notes that this heterogeneity has been mainly examined at retail rather than institutional level [29], leaving open how the same dynamics unfold when the deploying institution is a universal bank rather than a standalone platform. Human advisors are error-prone themselves [30], which provides the right baseline: comparing autonomous systems to a hypothetical perfect human advisor would systematically underestimate any real benefit of autonomy.

All sides agree automation can increase access. Disagreement as to whether it is autonomy or merely lower fees that causes the improvement. Implication. H3a, workflow heterogeneity, extends the finding that autonomy risk is concentrated in certain functions [27].

### 3.5 Bank performance and digital transformation

The evidence here is contradictory on its own terms. European data indicate a positive but non-linear relationship between digital transformation and profitability [6]; Chinese panel data indicate the opposite, with reduced net interest margins and return on assets [31]. To reconcile the two, a digitalisation-paradox mechanism associated with managerial myopia has been proposed [32]. Studies on emerging markets find positive effects operating thru non-interest income [33, 34]. AI-

powered adaptive response has been described as a performance strategy consistent with dynamic capability [14], supporting the theoretical framing in Section 2.1.

The nearest methodological antecedent of the present work is the text-based bank digital transformation index built from annual-report term frequency [8]. Crucially, research on financial intermediation and technological change warns that some digital-transformation claims are not associated with genuinely new capabilities, but are instead rebranded existing processes [35]. That distinction has direct bearing on the AI-washing concern addressed in Section 9.

None. Positive, negative and paradox-mediated findings coexist. Implication: H1a explicitly hypothesises non-linearity instead of assuming a simple positive coefficient.

### 3.6 Fraud detection and RegTech

One of the rising regulatory demands is explainability in machine-learning fraud detection reviews [36, 37]. Federated, explainable architectures target the balance of secrecy versus transparency [38]. The trade-off between predictive performance and explainability however, remains [39]. For example, RegTech is claimed to need careful governance in compliance automation [40]. Automation gains are often offset by new model risk and vendor-concentration risk [41], particularly when many institutions depend on a small number of underlying AI-platform vendors. That concentration, in turn, could be a source of correlated, system-wide operational risk if a single vendor's model fails across many client banks simultaneously.

Automation improves detection and compliance efficiency. Sustained tension: it presents explainability and vendor-concentration risks not fixed by automation**.** Implication: AAII text classification must be human-validated rather than automatically trusted (Section 6.4), and vendor concentration is recorded as a limitation (Section 11).

### 3.7 Central bank digital currencies and payment infrastructure

Models of CBDC design trade-offs generally conclude that design choices — remuneration, holding limits, interoperability — determine whether issuance meaningfully disintermediates commercial-bank deposit funding [42, 43, 44, 45]. Most directly relevant, event-study evidence on CBDC-related disintermediation finds that US market reactions differ according to deposit-franchise exposure [46], and broader CBDC-news market effects are documented across markets

[47]. Reviews of blockchain CBDC architectures describe the payment infrastructure within which agentic systems would ultimately operate, noting that permissioned-ledger designs offer greater control at the cost of single-point-of-failure and vendor-dependency risks analogous to those in Section 3.6 [48].

Agreement: changes in digital-payment infrastructure produce substantial and heterogeneous bank-level market reactions. *Implication:* the H4 event-study design and the H4a heterogeneity sub-hypothesis are adapted directly from these designs [46, 47].

### 3.8 Large language models and the emerging agentic-AI literature

Domain-specific pretraining improves financial natural language processing performance [49], and adoption frameworks weigh pretrained, fine-tuned and custom-trained options against data, compute and performance constraints [50]. A significant caveat is that generative-model-derived financial sentiment is vulnerable to adversarial manipulation [51]. Building on this base, AI has been cast as both productivity technology and priced risk factor [15]; large language models are argued to have reached genuine multi-step agentic capability [2]; agentic AI has been characterised as the next automation frontier for banking, distinguished from rule-based systems by the ability to plan, execute and self-correct multi-step workflows [1]; multi-agent credit-risk configurations have been modelled under explicit model-risk-management oversight [5]; and risk-management frameworks are argued not to have kept pace with deployment [52].

Agreement: domain-specific models outperform general ones, and agentic capability is a real and recent technical threshold. *Caveats:* AI-generated signals can be manipulated [51] and governance lags deployment [52]. *Implication:* this is the newest and thinnest literature, and the specific gap the AAII addresses. All model-assisted classification is validated against human coding (Section 6.4).

### 3.9 Algorithmic trading and efficiency measurement

Analysis of AI-driven trading demonstrates that AI agents can reach tacitly collusive outcomes absent explicit coordination [53] — the closest available analogue to this article's autonomy-risk hypothesis. On measurement, machine learning has been combined with data envelopment analysis (DEA) to evaluate bank efficiency and predict institutional stress, using several hundred branches as decision-making units [54, 55]. Methodological work advances machine-learning-

DEA hybrids further, notably into high-dimensional settings through LASSO-based variable selection [56, 57], which is directly relevant given the large control set the present panel specification requires. The underlying DEA formulations are the constant-returns and variable-returns models [58, 59].

Agreement: autonomous execution agents generate correlated, hard-to-detect risks distinct from advisory-AI risks; machine-learning-augmented DEA outperforms either method alone. *Implication:* DEA efficiency enters as an alternative outcome to the cost-to-income ratio in H1.

### 3.10 Financial inclusion and AI in emerging markets

AI-based underwriting can widen credit access for populations traditional banks find unprofitable to serve, since automated, low-marginal-cost decision-making removes the fixed per-application labour cost that makes small-value lending uneconomic under conventional underwriting [60]. This access-expansion argument stands in productive tension with the disparate-impact findings of Section 3.2: the same automation that lowers the cost of serving previously excluded borrowers can, if poorly designed, encode and scale biases present in historical training data.

Agreement: this is a necessary counterweight to the risk-oriented literatures above. *Implication:* H2 and H3 must be interpreted together, with distributional effects held in view.

### 3.11 The measurement gap

Read together, these literatures establish three things. AI and machine-learning adoption has well-documented efficiency and predictive benefits, frequently accompanied by fairness and systemic-risk costs. Adjacent literatures — digital-transformation indices, machine-learning-DEA hybrids, CBDC event studies — supply directly reusable measurement and identification templates. And the emerging agentic-AI literature names autonomy as distinct from general AI adoption but has not produced a validated, theoretically grounded, bank-level measure linked to outcomes under a credible endogeneity strategy.

Table 1 compares the AAII with the nearest available disclosure-based constructs. Two gaps are visible. First, no prior index separates explicitly autonomous-execution language from advisory or generic AI-adoption language; each either captures AI adoption as an undifferentiated whole [7] or captures digitalisation more broadly without an autonomy-specific lens [6, 8]. Second, no prior

index reports formal reliability or validity statistics of the kind specified in Section 6, leaving unresolved whether its classification scheme would replicate under independent coding.

The AAII's contribution is therefore incremental in construct definition — isolating autonomy specifically — and foundational in measurement rigour, supplying the validation apparatus this literature has so far lacked.

**Table 1** Comparison of the Agentic AI Intensity Index with existing disclosure-based technology constructs

MD&A management discussion and analysis EFA exploratory factor analysis CFA confirmatory factor analysis α Krippendorff's alpha κ Cohen's kappa

## 4 Assumptions and propositions

Based on Sections 2 and 3, four hypotheses and one consolidated moderation proposition are derived. They are all constructed to be testable against a particular documented empirical pattern rather than a general expectation of a positive or negative AAII coefficient. Table 2 shows the mapping of each estimator to its outcome variable, estimator and rejection criterion.

### 4.1 H1: Effectiveness

H1 Banks with higher AAII have lower cost-to-income ratios and higher DEA efficiency scores, holding constant size, share of digital channels, and prior technology investment. H1a. This is a non-linear and time-lagged relationship consistent with the sensing-seizing-transforming sequence [12] and conflicting evidence on digital transformation [32, 31]: there is a short-run cost increase during deployment before any long-run efficiency gain.

A rejection of H1, or an effect limited to H1a's short-run dip and no corresponding long-run gain, would be informative in itself.

It would indicate that agentic deployment functions more as costly signalling than as a genuine efficiency lever, consistent with the digitalisation-paradox mechanism [32].

### 4.2 H2: risk detection

*H2.* Banks with higher AAII exhibit lower fraud and anti-money-laundering (AML) false-positive rates.

*H2a.* Consistent with documented fairness costs of algorithmic underwriting [19, 18], these gains are accompanied by disparate impacts across borrower subgroups.

H2 and H2a are tested jointly rather than sequentially, to avoid reproducing the aggregate-accuracy-only framing that Section 3.2 identifies as a recurring limitation of the credit-scoring literature.

### 4.3 H3: autonomy risk

*H3.* Banks with higher AAII experience more frequent or more severe disclosed operational-risk events, consistent with Section 3.3 and with the algorithmic-collusion mechanism [53].

*H3a.* Consistent with evidence that automated rebalancing erodes performance in specific configurations [27], autonomy-risk effects concentrate in particular workflow types rather than distributing evenly. A bank-level average AAII may therefore understate risk concentrated in a small number of high-autonomy workflows such as payment execution, even where overall AAII appears moderate.

### 4.4 H4: market reaction

*H4.* Equity markets generate abnormal returns in response to credible agentic — as distinct from generative — AI announcements.

*H4a.* As documented for CBDC news [47], reactions vary with a bank's disintermediation-sensitive business lines: transaction-banking-heavy institutions differ from those with larger relationship-lending franchises.

### 4.5 H5: consolidated moderation proposition

*H5.* Organisational transformation capacity [12] and environmental context [13] jointly moderate the AAII-to-outcome relationships in H1–H3, operationalised across five dimensions: governance quality, AI maturity stage, regulatory stringency, digital capability and human-oversight ratio. Efficiency and risk-detection benefits (H1, H2) strengthen, and autonomy-risk costs (H3)

attenuate, as governance quality, AI maturity, digital capability and human oversight increase, and as the regulatory regime becomes more stringent.

Rather than five independently labelled sub-hypotheses, H5 is operationalised as a single standardised moderator index (Section 6.6) and, secondarily, through five individual interaction terms reported as a robustness decomposition (Section 8.5). This avoids the multiple-comparison inflation that five separately tested interaction hypotheses would introduce, while preserving the ability to see whether any result is driven by one dimension.

**Table 2.** Hypothesis-to-test mapping

IV instrumental variables, *GMM* generalised method of moments, *PSM* propensity-score matching, *CS-DiD* Callaway–Sant'Anna difference-in-differences, *CAR* cumulative abnormal return, *FE* fixed effects

## 5 Constructing the Agentic AI Intensity Index

### 5.1 Construct definition and design principles

We define agentic AI intensity as the extent to which a bank has delegated consequential financial decision and execution authority to autonomous computational systems operating without transaction-level human approval.

Three elements of this definition do work. *Consequential* excludes autonomy over trivial internal processes. *Decision and execution* is the operative distinction from generative AI, which produces outputs for human action. This identifies the boundary exactly without transaction-level human approval: a system that prepares a payment instruction for release by an officer is not agentic under this definition, while one that releases it within a mandate is.

Four design principles emerge from sections 2 and 3.

1. Authority, not activity. Scoring weights disclosure language by degree of delegated authority evidenced not by frequency of mention of AI (Section 5.3).

2. Difference, not absolute level. The index is defined against a generative-AI contrast dictionary, so that a bank that talks a lot about AI but only in advisory terms scores low (Section 5.4).

3. Specificity as evidence. Quantified workflow-named claims are more reliable than unquantified promotional language, following the logic of greenwashing detection [61] (Section 9.2).

4. Validate before use. No score is entered into an estimation sample until the acceptance thresholds in Section 6.7 are reached.

### 5.2 The source corpus

The corpus consists of three document classes per bank-year: the annual report including the management discussion and analysis (MD&A) section or its jurisdictional equivalent; regulatory filings, i.e. Form 10-K for US registrants and Pillar 3 plus annual report for EU institutions; and earnings-call transcripts for all quarters in the year, covering both prepared remarks and the analyst question-and-answer segment.

Q and A part is kept purposely. Investor relations prepares the remarks, but answers to unscripted questions from analysts are less rehearsed and have higher information content per word in the greenwashing-detection literature [61]. In Section 9.2, we use the gap between them as a diagnostic.

Sentences are segmented and only those containing at least one seed term (Appendix B) are retained for classification. The burden of human adjudication in Section 6.4 is kept tractable without throwing away candidate evidence because, by construction, a sentence that does not contain a seed term cannot be scored higher than zero.

### 5.3 Weighting dimension and authority

The AAII is made up of five dimensions that are based on the theoretical framework rather than inductive clustering of terms. These are listed in Table 3 with their authority weights.

**Table 3 Theoretical derivation and factor weights of AAII dimensions**

DCT Dynamic Capabilities Theory, TOE Technology–Organisation–Environment framework

Weights are theory-derived priors, not free parameters. The highest weights are given to D1 and D2 , where the construct definition differentiates the delegated execution and the binding decision authority. D3-D5 represent the transformation and the environmental conditions under which the authority is exercised. Section 6.3 tests whether the empirical factor structure is consistent with

this allocation, and Section 8.5 reports an equal-weights alternative as a robustness check. Should the confirmatory factor analysis reject the five-dimension structure, the weights are re-derived from standardised loadings rather than retained by assertion; this contingency is stated in advance to avoid post-hoc weight selection.

Within each dimension, each retained sentence *s* is assigned an authority tier:

- **Tier 3 (score 1.00)** — explicit delegated execution, past or present tense, with a named workflow. Example construction: a system that *executes* or *releases* within stated limits.
- **Tier 2 (score 0.60)** — autonomous decision without execution, or execution under case-level review.
- **Tier 1 (score 0.30)** — planned, piloted or prospective autonomy; future tense or conditional mood.
- **Tier 0 (score 0.00)** — advisory, assistive, generative or human-in-the-loop framing; seed term present but authority absent.

Tier 0 is essential. It is what prevents the index from rewarding banks that discuss AI at length in purely advisory terms, and it is the mechanism by which the generative-contrast requirement of Section 5.1 is enforced at sentence level.

### 5.4 Aggregation and normalisation

For bank *i* in year *t*, the raw dimension score is the authority-weighted sum of tier scores over retained sentences in that dimension:

$$R_{i,t,d} = \sum_{s \in S_{i,t,d}} \tau_s$$

where $\tau_s \in \{0, 0.30, 0.60, 1.00\}$ is the tier score of sentence *s*.

Raw counts are not comparable across banks, because disclosure documents differ greatly in length and because banks differ in how much they discuss AI at all. Two successive normalisations address this.

**Length normalisation.** Divide by total disclosure length $L_{i,t}$ in thousands of words, giving an intensity per unit of disclosure:

$$N_{i,t,d} = \frac{R_{i,t,d}}{L_{i,t}/1000}$$

**AI-baseline normalisation.** Divide by the bank's overall AI-mention rate $A_{i,t}$, the count per thousand words of any AI-related term, whether agentic or not:

$$\widetilde{N}_{i,t,d} = \frac{N_{i,t,d}}{A_{i,t} + \kappa}$$

The constant $\kappa = 1$ prevents division by near-zero for banks with minimal AI discussion. This second step is what distinguishes the AAII from a general AI-adoption proxy. Without it, the index would largely recover the general AI-adoption construct already studied elsewhere [7], and its incremental contribution would be nil. Section 6.5 tests this directly: the discriminant-validity requirement is that the normalised AAII correlate only moderately with a general AI-mention index.

The composite index is the weighted sum of dimension scores, rescaled to the unit interval across the estimation sample:

$$\mathrm{AAII}_{i,t} = \frac{\sum_{d=1}^{5} w_d \, \widetilde{N}_{i,t,d} - \min(\cdot)}{\max(\cdot) - \min(\cdot)}$$

Rescaling uses sample minimum and maximum computed on the pooled sample and held fixed thereafter, so that year-on-year changes in a bank's score reflect changes in its disclosures rather than changes in the peer distribution.

### 5.5 Workflow-level disaggregation

H3a requires variation below the bank level. Each Tier 2 or Tier 3 sentence is therefore additionally tagged to one of six workflow classes: payments and settlement; credit adjudication; fraud and AML; customer onboarding and KYC; regulatory reporting; and treasury and liquidity management. This yields a workflow-level AAII vector alongside the composite, allowing the test of whether autonomy risk concentrates in payment execution as H3a predicts. Sentences that cannot be assigned to a single class are tagged *unclassified* and excluded from the disaggregated analysis while remaining in the composite.

## 6 Validation protocol

This section specifies the five-step protocol referred to in Section 5.1, together with the acceptance thresholds that govern whether the resulting scores may be used. It constitutes the article's principal methodological contribution, and is reported in the main text rather than as an appendix for that reason.

### 6.1 Step 1: keyword-pool generation

An initial pool of approximately 60 candidate terms is generated from three sources: the agentic-AI-specific literature [1, 2, 52, 5]; a generative-AI contrast set drawn from the large language model literature [51, 50, 49], used to define Tier 0 rather than to score positively; and existing digital-transformation index vocabularies [6, 8], used to identify terms that must be *excluded* as insufficiently autonomy-specific.

This third source is as important as the first. Terms such as *digital transformation*, *automation* and *machine learning* appear in agentic disclosures but are not diagnostic of autonomy, and their inclusion would collapse the AAII into the constructs it is meant to improve upon. The seed pool, exclusion set and contrast set are given in Appendix B.

Terms are generated by two researchers working independently against the same source set, with the union taken forward to Step 2. Union rather than intersection is used at this stage because Delphi refinement is designed to remove weak candidates, and premature exclusion cannot be recovered later.

### 6.2 Step 2: two-round expert Delphi refinement

The candidate pool is refined through a two-round Delphi process following established procedure for information-systems research [62].

**Panel composition.** Twelve to fifteen experts, drawn to include at least four academic researchers in banking technology or financial econometrics, at least four practitioners holding model-risk, technology-risk or AI-governance roles in banks, and at least two supervisory or central-bank staff. Panel size in this range is standard for information-systems Delphi work and balances judgement diversity against attrition risk. Practitioners and supervisors are included because the distinction between delegated and supervised execution is an operational one on which academic panellists alone would have limited direct knowledge.

**Round 1.** Panellists rate each candidate term on a seven-point scale for (i) diagnosticity of autonomous execution and (ii) resistance to promotional use, and assign it to one of the five dimensions in Table 3 or mark it unassignable. Free-text addition of missing terms is invited.

**Round 2.** Panellists receive the anonymised distribution of Round 1 ratings together with their own response, and re-rate. Terms are retained where the median diagnosticity rating is $\geq 5$ and the interquartile range is $\leq 2$, the conventional Delphi consensus criterion. Terms added by three or more panellists in Round 1 enter Round 2 as new candidates.

**Dimension assignment.** A term is assigned to the dimension receiving $\geq 70\%$ of panellist allocations. Terms not reaching this threshold are either split into more specific variants and returned for a supplementary round, or dropped.

Panel composition, round-level response rates, attrition and the full pre- and post-Delphi term lists are reported as Supplementary Information, so that the derivation is auditable rather than asserted.

### 6.3 Step 3: reliability assessment and dimensional structure

**Inter-rater reliability.** Three trained coders independently code a stratified random sample of at least 600 sentences — 120 per dimension, stratified across bank size, jurisdiction and year — using the coding manual excerpted in Appendix C. Agreement is assessed on the tier assignment, which is ordinal, using Krippendorff's alpha [63] with ordinal difference weighting, and pairwise using Cohen's kappa [64].

Acceptance requires $\alpha \geq 0.80$ for the tier assignment and $\alpha \geq 0.75$ for dimension assignment; kappa is interpreted against the conventional benchmarks of Landis and Koch [65], with values below 0.61 treated as insufficient. Where a threshold is not met, the coding manual is revised, coders retrained, and a fresh sample drawn — the previously coded sample is not reused, since re-coding the same sentences after discussion inflates agreement.

**Internal consistency.** Cronbach's alpha [66] is computed across the five dimension scores, with $\alpha \geq 0.70$ required for the composite to be treated as a single scale. A materially lower value would indicate that the dimensions do not share sufficient common variance to justify aggregation, in which case dimensions are reported separately rather than combined.

**Dimensional structure.** Exploratory factor analysis is conducted on the dimension scores using principal-axis factoring with oblique (promax) rotation, oblique rotation being appropriate because the dimensions are theoretically expected to correlate [67]. Sampling adequacy is assessed by the Kaiser–Meyer–Olkin measure [68], requiring $\geq 0.70$, and factorability by Bartlett's test of sphericity [69] at $p < 0.05$. Factor retention follows parallel analysis rather than the Kaiser eigenvalue-greater-than-one rule, which over-retains.

Confirmatory factor analysis on a held-out half-sample then tests the five-dimension structure against a single-factor alternative, with acceptance at comparative fit index $\geq 0.90$, root mean square error of approximation $\leq 0.08$ and standardised root mean square residual $\leq 0.08$. The sample is split before exploratory analysis, not after, so that the confirmatory test is genuinely out-of-sample.

### 6.4 Step 4: model-assisted classification with human adjudication

Sentence-level tier and dimension assignment at corpus scale is performed with large language model assistance under a fixed, published prompt, applied identically to every sentence. Given documented vulnerability of model-derived financial text signals to manipulation and instability [51], model output is treated as a proposal requiring validation, never as a final score.

The validation design is as follows.

**Calibration set.** A stratified random sample of 1,000 sentences is coded independently by humans and by the model. Model-human agreement is assessed with the same statistics and thresholds as Section 6.3. Deployment requires $\alpha \geq 0.80$ against the human gold standard. Where this is not met on the first attempt, the prompt is revised and a fresh calibration sample drawn; the number of revision rounds is reported, since undisclosed prompt iteration against a fixed sample is a form of overfitting.

**Full-corpus adjudication**. Every sentence the model assigns to Tier 2 or Tier 3 is reviewed by a human coder, because these are the sentences that generate index variation and therefore drive every downstream estimate. Tier 0 and Tier 1 assignments are audited on a 10% random sample. Where audit disagreement exceeds 10%, the full Tier 0–1 set is reviewed.

**Adjudication of disagreement**. Model-human disagreements are resolved by a second human coder blind to both prior assignments. Disagreement rates are reported by dimension, by jurisdiction and by year, since a rate that varies systematically by jurisdiction would indicate that disclosure-language differences rather than autonomy differences are driving classification.

**Stability.** The full classification is repeated with a fixed random seed and, separately, with a paraphrased prompt of equivalent content. Rank correlation between the resulting bank-year indices is reported, with $\rho \geq 0.95$ required. This addresses the concern that model-derived measures may not replicate under trivially different implementations.

### 6.5 Step 5: construct validity

Reliability establishes that the measure is reproducible; validity establishes that it measures agentic autonomy rather than something correlated with it. Four assessments are specified, following standard multitrait practice [70, 71].

**Convergent validity.** AAII is correlated with independent indicators of agentic deployment not derived from the same text: the count of AI or machine-learning engineering job postings mentioning autonomous-agent or orchestration skills; patent filings classified to autonomous decision systems; and disclosed vendor contracts for agentic platforms. Expected correlation $\rho \geq 0.50$. These are imperfect indicators individually, which is why three are used and convergence across them, rather than any single correlation, is the criterion.

**Discriminant validity.** AAII is correlated with a general AI-mention index constructed from the same corpus without authority weighting. The requirement is a *moderate* correlation, $0.30 \leq \rho \leq 0.70$. A correlation above 0.70 would indicate the index has not separated autonomy from AI discussion generally and would trigger a return to Section 5.3 weighting; a correlation below 0.30 would suggest the index is capturing something unrelated to AI, indicating a classification fault. Both bounds are stated in advance.

**Criterion validity.** AAII is evaluated against known agentic-deployment events confirmed by independent sources: regulatory sandbox admissions for autonomous systems, supervisory disclosures and vendor announcements confirmed by the bank. The requirement is a large positive step change in year of confirmed deployment.

**Known-groups validity:** Banks identified by supervisory or industry sources as agentic-AI leaders should score significantly higher than comparable size and jurisdiction non-leaders. This is the weakest of the four tests, because the independent classifications are themselves partially derived from the disclosure, and it is reported as corroborative rather than conclusive.

### 6.6 Construction of the moderator

The five H5 moderators are constructed as follows and standardised to mean zero and `unit variance, prior to entering interaction terms:

Governance quality is coded on a 0- 4 scale based on disclosures: (1) existence of a board-level technology or AI committee, (2) existence of a model-risk function with explicit authority over autonomous systems, (3) disclosed independent validation of agentic models, and (4) disclosed protocol for incident escalation with stated intervention time.

AI maturity stage, 1–4 coded from disclosed deployment history: pilot only, single-workflow production, multi-workflow production, enterprise-wide orchestration.

The stringency of regulation is measured at the jurisdiction-year level by the relevant regime, primarily the staged rollout of the EU AI Act and the Digital Operational Resilience Act (coded 1–3).

**Digital capability** is the digital-channel transaction share, taken from disclosures where reported and from supervisory statistics otherwise.

**Human-oversight ratio** is the disclosed ratio of risk and compliance full-time equivalents to the count of production agentic workflows. Where the denominator is not disclosed, the workflow count from Section 5.5 is substituted, and this substitution is flagged in the observation.

The composite moderator index is the mean of the five standardised components. Because the components are conceptually distinct rather than reflective indicators of a single latent trait, the composite is treated as a formative index, and no internal-consistency requirement is imposed on it. This is stated explicitly because applying Cronbach's alpha to a formative index is a common error.

### 6.7 Acceptance thresholds

Table 4 consolidates the thresholds that must be met before AAII scores enter any estimation sample. The purpose of stating them in advance is to remove discretion at the point where the analyst can already see the results.

**Table 4** Validation acceptance thresholds

*EFA* exploratory factor analysis, *CFA* confirmatory factor analysis, *CFI* comparative fit index, *RMSEA* root mean square error of approximation, *SRMR* standardised root mean square residual

## 7 Sample design and variables

### 7.1 Sample construction

The estimation sample comprises global systemically important banks (G-SIBs) together with a matched set of regional and challenger banks, observed annually and, for the event study, at disclosure-date frequency, from 2020 through 2026. The period splits into a 2020–2024 pre-inflection baseline and a 2025–2026 agentic-deployment period. Following established practice in the digital-transformation panel literature [6], banks with fewer than three consecutive years of disclosure data are excluded.

A target of at least 150 banks across five or more years, giving 750 or more bank-year observations, is specified as the minimum for adequately powered fixed-effects estimation. This target is derived from effect sizes documented in comparable digital-transformation studies [6, 31, 8]: at the smaller end of that reported range, and with within-bank variation of the magnitude the 2025–2026 inflection is expected to generate, 750 observations is approximately what is required to detect the H1 coefficient at conventional power. Because that calculation depends on an assumed effect size, the realised minimum detectable effect should be reported alongside the estimates rather than treated as adequate by assumption.

Including regional and challenger banks alongside G-SIBs is deliberate. A G-SIB-only sample would confound AAII variation almost entirely with institution size, since large-scale agentic deployment has so far concentrated among the largest and best-resourced institutions (Section 3.8). Including smaller institutions with varying degrees of adoption, even if imperfectly matched on every dimension, is what identifies the AAII coefficient separately from a pure size effect. It is

also why the matching design in Section 8.4 matches explicitly on pre-adoption size and capital ratio rather than treating size as a linear control.

### 7.2 Variables

Appendix A gives full definitions and sources. In summary: efficiency outcomes use accounting ratios with a DEA score as robustness; risk outcomes use operational-loss and fraud or AML data; market outcomes use event-study abnormal returns; moderators are coded per Section 6.6; and controls comprise size, capital adequacy, digital-channel share, technology capital expenditure, fintech competitive pressure and CBDC exposure.

The two instruments used in Section 8.3 — regional AI-talent density and distance-weighted peer-bank AAII — are constructed separately from this control set and are deliberately excluded from it, since including an excluded instrument as a regressor destroys its identifying role.

## 8 Identification strategy

### 8.1 Baseline panel specification

The baseline is a bank-and-year two-way fixed-effects model:

$$Y_{i,t} = \beta\,\mathrm{AAII}_{i,t} + \delta\,(\mathrm{AAII} \times M)_{i,t} + \Gamma'\mathrm{Controls}_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t}$$

where $Y$ is the cost-to-income ratio or DEA score (H1), a fraud or AML proxy (H2), or an operational-risk event count (H3), and $M$ is the composite moderator of Section 6.6. Standard errors are clustered at bank level. A squared AAII term and distributed lags test H1a. For the count outcome in H3, a negative binomial fixed-effects specification replaces the linear model, since operational-risk events are non-negative integers with variance exceeding the mean.

Bank fixed effects absorb time-invariant institutional characteristics — business-model mix, historical risk culture — that would otherwise confound the AAII coefficient when pooling across institutions. Year fixed effects absorb sector-wide shocks common to all banks in a year, such as the interest-rate cycle or the phased rollout of the EU AI Act.

### 8.2 Event study

For H4, cumulative abnormal returns (CARs) are estimated over [−1, +1] and [−5, +5] windows around agentic-disclosure dates using a market model estimated over a [−250, −30] window, following standard event-study procedure [72] and extending prior applications to banking-technology announcements [46]. CARs are interacted with deposit-franchise exposure to test H4a.

The two windows serve different purposes. The narrow window captures the immediate reaction net of confounding news. The wider window allows for the market taking several days to process implications, which is particularly relevant where a disclosure is embedded in a longer earnings-call transcript rather than issued as a standalone announcement — a distinction recorded for every event and used as a sub-sample split.

The essential comparison for H4 is not against zero but against a matched sample of *generative-only* AI announcements by the same banks. Absent that comparison, a positive CAR would be uninformative as to whether markets distinguish assistance from authority, which is the actual question.

### 8.3 Staggered adoption and endogeneity

**Staggered-adoption difference-in-differences.** Where adoption timing is sufficiently staggered, the Callaway and Sant'Anna estimator [75] compares banks crossing an AAII threshold against not-yet-treated banks. This avoids the negative-weighting and heterogeneous-treatment-timing bias that affects naive two-way fixed-effects designs under staggered adoption [73, 74], which is the expected pattern here given the uneven pace of agentic rollout documented in Section 3.8. Treatment is defined as crossing the sample median AAII; robustness to alternative thresholds at the 60th and 75th percentiles is reported, since threshold choice is otherwise a researcher degree of freedom.

**Reverse causality.** Better-performing or better-capitalised banks may simply have more resources to invest in agentic AI. Four remedies are specified.

**Instrumental variables.** Regional AI and machine-learning job-posting density, and a distance-weighted peer-bank AAII, instrument for own-bank AAII. First-stage strength is assessed against the weak-instrument critical values of Stock and Yogo [76], and over-identification tested where both instruments are used jointly.

**System GMM.** Lagged levels and differences serve as internal instruments to address dynamic endogeneity [77, 78], implemented with the Windmeijer finite-sample correction [80] and with instrument-count restriction to avoid the proliferation that renders the Hansen test uninformative [79]. Both the Hansen *J* and the second-order autocorrelation test are reported, along with the instrument count.

**Propensity-score matching.** Banks are matched on pre-adoption covariates — principally size, capital adequacy and prior technology capital expenditure — using the propensity score [81], with covariate balance reported before and after matching.

**Synthetic control**. For individual large-bank cases, a synthetic counterfactual is constructed from a donor pool of non-adopting banks following Abadie et al. [82], providing a transparent single-unit comparison where the panel estimators are least informative.

These remedies are deliberately overlapping rather than alternatives. Each rests on a different identifying assumption, and none is beyond challenge in this setting. The instrumental-variable approach requires that regional AI-talent supply and peer-bank AAII affect outcomes only through a bank's own AAII — an exclusion restriction that cannot be directly tested and is discussed as a limitation in Section 11. System GMM relies instead on the weaker but still non-trivial assumption that idiosyncratic outcome shocks are not serially correlated beyond the specified lag structure. Matching relies on selection on observables, which will not hold perfectly but reduces bias along the dimensions where early and late adopters most differ. Reporting all four allows a reader to judge whether the AAII-outcome relationship is robust across differing identifying assumptions or fragile to any one of them.

### 8.4 Robustness checks

Index of placebo The specification is re-estimated with a generic machine-learning and artificial intelligence index that excludes agentic-specific terms. A significant coefficient on the placebo would suggest that the effect is simply capturing general AI enthusiasm already addressed by existing constructs [7] rather than agentic intensity per se.

This is the single most informative robustness check in the design.

**Alternative efficiency measure.** A machine-learning-DEA hybrid efficiency score [55, 57] replaces the accounting cost-to-income ratio, addressing the concern that cost-to-income ratios are affected by one-off accounting items unrelated to operational efficiency.

**Sub-sample splits.** Results are split by bank size and by jurisdiction (EU against US), testing whether findings are driven by the EU AI Act's differentiated treatment of high-risk use cases or generalise across regimes.

**Weighting.** The theory-derived weights of Table 3 are replaced with equal weights and with weights derived from standardised confirmatory factor loadings.

**Moderator decomposition.** The composite H5 moderator is decomposed into its five interaction terms, allowing assessment of whether results are driven evenly across dimensions or concentrated in one or two — most plausibly governance quality and human oversight, given the theoretical emphasis of Section 2.1. Multiple-comparison adjustment is applied to this decomposition, since it involves five simultaneous interaction tests.

### 8.5 Pre-specification

Because the AAII involves substantial construction discretion, we recommend that the outcome models, thresholds and robustness set be registered before the index is linked to outcome data. The classification work of Section 6 can be completed and reported independently of any outcome variable, and doing so removes the possibility that index-construction choices are made, however unconsciously, with reference to the estimates they produce. This is a design feature available here that is unavailable in most observational settings, and it should be used.

## 9 Disclosure-based measurement: threats and mitigations

The AAII is built from public disclosures and is therefore exposed to strategic disclosure, promotional AI language and cross-jurisdictional differences in disclosure practice. These are treated as design constraints requiring explicit mitigation rather than as caveats to be noted and set aside.

### 9.1 Triangulation against non-disclosure indicators

Third-party indicators — AI job postings, patent filings, disclosed vendor contracts — are triangulated against the AAII, with material divergence flagged for review. A bank with a sharp rise in AAII without a corresponding rise in agentic-relevant hiring and vendor contracting is flagged. This reduces dependence on a single disclosure channel that an investor-relations function might optimise for tone rather than accuracy.

### 9.2 Specificity scoring

A specificity score distinguishes quantified, verifiable claims from unquantified promotional language, following the logic of greenwashing detection in the ESG disclosure literature [61]. A bank stating a specific percentage reduction in onboarding time attributable to a named workflow scores differently from one making general claims of AI leadership. Two diagnostics follow: the ratio of specific to general claims is reported per bank-year, and the gap between prepared remarks and unscripted question-and-answer language (Section 5.2) is reported as a supplementary indicator, since promotional framing is harder to sustain under unscripted questioning.

### 9.3 Jurisdictional disclosure norms

Jurisdiction fixed effects and jurisdiction-split robustness checks address the fact that disclosure verbosity and specificity norms differ materially between, for example, US 10-K filings and EU annual-report conventions, independently of underlying deployment. The classification-disagreement rates reported by jurisdiction in Section 6.4 provide a direct diagnostic of whether this is operating.

### 9.4 Longitudinal consistency

Implausible year-on-year AAII discontinuities unsupported by corresponding capital-expenditure or headcount disclosures are flagged, on the reasoning that genuine agentic-capability build-out is unlikely to jump discontinuously in ways the underlying investment record does not support. Flagged observations are retained but reported separately, and results are re-estimated excluding them.

None of these mitigations eliminates disclosure-based measurement error. Each is intended to bound it and to make its likely direction and magnitude visible, rather than to support a claim that the resulting index is measurement-error-free. The expected direction of the residual error is worth

stating: if promotional disclosure is more prevalent among banks with weaker actual deployment, the AAII will be attenuated towards the mean and the H1 and H3 coefficients biased towards zero, making the design conservative with respect to those hypotheses and anti-conservative with respect to a null finding.

## 10 Anticipated findings and their interpretation

We anticipate that agentic AI intensity is associated with measurable efficiency and risk-detection gains (H1, H2) while also being associated with a higher incidence of disclosed operational-risk events (H3). Such a pattern would refine the practitioner narrative, which reports efficiency gains without documenting the accompanying autonomy-related risk.

If H1 and H3 are jointly supported, banks face a genuine trade-off: efficiency gains that are, in part, a transfer from operating costs to contingent, later-disclosed risk costs. H5 then tests whether that trade-off is mitigable through governance quality and human oversight, or is an inescapable feature of agentic deployment. Any aggregate H2 gain must be checked against H2a's disparate-impact test before being read as an unambiguous improvement.

For H4, the informative question is not whether abnormal returns are positive but whether they are distinguishable from reactions to generative-only disclosures, and whether they vary with deposit-franchise exposure as documented for CBDC news [47]. A null finding — markets not yet pricing the assistance-to-authority distinction — would itself be meaningful, indicating a gap between operational and regulatory reality and current analyst practice.

The H5 proposition may reshape the interpretation of H1–H3 rather than functioning as a peripheral robustness check. If efficiency and risk-detection gains cluster heavily among high-governance, high-maturity, high-oversight banks while autonomy-risk costs cluster among their low-governance counterparts, this would indicate that agentic AI's net effect on the banking system depends less on the technology than on the organisational and regulatory apparatus built around it. That interpretation aligns with Dynamic Capabilities Theory's emphasis on organisational capacity as the binding constraint [12], and carries a materially different policy implication — favouring governance-quality and oversight-ratio standards over blanket restrictions on agentic deployment — than a finding in which moderation effects are weak and average effects dominate regardless of context.

Taken as a whole, the anticipated pattern is not that agentic AI is good or bad for banks, but that it is conditional: efficiency and risk-detection benefits that are real but neither automatic nor unconditional; an autonomy-risk cost that is real but concentrated and governable rather than diffuse and inescapable; and a market that may or may not yet be pricing any of this correctly. These remain hypothesised directions. Establishing which conditional pattern actually holds is what the strategy in Section 8 is designed to test once the data are assembled.

## 11 Limitations

**Selection towards large adopters**. Agentic deployment at scale remains concentrated among institutions with the resources to fund it. The matching and instrumental-variable designs address this only partially, and estimates should be read as most reliable for the size range actually represented in the sample.

**Residual measurement error**. Despite the Section 6 protocol and the Section 9 mitigations, disclosure-based text measures capture only what banks choose to disclose, which may lag or diverge from internal deployment for competitively sensitive systems such as fraud and AML models — precisely the systems most relevant to H2.

**Data availability.** Fraud, AML and operational-loss variables are frequently proprietary or supervisory-confidential. Testing H2 and H3 is likely to require a smaller set of banks with voluntary disclosure, reducing power relative to H1 and H4. This is the most binding practical constraint on executing the design.

**Untestable exclusion restriction**. The instrumental-variable strategy requires that regional AI-talent density and peer-bank AAII affect bank outcomes only through own-bank AAII. Regional talent density may plausibly correlate with regional economic conditions that affect bank performance directly, and peer-bank AAII may reflect common shocks to a competitive market. Neither channel can be ruled out by test, which is why system GMM and matching are reported alongside rather than as supplements.

**Residual endogeneity**. Unobserved time-varying confounders — a leadership change simultaneously driving both AAII and performance — are mitigated but not eliminated by instruments and matching, absent a randomised rollout unavailable in this setting.

**Regulatory heterogeneity**. The EU AI Act's phased application and the Digital Operational Resilience Act's incident-reporting requirements may shift disclosure patterns independently of true deployment. Jurisdiction controls separate genuine regulatory effects from disclosure-practice effects only partially.

**Vendor concentration**. The design measures bank-level autonomy but not the concentration of underlying vendor platforms across banks. If many institutions depend on a small number of agentic-AI vendors, the relevant systemic-risk unit may be the vendor rather than the bank, and a bank-level measure will not capture it (Section 3.6).

**Construct scope**. The AAII measures disclosed delegated authority, not the quality or safety of the underlying systems. Two banks with identical AAII scores may have very different actual control environments; the governance moderator addresses this only to the extent that control quality is itself disclosed.

## 12 Future research, implications and conclusion

### 12.1 Extensions beyond banking

The AAII framework and its identification strategy extend to adjacent domains, each requiring its own construct-validation cycle rather than direct transplant of the banking dictionary.

In **insurance**, agentic claims-processing and underwriting raise analogous efficiency-versus-autonomy-risk questions, with the added complexity of actuarial fairness regulation and the long-tail, delayed-discovery nature of claims risk, which may require a longer post-adoption observation window than the panel structure proposed here.

In **capital markets**, agentic execution and market-making connect directly to the algorithmic-collusion literature [53] and could be studied with the same event-study and difference-in-differences apparatus, substituting trading-venue-level operational-risk disclosures for bank-level measures.

In **fintech and non-bank lending**, agentic underwriting at firms unconstrained by bank capital regulation may show sharper efficiency-risk trade-offs, offering a natural comparison for testing whether the regulatory-stringency moderator generalises outside the banking perimeter.

In **central banking**, agentic tools for supervision and policy communication [4] merit their own disclosure-based measure, distinct from the commercial-bank construct given central banks' different disclosure norms and objectives.

Finally, **cross-country comparison** contrasting EU AI Act jurisdictions against lighter-touch environments would allow the regulatory-stringency moderator to be tested as a between-country design rather than a within-sample control, strengthening identification of the regulatory channel and providing external validity beyond the US- and EU-weighted sample this design implies.

### 12.2 Theoretical and practical implications

Rather than developing a new vocabulary for agentic banking, this article provides a theoretical re-positioning of AI adoption in banking as a continuum of delegated authority, based on Dynamic Capabilities Theory and the TOE framework, with robo-advisory and digital-transformation constructs nested as special cases. This nesting allows for the import of hypotheses from adjacent literatures—disparate impact (Section 3.2), autonomy eroding performance [27], non-linear efficiency effects (Section 3.5)—as testable sub-hypotheses rather than as unrelated findings from a different technology era. It also suggests that the productivity-and-risk duality of AI [15] should be thot about together and not in sequence.

In practice, a bank with a high AAII score but low governance quality is a candidate for the model-risk-management interventions discussed in the agentic credit-risk literature [5]; a high AAII, low-specificity-score bank is a candidate AI-washing case (Section 9.2). The AAII is a disclosure-based instrument, providing a scalable, non-intrusive supervisory monitoring tool, which does not require access to proprietary model architectures – addressing a proportionality concern implicit throughout the Reg Tech literature (Section 3.6). For investors, the H4 and H4a results have direct valuation implications if markets have not yet priced the difference between aid and authority: analysts incorporating bank-specific AAII trends would then have an informational advantage over those relying on generic AI-adoption narratives.

### 12.3 Conclusion

This article has identified a validated, theory-derived instrument to measure agentic AI intensity in banking and the identification machinery needed to estimate its effects. The measurement contribution is the five step construction and validation protocol of Sections 5 and 6, with stated

acceptance thresholds at every step -- an apparatus that existing disclosure-based technology indices have not provided. The identification contribution is a design of panel fixed effects, event studies, staggered-adoption difference-in-differences, instrumental variables, system GMM, matching and synthetic control that addresses reverse causality with multiple independent identifying assumptions rather than one.

The article reports rather than specifies empirical estimates, and this is done on purpose. The protocol is intended to be directly executable by other researchers on newly assembled data, and it is separable: the classification and validation work of Section 6 constitutes a self-contained methodological output that can be completed, reported and audited before any outcome variable is introduced. That separation is what makes pre-specification feasible here, and it is the strongest available safeguard against a measure whose construction is shaped by the results it produces.

**Declarations**

**Funding.** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests.** The authors declare no competing financial or non-financial interests in relation to the work described.

**Ethics approval.** Not applicable. This methodological study involved no human participants, surveys, questionnaires, interviews, experiments, or collection of personal data. The article is based on published literature and publicly available secondary sources; ethics approval was therefore not required. Should the Delphi panel described in Section 6.2 be convened in subsequent empirical work, institutional ethics review will be sought for that stage and expert participants will provide informed consent, as the panel constitutes research involving human participants.

**Consent to participate.** Not applicable. No human participants were involved in the preparation of this article.

**Consent for publication.** Not applicable. This article contains no data relating to any identifiable individual.

**Data availability.** All data generated or analysed during this study are included in this published article and its supplementary information files. Specifically, the seed keyword pool, exclusion set and generative-contrast set are provided in Appendix B; the coding manual used for reliability

assessment is provided in Appendix C; and full variable definitions together with their prospective sources are provided in Appendix A. No bank-level dataset was generated or analysed, because the article specifies a measurement and identification protocol prior to its execution. Upon execution of the protocol in subsequent work, the full dictionary, coding manual, classification prompt and resulting bank-year index values will be deposited in a public repository, subject to the licensing terms of the underlying commercial disclosure data.

**Code availability.** Not applicable to the present article, as no code was executed to produce results reported here. Classification and estimation code will be deposited in the same public repository as the data, upon execution of the protocol.

**Author contributions. Ramyashree M S:** conceptualisation, methodology, investigation, writing — original draft, writing — review and editing. **Jayashree N:** conceptualisation, investigation, theoretical framework development, writing — original draft, writing — review and editing. **Pavithra Salanke:** conceptualisation, methodology, investigation, hypothesis development, writing — original draft, writing — review and editing. All authors contributed to development of the conceptual framework, critically reviewed and revised the manuscript, and approved the final version.

Use of generative AI. SELECT ONE BEFORE SUBMISSION — (a) The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript; or (b) The authors used [tool name] for [language editing/reference formatting] during manuscript preparation and take full responsibility for the content of the published article. The use of large language models described in Section 6.4 is a research instrument within the proposed measurement protocol and is distinct from the preparation of this manuscript.

**References**

1. Kalyanasundaram, S., Balasubramanian, G., & Jayaram, R. (2026). Agentic AI: The new engine of cost-optimized banking. *The Management Accountant Journal, 61*(4), 37–43.

2. Korinek, A. (2025). *AI agents for economic research* (NBER Working Paper No. 34202). National Bureau of Economic Research.

3. International Monetary Fund. (2026). *How agentic AI will reshape payments* (IMF Notes No. 2026/004).

4. de Araujo, D. K. G., Doerr, S., Gambacorta, L., & Tissot, B. (2024). *Artificial intelligence in central banking* (BIS Bulletin No. 84). Bank for International Settlements.

5. Okpala, I., Golgoon, A., & Kannan, A. R. (2025). Agentic AI systems applied to tasks in financial services. *arXiv*. https://arxiv.org/abs/2502.05439

6. Citterio, A., King, T., & Locatelli, R. (2024). Is digital transformation profitable for banks? Evidence from Europe. *Finance Research Letters, 70*, 106269. https://doi.org/10.1016/j.frl.2024.106269

7. Singh, R. K., Mishra, R., Kumar, S., & Bag, S. (2025). Applications of artificial intelligence for optimizing banking and financial operations: A systematic literature review and future research agenda. *Journal of Economic Surveys, 39*(5), 2284–2302. https://doi.org/10.1111/joes.12693

8. Zhu, Y., & Jin, S. (2025). The impact of bank digital transformation on the financial performance of commercial banks. *SAGE Open*.

9. Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics, 131*(4), 1593–1636.

10. Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics, 134*(4), 2135–2202.

11. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.

12. Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal, 18*(7), 509–533.

13. Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington Books.

14. Sullivan, Y., & Wamba, S. F. (2024). Artificial intelligence and adaptive response to market changes: A strategy to enhance firm performance and innovation. *Journal of Business Research, 174*, 114500.

15. Eisfeldt, A. L., & Schubert, G. (2024). *AI and finance* (NBER Working Paper No. 33076). National Bureau of Economic Research.

16. Jiang, J., Liao, L., Lu, X., Wang, Z., & Xiang, H. (2021). Deciphering big data in consumer credit evaluation. *Journal of Empirical Finance, 62*, 28–45.

17. Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192.

18. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5–47.

19. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56.

20. Ozgur, O., Karagol, E. T., & Ozbugday, F. C. (2021). Machine learning approach to drivers of bank lending: Evidence from an emerging economy. *Financial Innovation, 7*, 20.

21. Teply, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance, 51*, 100904.

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

23. Danielsson, J., & Uthemann, A. (2025). Artificial intelligence and financial crises. *Journal of Financial Stability, 80*, 101453.

24. Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L. T., Vodenska, I., & Trajanov, D. (2022). Ethically responsible machine learning in fintech. *IEEE Access, 10*, 97531–97554.

25. Capponi, A., Olafsson, S., & Zariphopoulou, T. (2022). Personalized robo-advising: Enhancing investment through client interaction. *Management Science, 68*(4), 2485–2512.

26. Rossi, A. G., & Utkus, S. P. (2024). Who benefits from robo-advising? Evidence from machine learning. *Journal of Financial Economics*.

27. Horn, M., & Oehler, A. (2020). Automated portfolio rebalancing: Automatic erosion of investment performance? *Journal of Asset Management, 21*(6), 489–505.

28. Helms, N., Hölscher, R., & Nelde, M. (2022). Automated investment management: Comparing the design and performance of international robo-managers. *European Financial Management, 28*(4), 1028–1078.

29. Cardillo, G., & Chiappini, H. (2024). Robo-advisors: A systematic literature review. *Finance Research Letters, 62*, 105119. https://doi.org/10.1016/j.frl.2024.105119

30. Linnainmaa, J. T., Melzer, B. T., & Previtero, A. (2021). The misguided beliefs of financial advisors. *The Journal of Finance, 76*(2), 587–621.

31. Zhang, S. (2026). The influence of digital transformation on Chinese bank profitability performance. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0340249

32. Guo, X., Li, M., Wang, Y., & Mardani, A. (2023). Does digital transformation improve the firm's performance? From the perspective of digitalization paradox and managerial myopia. *Journal of Business Research, 163*, 113868.

33. Do, T. D., Pham, H. A. T., Thalassinos, E. I., & Le, H. A. (2022). The impact of digital transformation on performance: Evidence from Vietnamese commercial banks. *Journal of Risk and Financial Management, 15*(1), 21.

34. Nguyen, Q. T. T., & Ho, L. T. H. (2023). Digitalization and bank profitability: Evidence from an emerging country. *International Journal of Bank Marketing*.

35. Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). Fintech: What's old, what's new? *Journal of Financial Stability, 53*, 100836. https://doi.org/10.1016/j.jfs.2020.100836

36. Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review, 40*, 100402.

37. Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences, 12*(19), 9637.

38. Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management, 18*(4), 179.

39. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*(8), 216.

40. Papantoniou, A. (2022). RegTech: Steering the regulatory spaceship in the right direction? *Journal of Banking and Financial Technology, 6*, 1–16.

41. Teichmann, F., Boticiu, S. R., & Sergi, B. S. (2023). RegTech — Potential benefits and challenges for businesses. *Technology in Society, 72*, 102150.

42. Agur, I., Ari, A., & Dell'Ariccia, G. (2022). Designing central bank digital currencies. *Journal of Monetary Economics, 125*, 62–79.

43. Andolfatto, D. (2021). Assessing the impact of central bank digital currency on private banks. *The Economic Journal, 131*(634), 525–540.

44. Barrdear, J., & Kumhof, M. (2022). The macroeconomics of central bank digital currencies. *Journal of Economic Dynamics and Control, 142*, 104148.

45. Kumhof, M., & Noone, C. (2021). Central bank digital currencies — Design principles for financial stability. *Economic Analysis and Policy, 71*, 553–572.

46. Beckmann, L., Debener, J., Hark, P. F., & Pfingsten, A. (2024). *CBDC and the shadow of bank disintermediation: US stock market insights on threats and remedies* [Working paper].

47. Wang, Y., Lucey, B. M., Vigne, S. A., & Yarovaya, L. (2022). The effects of central bank digital currencies news on financial markets. *Technological Forecasting and Social Change, 180*, 121715.

48. Sethaput, V., & Innet, S. (2023). Blockchain application for central bank digital currencies (CBDC). *Cluster Computing, 26*(4), 2183–2197.

49. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv*. https://arxiv.org/abs/2303.17564

50. Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the 4th ACM International Conference on AI in Finance* (pp. 374–382). Association for Computing Machinery.

51. Leippold, M. (2023). Sentiment spin: Attacking financial sentiment with GPT-3. *Finance Research Letters, 55*, 103957.

52. Ogundimu, O. (2025). Artificial intelligence in financial and operational risk management. *Cognizance Journal of Multidisciplinary Studies, 5*, 484–506.

53. Dou, W. W., Goldstein, I., & Ji, Y. (2025). *AI-powered trading, algorithmic collusion, and price efficiency* [Wharton School research paper]. University of Pennsylvania.

54. Appiahene, P., Missah, Y. M., & Najim, U. (2020). Predicting bank operational efficiency using machine learning algorithm: Comparative study of decision tree, random forest, and neural networks. *Advances in Fuzzy Systems, 2020*, 8581202.

55. Kumar, A., et al. (2022). Performance evaluation of Indian banks using feature selection data envelopment analysis: A machine learning perspective. *Journal of Public Affairs, 22*(4), e2686. https://doi.org/10.1002/pa.2686

56. Chen, Y., Tsionas, M. G., & Zelenyuk, V. (2021). LASSO+DEA for small and big wide data. *Omega, 102*, 102419.

57. Zhu, N., Zhu, C., & Emrouznejad, A. (2021). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies. *Journal of Management Science and Engineering, 6*(4), 435–448.

58. Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078–1092.

59. Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.

60. Kshetri, N. (2021). The role of artificial intelligence in promoting financial inclusion in developing countries. *Journal of Global Information Technology Management, 24*(1), 1–6.

61. Yu, E. P., Luu, B. V., & Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance, 52*, 101192.

62. Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management, 42*(1), 15–29.

63. Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

64. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

65. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

66. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

67. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

68. Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.

69. Bartlett, M. S. (1954). A note on the multiplying factors for various chi-square approximations. *Journal of the Royal Statistical Society: Series B, 16*(2), 296–298.

70. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

71. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.

72. MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature, 35*(1), 13–39.

73. Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.

74. Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.

75. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.

76. Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models* (pp. 80–108). Cambridge University Press.

77. Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies, 58*(2), 277–297.

78. Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics, 87*(1), 115–143.

79. Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal, 9*(1), 86–136.

80. Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics, 126*(1), 25–51.

81. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

82. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association, 105*(490), 493–505.

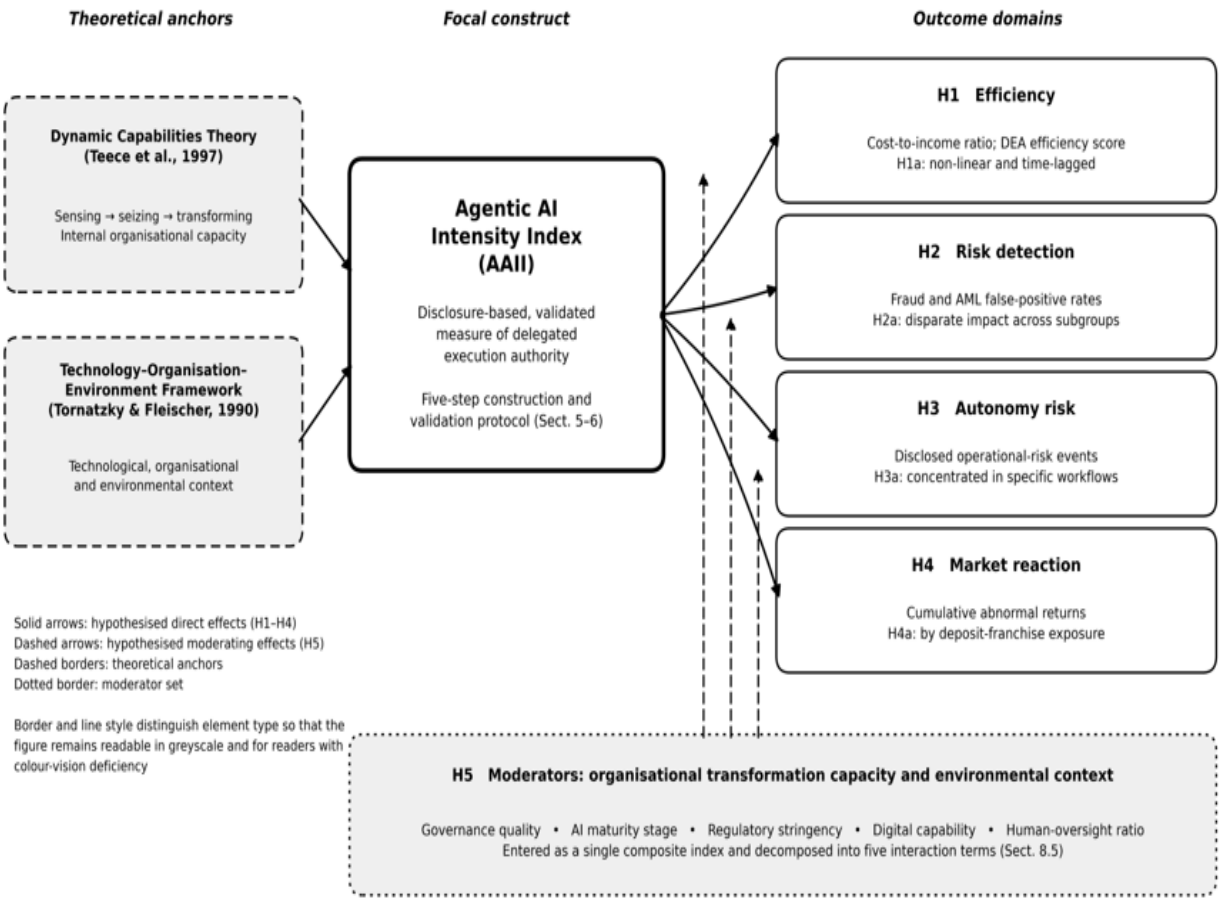



**Fig. 1.** Conceptual framework linking Agentic AI Intensity (AAII) to four outcome domains. Dashed-border elements are the theoretical anchors: Dynamic Capabilities Theory and the Technology–Organisation–Environment framework. The heavy-border element is the focal construct. Solid arrows denote hypothesised direct effects H1–H4. Dashed arrows denote the moderating effects of the H5 moderator set, shown in the dotted-border band, which conditions the H1–H3 paths. Border and line style, rather than colour, distinguish element types

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx