COSTING: Journal of Economic, Business and Accounting
Volume 9 Nomor 3, Tahun 2026
e-ISSN : 2597-5234

GOVERNING AGENTIC AI IN CENTRAL BANK: A MULTI-LAYER INSTITUTIONAL RESILIENCE ARCHITECTURE

PENGATURAN AI AGENTIK DALAM BANK SENTRAL: ARSITEKTUR KETAHANAN INSTITUSIONAL MULTI-LAPIS

Trioksa Siahaan[1*], Mulya Effendi Siregar[2], Achmad Fauzi[3], Retno Wahyuni Wijayanti[4], Edy Setiadi[5], Gerhad Lanuharsa[6]
Sekolah Tinggi Ilmu Ekonomi Dharma Bumi Putra[1]
Lembaga Pengembangan Perbankan Indonesia[2,3,4,5,6]
trioksa@stiebumiputera.ac.id[1]

ABSTRACT

Agentic artificial intelligence is no longer a distant prospect for central banks. It is already being deployed across macroprudential surveillance, payment system oversight, and predictive analytics, bringing genuine supervisory gains alongside governance risks that existing frameworks were not designed to handle. Autonomous reasoning, multi-step decision-making, and adaptive learning introduce challenges that go well beyond conventional model risk, including algorithmic contagion, delegation drift, and accountability gaps that are difficult to trace once a system has acted. This study takes that problem seriously. Using a Design Science Research methodology, it develops and validates the Layered Institutional Resilience Architecture for Agentic AI (LIRAA), a four-layer governance framework built around the core central bank mandate domains of monetary policy, payment system policy, and financial stability policy, with banking supervision included in jurisdictions where it applies. LIRAA integrates institutional authority anchoring, algorithmic accountability, adaptive supervisory escalation, and systemic risk containment into a coherent governance structure. Evaluation through structured expert assessment and scenario-based stress testing shows strong institutional feasibility and meaningful risk mitigation capacity, though experts consistently noted the need for clearer escalation thresholds and stronger legal codification. What the findings suggest, perhaps more importantly, is that financial stability in AI-mediated supervisory environments is not primarily a computational problem. It depends, in no small part, on whether institutional design is capable of preventing the kind of synchronized decision amplification that algorithmic systems can quietly produce at scale. The study contributes to AI governance literature by extending systemic risk theory into the domain of algorithmic contagion governance, and by positioning institutional architecture, rather than algorithmic performance alone, as the central determinant of stability in an era of increasingly autonomous financial intelligence.

***Keywords**: agentic artificial intelligence; central bank; design science research; ai governance; institutional resilience; systemic risk; macroprudential regulation; algorithmic contagion; supervisory technology (suptech)*

ABSTRAK

Kecerdasan buatan berbasis agen bukan lagi sekadar wacana bagi bank sentral. Sistem ini telah diterapkan secara nyata dalam pengawasan makroprudensial, pemantauan sistem pembayaran, dan analitik prediktif, membawa manfaat pengawasan yang signifikan sekaligus risiko tata kelola yang belum sepenuhnya diantisipasi oleh kerangka regulasi yang ada. Penalaran otonom, pengambilan keputusan berlapis, dan pembelajaran adaptif memunculkan tantangan yang melampaui risiko model konvensional, mencakup penularan algoritmik, pergeseran delegasi, dan celah akuntabilitas yang sulit dilacak setelah sistem bertindak. Studi ini mengangkat persoalan tersebut secara serius. Dengan menggunakan metodologi Riset Ilmu Desain (DSR), studi ini mengembangkan dan memvalidasi Arsitektur Ketahanan Kelembagaan Berlapis untuk AI Berbasis Agen (LIRAA), sebuah kerangka tata kelola empat lapis yang dibangun di atas domain mandat inti bank sentral, yaitu kebijakan moneter, kebijakan sistem pembayaran, dan kebijakan stabilitas keuangan, dengan pengawasan perbankan yang turut diintegrasikan pada yurisdiksi yang relevan. LIRAA memadukan penguatan otoritas kelembagaan, akuntabilitas algoritmik, eskalasi pengawasan adaptif, dan penahanan risiko sistemik ke dalam satu struktur tata kelola yang koheren. Evaluasi melalui penilaian ahli terstruktur dan pengujian stres berbasis skenario menunjukkan kelayakan kelembagaan yang kuat serta kapasitas mitigasi risiko yang bermakna, meskipun para ahli secara konsisten mencatat perlunya ambang batas eskalasi yang lebih jelas dan kodifikasi hukum yang lebih kuat. Yang lebih penting dari temuan ini adalah bahwa stabilitas keuangan dalam sistem pengawasan berbasis AI tidak semata-mata

bergantung pada tingkat kecanggihan komputasi. Stabilitas tersebut sangat ditentukan oleh seberapa kuat desain kelembagaan mampu mencegah terjadinya amplifikasi keputusan algoritmik secara serentak yang dapat berlangsung tanpa disadari dalam skala besar. Studi ini berkontribusi pada literatur tata kelola AI dengan memperluas teori risiko sistemik ke ranah tata kelola penularan algoritmik, sekaligus memposisikan arsitektur kelembagaan, bukan sekadar kinerja algoritma, sebagai penentu utama stabilitas di era kecerdasan keuangan yang semakin otonom.
Kata kunci: Kecerdasan buatan agen; bank sentral; penelitian ilmu desain; tata kelola AI; ketahanan kelembagaan; risiko sistemik; regulasi makroprudensial; penularan algoritmik; teknologi pengawasan (suptech)

INTRODUCTION

The rapid advancement of artificial intelligence (AI) is fundamentally reshaping institutional decision-making across the global financial system. Recent empirical and survey-based research demonstrates that AI adoption significantly enhances predictive modelling, financial risk analytics, and supervisory monitoring capabilities within banking and macroprudential environments (Cao, 2022). As central banks increasingly integrate AI into supervisory analytics, macroprudential surveillance, and policy modelling, their capacity to process high-volume financial datasets, generate forward-looking insights, and strengthen early-warning systems has expanded considerably (Daníelsson et al., 2022).

At the same time, regulatory modernization underscores the expanding role of RegTech and SupTech as core digital governance instruments. These technologies enhance supervisory agility and enable near real-time oversight, but they also introduce new institutional challenges, including algorithmic opacity, accountability gaps, cyber risk exposure, and cross-institutional coordination complexity. As supervisory and regulatory functions become progressively mediated by algorithmic systems, concerns surrounding institutional legitimacy, transparency, and decision authority become increasingly pronounced (Bagherifam et al., 2025)

Within the Indonesian financial governance landscape, digital transformation has been actively institutionalized through regulatory and policy frameworks issued by Bank Indonesia and the Financial Services Authority (OJK). The Blueprint Sistem Pembayaran Indonesia 2025 articulates the strategic direction for digital payment integration, interoperability, and advanced analytics adoption in payment system oversight (Bank Indonesia, 2019). In parallel, Bank Indonesia Regulation No. 23/11/PBI/2021 establishes governance standards emphasizing operational resilience, cybersecurity safeguards, and systemic risk mitigation in digital financial infrastructures (Bank Indonesia, 2021). On the prudential side, OJK strengthened digital risk governance through POJK No. 38/POJK.03/2016 concerning IT risk management in commercial banks (Otoritas Jasa Keuangan, 2016). More recently, POJK No. 19 Tahun 2024 on Digital Risk Management and Information Technology Governance reinforces board-level oversight, institutional accountability, cybersecurity resilience, and AI-related risk controls within financial institutions, signaling a shift toward more comprehensive digital governance regimes (POJK No. 19 of 2024 Concerning the Implementation of Information Technology and Digital Risk Management, 2024).

At the global level, multilateral institutions have consistently emphasized that AI adoption in finance must be accompanied by robust governance safeguards. The Bank for

International Settlements (BIS) highlights that AI-driven financial systems may amplify systemic spillovers if model risk management, transparency, and cross-border coordination mechanisms remain inadequate (Bank for International Settlements, 2025). Similarly, the International Monetary Fund (IMF) warns that algorithmic opacity, concentration risk, and supervisory capacity gaps may undermine financial stability if governance frameworks lag behind technological innovation (International Monetary Fund, 2024). These perspectives collectively underscore that technological capability without corresponding institutional governance may introduce new forms of systemic fragility.

Concurrently, the frontier of institutional automation is shifting beyond conventional machine learning toward agentic artificial intelligence. Agentic AI systems are characterized by autonomous reasoning, adaptive learning, multi-step planning, and goal-oriented execution capabilities. Such architectures enable complex human AI orchestration across supervisory simulations, crisis modelling, and regulatory scenario analysis, thereby redefining how decisions are generated, validated, and executed within institutional settings (Wang et al., 2024). Within financial governance contexts, these capabilities fundamentally reshape the locus of decision authority and intensify the need for robust institutional accountability structures.

The growing integration of AI into financial systems is also closely linked to transformations in systemic risk dynamics. Empirical evidence suggests that algorithmic synchronization, interconnected decision infrastructures, and accelerated feedback loops may intensify contagion effects during periods of financial stress (Daníelsson et al., 2022). Complementing this perspective, data governance scholarship emphasizes that digital regulatory architectures must evolve to ensure accountability, transparency, and institutional legitimacy in increasingly automated environments (McNulty et al., 2023). Central banking scholarship, including contributions in the Journal of Central Bank Law and Institutions, further highlights that AI integration reshapes supervisory legitimacy and regulatory trust within emerging financial ecosystems (Damaris et al., 2025).

However, the efficiency gains associated with AI adoption are accompanied by non-trivial systemic vulnerabilities. Evidence indicates that AI may amplify interconnected exposures, accelerate feedback loops, and increase the likelihood of synchronized institutional responses under stress conditions (Daníelsson et al., 2022). In parallel, scholarship on data-age financial regulation underscores that new data access and governance techniques can transform oversight logics while simultaneously intensifying risks related to concentration, opacity, and coordination if accountability mechanisms do not evolve accordingly. Concerns regarding fairness and legitimacy further reinforce that institutional trust cannot be treated as a by-product of model accuracy alone, particularly when AI systems are embedded in high-stakes economic decision-making processes (Das et al., 2023). Recent reviews of generative AI in financial economics similarly emphasize that rapidly evolving model capabilities require equally adaptive governance frameworks to prevent fragility, misuse, and over-reliance (Mo, 2025).

Apart from the growing body of research on AI applications in finance and supervisory modernization, significant conceptual and institutional gaps remain. First, existing literature remains predominantly technology-centric, focusing on model performance, algorithmic efficiency, or risk analytics, while under-theorizing the institutional governance implications of autonomous AI deployment. Second, studies on financial regulation and SupTech modernization emphasize compliance enhancement and supervisory innovation but provide limited guidance on how agentic AI should be embedded within formal accountability, liability, and escalation structures inside central banks. Third, while systemic risk scholarship recognizes AI as a potential amplifier of financial instability, it does not offer operational governance designs capable of preventing algorithmic contagion across institutional decision networks. Finally, there is an absence of integrated frameworks that connect AI oversight, inter-authority coordination, model governance, and macroprudential risk containment within a unified institutional architecture.

Addressing these gaps, this study proposes a Multi Layer Institutional Resilience Architecture for governing agentic AI in central bank. The framework conceptualizes AI governance as a layered institutional defense system integrating supervisory guardrails, organizational accountability mechanisms, regulatory coordination platforms, and systemic-risk containment protocols. By repositioning governance architecture rather than technological capability as the primary determinant of financial system stability, this study advances both theoretical and policy discourse in the age of autonomous financial intelligence.

To address the identified governance gap, this study formulates the following research questions:

1. What governance risks are uniquely associated with agentic AI deployment in monetary and supervisory authorities, beyond conventional model risk frameworks?
2. What institutional design principles are required to mitigate algorithmic contagion, delegation drift, and model monoculture in systemic regulatory environments?
3. How can a multi-layered governance architecture operationalize graduated autonomy, cross-authority coordination, and systemic risk containment for agentic AI?

Beyond technological and supervisory considerations, the integration of agentic AI into central bank raises fundamental legal and institutional mandate questions. Central banks operate under statutory authority grounded in monetary sovereignty, accountability doctrines, and administrative law principles. The delegation of analytical or supervisory functions to AI systems, even when positioned as advisory, introduces unresolved issues concerning responsibility attribution, decision authority boundaries, and mandate compliance.

In jurisdictions such as Indonesia, the mandate of Bank Indonesia is explicitly anchored in statutory provisions governing its core functions: monetary policy, payment system policy, and financial stability policy. In some central banks, banking supervision is also included as a core function. The incorporation of autonomous AI systems within these domains must therefore be assessed not only through efficiency and risk lenses, but also through institutional legality and accountability frameworks.

Accordingly, the central research problem addressed in this study is twofold: (1) how to prevent systemic risk amplification arising from autonomous AI-driven decision infrastructures, and (2) how to preserve sovereign decision authority and legal accountability within AI-mediated supervisory environments. In response, this study proposes a governance architecture that embeds legal authority anchoring, escalation sovereignty principles, and systemic containment safeguards directly into institutional design, thereby aligning technological advancement with foundational principles of public authority and financial stability.

THEORETICAL FOUNDATION AND DESIGN REQUIREMENTS

Agentic Artificial Intelligence in Financial Systems

The evolution of artificial intelligence (AI) in financial systems reflects a transition from static analytical tools toward adaptive, decision-capable infrastructures. Early applications focused on predictive modelling, fraud detection, and credit risk analytics, whereas recent developments emphasize systems capable of autonomous reasoning, adaptive learning, and goal-oriented execution (Cao, 2022). This shift represents a transformation from assistive analytics toward institutional decision augmentation and hybrid governance systems (Babina et al., 2024).

Agentic artificial intelligence refers to systems designed to operate as autonomous or semi-autonomous agents capable of multi-step planning, environmental interaction, tool utilization, and long-horizon task execution (Wang et al., 2024). . Unlike conventional models, agentic architectures incorporate persistent memory, iterative reasoning, and adaptive feedback mechanisms, enabling continuous interaction within complex institutional environments. Built upon large language models and multi-agent coordination frameworks, these systems facilitate distributed decision orchestration across organizational processes (Nguyen & Pham, 2024).

Within financial governance contexts, agentic AI enables advanced supervisory capabilities, including stress-testing simulations, macroprudential forecasting, and real-time compliance monitoring. Empirical evidence suggests that such systems improve predictive accuracy and enhance systemic risk detection within banking networks (Z. Chen et al., 2023). However, these capabilities also shift the locus of decision authority, as human actors increasingly supervise rather than directly execute analytical processes.

This transition introduces governance challenges. Algorithmic infrastructures may dilute human oversight, complicate liability attribution, and reduce decision explainability, reinforcing the need for robust accountability frameworks (Kerry, 2020). As decision processes become mediated by layered machine reasoning, governance systems must ensure transparency, traceability, and controllability across the full decision lifecycle.

From an organizational perspective, agentic AI reshapes institutional authority structures. Human roles evolve toward supervisory orchestration, encompassing model validation, ethical oversight, escalation governance, and auditability functions. Research on human–AI collaboration indicates that optimal outcomes emerge from calibrated hybrid intelligence systems balancing automation efficiency with human judgment (Rai et al., 2019).

At the systemic level, agentic AI introduces coordination risks. Multi-

agent architectures interacting across institutions may generate networked decision ecosystems characterized by interdependence and synchronization. While such integration enhances coordination, it may also amplify systemic risk through algorithmic convergence and procyclical responses during financial stress (Daníelsson et al., 2022).

Recent financial economics research further highlights risks related to model concentration, data dependency, and reliance on shared computational infrastructures (Mo, 2025). These developments suggest that governance must extend beyond model-level oversight toward infrastructure-level resilience and cross-institutional accountability.

Despite these advances, the literature remains fragmented. Technical research on agentic AI focuses on system capabilities, while governance scholarship emphasizes compliance and risk management, with limited integration between these domains. This gap is particularly pronounced in central bank environments, where supervisory authority, systemic stability mandates, and institutional accountability intersect.

Central Bank Governance and AI Integration

Central bank governance has evolved significantly in response to financial globalization, digitalization, and systemic risk complexity. Traditionally focused on monetary policy independence and price stability, central bank mandates now encompass three core policy functions: monetary policy, payment system policy, and financial stability policy. In addition, banking supervision is included as a core function in some central banks. Governance frameworks must therefore address each of these mandate domains, including macroprudential supervision, financial stability coordination, and systemic risk oversight (Goodhart & Lastra, 2018).

As financial systems become increasingly data-driven, central banks are transforming into hybrid institutions that function not only as monetary authorities but also as data governors, systemic risk monitors, and digital infrastructure supervisors (Bholat, 2020). This transformation redefines governance as a multi-layer coordination problem across interconnected financial ecosystems.

The integration of AI into supervisory and macroprudential domains further complicates governance structures. AI-enabled SupTech infrastructures enable real-time surveillance, anomaly detection, and predictive financial monitoring at scale (Auer et al., 2022). While these capabilities enhance efficiency, they also introduce governance challenges related to model opacity, explainability, and accountability.

Institutional legitimacy remains central to central bank effectiveness. Trust depends not only on policy outcomes but also on the transparency and accountability of decision-making processes, particularly when algorithmic systems influence supervisory judgments (de Haan et al., 2020).

Global institutions emphasize similar concerns. The Bank for International Settlements highlights the need for governance safeguards, model validation mechanisms, and cross-border coordination (Bank for International Settlements, 2025), while the International Monetary Fund underscores the importance of institutional redesign and regulatory capacity development in response to AI adoption (International Monetary Fund, 2024). These perspectives reinforce the

necessity of integrating technological capability with institutional governance.

Recent scholarship further emphasizes explainability, accountability allocation, and supervisory override mechanisms as core components of AI governance (McNulty et al., 2023). This reflects a shift from hierarchical governance toward distributed decision systems involving both human and computational actors.

Moreover, systemic risk governance increasingly requires coordination across institutional boundaries. Macroprudential oversight depends on collaboration among central banks, financial regulators, and international bodies to monitor interconnected vulnerabilities (Auer et al., 2022). As AI systems operate across these domains, the need for harmonized governance frameworks becomes more critical.

Despite these advances, existing governance models remain insufficiently equipped to address autonomous decision agents embedded within supervisory infrastructures. Questions regarding liability, escalation authority, and decision sovereignty remain unresolved. This gap highlights the need for governance architectures capable of supervising agentic AI while preserving institutional legitimacy and accountability.

AI Amplified Systemic Risk and Governance Implications

The integration of AI into financial systems is transforming systemic risk dynamics by introducing new channels of amplification, synchronization, and opacity. While AI enhances analytical capabilities, it may also increase model monoculture, where institutions rely on similar data and algorithms, raising the likelihood of correlated decision-making (Mo, 2025).

In central bank contexts, these risks are particularly significant, as AI adoption spans all core central bank functions—monetary policy, payment system policy, financial stability policy, and banking supervision. Among these, payment system policy has experienced the most extensive AI integration, with automated clearing, real-time gross settlement, and digital currency oversight increasingly driven by agentic systems (Auer et al., 2022). This creates tighter coupling between institutional responses and algorithmic outputs, increasing the potential for synchronized policy actions.

Systemic risk research shows that AI can accelerate contagion dynamics by compressing reaction times and reducing opportunities for human intervention. Shared models may reinforce procyclical behavior, where institutions expand simultaneously during economic upswings and contract during downturns (Chen, 2022).

Governance research further highlights that digital regulatory architectures may increase concentration and coordination risks if accountability mechanisms lag behind technological adoption (McNulty et al., 2023). This suggests that systemic risk governance must evolve toward monitoring interactions among algorithmic systems rather than focusing solely on individual institutions.

Effective governance therefore requires system-level mechanisms, including correlation stress testing, monitoring of shared infrastructure dependencies, and escalation protocols that preserve human override authority (Das et al., 2023). From a policy perspective, AI risk management becomes a macroprudential concern when model concentration, opacity, and

institutional synchronization increase the likelihood of cascading failures (Mo 2025).

Although existing frameworks developed by global institutions emphasize transparency and model risk management, they remain largely model-centric and do not provide integrated institutional architectures capable of addressing agentic autonomy and systemic synchronization dynamics.

Synthesis and Design Requirements for Governance Architecture

Synthesizing the literature across AI systems, central bank governance, and systemic risk reveals a consistent pattern of fragmentation. Existing research addresses AI capabilities, regulatory frameworks, and systemic risk dynamics in isolation, but does not provide an integrated approach to governing autonomous decision agents within institutional environments.

This fragmentation produces four key gaps. First, AI research emphasizes technical performance while under-theorizing institutional governance implications. Second, central bank governance frameworks focus on mandates and coordination but do not address autonomous decision systems. Third, systemic risk scholarship identifies AI as a risk amplifier without offering operational governance mechanisms. Fourth, existing policy frameworks remain model-centric and lack multi-layer institutional design.

These gaps inform a set of design requirements that guide the development of the governance artifact in this study:

DR1: Multi-step auditability, ensuring traceability of AI reasoning and decision pathways.

DR2: Graduated autonomy control, enabling risk-based oversight rather than binary supervision.

DR3: Escalation sovereignty, preserving institutional authority to override AI decisions.

DR4: Cross-authority coordination, supporting governance across institutional boundaries.

DR5: Model convergence monitoring, detecting correlated behavior across systems.

DR6: Delegation drift prevention, limiting uncontrolled expansion of AI roles.

DR7: Accountability mapping, defining decision rights and liability structures.

DR8: Infrastructure and vendor risk control, addressing dependency on shared technologies.

These requirements serve as the foundation for the Design Science Research (DSR) process. In the subsequent section, they are translated into design principles and instantiated within a multi-layer governance architecture, enabling systematic development, demonstration, and evaluation of the proposed framework.

RESEARCH METHODS

This study adopts a Design Science Research (DSR) approach, in reference to Gregor and Zwikael (2024), with the intention to develop, justify, and validate a Multi-Layer Institutional Resilience Governance Framework for governing agentic AI in central bank. DSR is suitable because the main contribution of this article is an actionable governance artifact, a framework with design principles, controls, and escalation mechanisms, intended to address a real-world institutional problem, namely AI-enabled supervisory and macroprudential decision systems that increasingly exhibit autonomy, opacity, and cross-agency spillover risk. Recent DSR scholarship further emphasizes that

publishable design research must demonstrate (i) transparent design logic, (ii) defensible evaluation, and (iii) credible knowledge contributions beyond a "conceptual proposal" (Akoka et al., 2023).

DSR Process and Outputs. The research follows a staged DSR logic aligned with contemporary quality expectations: problem framing, artifact design, and evaluation, with explicit transparency practices and validity claims (Brendel et al., 2021). To better manage the inherent complexity of multi-actor governance settings (central bank–regulator–industry–international bodies), the study adopts a "complexity-aware" DSR structuring approach, ensuring that artifact scope, layers, and mechanisms are explicitly bounded and traceable (Hevner et al., 2024).

Stage 1: Problem identification and motivation

The researchers synthesize institutional and regulatory challenges of agentic AI in central bank through (a) targeted document analysis (BIS/IMF guidance; BI/OJK regulations), and (b) structured literature mapping on AI governance, SupTech/RegTech, and systemic risk. This stage produces a clearly articulated problem statement alongside a set of design requirements, including auditability, accountability, escalation capacity, cross-authority coordination, and systemic risk containment.

Stage 2: Defining objectives of a governance solution

The study translates these requirements into design objectives and principles, such as "layered accountability," "model–data governance coupling," "institutional escalation and kill-switch protocols," and "inter-authority coordination interfaces." To strengthen the theoretical contribution, DSR contribution typologies are explicitly applied to ensure the artifact delivers a structured knowledge output (framework, principles, and instantiation logic), rather than remaining a narrative-only argument (Larsen et al., 2025).

Stage 3: Design and development (artifact construction)

The artifact is developed as a multi-layer governance architecture consisting of:

1) Institutional oversight layer (mandate, roles, accountability, escalation),
2) Model and data governance layer (validation, monitoring, documentation, traceability),
3) Operational control layer (human-in-the-loop design, override, incident response), and
4) Systemic-risk containment layer (stress pathways, coordination triggers, cross agency playbooks) (Matheus et al., 2021).

The design logic underpinning these layers is explicitly documented to ensure traceability between identified problems, design choices, and governance mechanisms, in line with DSR transparency expectations (Tuunanen et al., 2024).

Stage 4: Demonstration

The framework is demonstrated through scenario mapping, such as an agentic supervisory system recommending macroprudential actions under stress conditions or supervisory simulations producing synchronized institutional responses. This demonstration generates traceable governance pathways, illustrating how decisions propagate across layers and where specific controls, overrides, or escalation mechanisms are activated.

Stage 5: Evaluation (multi-method, governance-appropriate).

Because the artifact is a governance framework rather than a software prototype, evaluation employs a combination of methods: expert evaluation (structured review by central bank and regulatory governance experts), scenario-based evaluation (stress-case walkthroughs and failure mode analysis), and replication logic to clarify conditions for applicability across jurisdictions and institutional arrangements (Matheus, Janssen, and Janowski 2021). Evaluation reporting is aligned with contemporary DSR validity standards by explicitly linking claims to evidence and defining the contextual boundaries of generalization (Tuunanen et al., 2024).

To strengthen rigor, the framework was subjected to structured expert validation. The expert panel comprised seven (7) senior practitioners selected through purposive sampling based on the following eligibility criteria: (i) a minimum of ten years of professional or academic experience in central bank governance, financial regulation, AI risk management, or a related domain; (ii) demonstrable familiarity with supervisory technology, digital financial infrastructure, or institutional AI governance; and (iii) active engagement in either policy formulation, regulatory design, or applied research in these domains at the time of the study. Purposive sampling was employed to ensure that the panel reflected the interdisciplinary nature of the governance challenges addressed by LIRAA, spanning monetary authority practice, prudential supervision, AI risk, and legal-institutional frameworks. The seven-expert panel size is consistent with DSR evaluation norms for governance oriented artifact assessment, where depth of domain knowledge is prioritized over sample breadth (Matheus et al., 2021).

Table 1. Expert Panel Profile

No	Position / Role	Institution / Domain	Area of Expertise	Years of Experience
1	Senior Director, Macroprudential Policy	Central Bank (Monetary Authority)	Macroprudential Policy, Systemic Risk	18 years
2	Head of SupTech & Digital Supervision	Financial Services Authority (OJK)	SupTech, Digital Regulatory Governance	14 years
3	AI Risk Management Lead	Multilateral Financial Institution	AI Risk, Model Governance, Financial Stability	16 years
4	Professor of Financial Law & Regulatory Governance	National University (Law Faculty)	Administrative Law, AI Accountability, Regulatory Design	22 years

No	Position / Role	Institution / Domain	Area of Expertise	Years of Experience
5	Chief Risk Officer (Banking Sector)	Systemically Important Bank	Operational Risk, Credit Risk, Basel III Implementation	19 years
6	Associate Professor, Digital Economy & AI Policy	Graduate School of Economics	AI Policy, Fintech Regulation, DSR Methodology	12 years
7	Director of Institutional Technology Governance	Regional Regulatory Body	IT Governance, Cybersecurity, Cross-border Regulatory Coordination	15 years

Collectively, the seven experts represent a deliberately balanced cross-section of institutional actors relevant to the governance of agentic AI in central bank: two practitioners from monetary and prudential regulatory authorities (E1, E2), one expert from a multilateral financial institution (E3), two academics specializing in financial law and AI policy (E4, E6), one practitioner from the regulated banking sector (E5), and one expert from a regional regulatory coordination body (E7). This composition ensures that the evaluation captures both practitioner implementability and theoretical rigor across the full governance lifecycle addressed by LIRAA.

Assessment Instrument Design. The evaluation instrument was developed in three iterative steps. First, a preliminary structured questionnaire was derived directly from the eight design requirements (DR1–DR8) and the four LIRAA governance layers, producing an initial item pool of thirty-two assessment statements. Second, the instrument underwent content review by two researchers outside the study team to assess clarity, completeness, and alignment with DSR evaluation standards (Larsen et al., 2025), resulting in consolidation into twenty evaluation items grouped under five dimensions: (i) institutional feasibility, (ii) risk mitigation adequacy, (iii) escalation clarity, (iv) systemic containment robustness, and (v) legal accountability coherence. Third, each item was scored using a five-point Likert-type response scale anchored at 1 = Strongly Disagree and 5 = Strongly Agree, allowing quantitative aggregation alongside qualitative commentary. Experts were also invited to provide open-ended observations on gaps, implementation barriers, and jurisdictional applicability, which were subsequently analyzed using thematic coding. The instrument was administered through structured individual review sessions to minimize anchoring effects from peer interaction. Inter-rater consistency across the panel was assessed using Kendall's coefficient of concordance (W), confirming acceptable agreement across evaluation dimensions.

ASEM Parameter Specification. The Adaptive Supervisory Escalation Matrix (ASEM) operationalizes Layer III of LIRAA by replacing binary human in the loop triggers with a continuous, risk-sensitive escalation function. ASEM evaluates each AI generated decision or recommendation against three operationally defined parameters. The first parameter, *Decision Impact Level (DIL)*, classifies the potential institutional consequence of an AI recommendation on a three-tier ordinal scale: operational (routine, bounded-scope actions), supervisory (policy-relevant recommendations affecting one or more institutions), or systemic (recommendations with macroprudential or cross-authority implications). The second parameter, *Confidence Score (CS)*, captures the AI system's internal model certainty for a given output, expressed as a normalized probability score between 0 and 1. Low confidence scores below a calibrated threshold trigger mandatory escalation regardless of decision impact level, reflecting the principle that uncertain outputs in high-stakes institutional environments require human validation before action. The third parameter, *Systemic Relevance Index (SRI)*, quantifies the degree to which a decision may interact with or influence the behavior of other institutions or AI systems operating in the same supervisory ecosystem. SRI draws on network topology data, cross-institutional exposure maps, and historical correlation patterns to assign a continuous relevance score. When DIL is systemic and SRI exceeds the defined threshold, ASEM automatically elevates the escalation trajectory to institutional committee review, bypassing standard supervisory validation. The three parameters are jointly evaluated at each decision node, with the most conservative escalation requirement taking precedence, thereby ensuring that risk tiered oversight is always binding at the highest applicable level.

MCRI Computation. The Model Convergence Risk Indicator (MCRI) is designed to detect systemic vulnerabilities arising from correlated AI behavior across institutions. MCRI is computed as a composite score integrating three sub indicators. The first sub-indicator, *Output Similarity Score (OSS)*, measures the pairwise cosine similarity of decision recommendations or analytical outputs generated by AI systems deployed across monitored institutions within a defined observation window. The second sub-indicator, *Infrastructure Dependency Index (IDI)*, quantifies the proportion of monitored institutions relying on shared computational infrastructure, model families, or third-party AI vendors, thereby capturing concentration risk at the platform level. The third sub-indicator, *Decision Synchronization Rate (DSR)*, tracks the temporal clustering of AI generated decisions across institutions, measuring the degree to which comparable outputs are generated within a narrow time window a pattern that would indicate algorithmic synchronization rather than independent institutional reasoning. The composite MCRI is computed as a weighted average of the three normalized sub-indicators: $MCRI = w_1 \cdot OSS + w_2 \cdot IDI + w_3 \cdot DSR$, where weights ($w_1$, w_2, w_3) are calibrated by the supervisory authority to reflect the relative systemic importance of each convergence channel in a given institutional context, subject to the constraint that $w_1 + w_2 + w_3 = 1$. An MCRI value exceeding the defined red-zone threshold (provisionally set at 0.70 on a normalized 0–1 scale) triggers a supervisory alert within Layer II and activates the ASEM escalation protocol in Layer III. MCRI is computed on a

rolling basis across defined observation windows, enabling real-time monitoring of convergence dynamics rather than relying on periodic audit cycles.

Figure 1. LIRAA Governance Architecture: Four Layer Conceptual Diagram

LAYER I	**Institutional Authority Anchor** Sovereign mandate • Decision-rights allocation • Escalation sovereignty • Kill switch authority	**Design Requirements Met:** DR3 (Escalation Sovereignty) • DR6 (Delegation Drift Prevention) • DR7 (Accountability Mapping)
LAYER II	**Algorithmic Accountability and Model Integrity** MCRI monitoring • Model inventory & traceability • Audit logging • Convergence alerts	**Design Requirements Met:** DR1 (Multi-step Auditability) • DR5 (Model Convergence Monitoring) • DR7 (Accountability Mapping)
LAYER III	**Adaptive Supervisory Escalation Matrix (ASEM)** DIL • Confidence Score • Systemic Relevance Index • Risk-tiered human review	**Design Requirements Met:** DR2 (Graduated Autonomy) • DR3 (Escalation Sovereignty) • DR6 (Delegation Drift Prevention)
LAYER IV	**Systemic Risk Containment & Cross-Authority Coordination** Joint AI risk registers • Coordinated stress testing • Cross-authority playbooks • Contagion circuit-breakers	**Design Requirements Met:** DR4 (Cross-authority Coordination) • DR5 (Convergence Monitoring) • DR8 (Infrastructure Risk Control)

Layers are interdependent and vertically integrated. ASEM (Layer III) draws escalation triggers from MCRI alerts (Layer II). Cross-authority coordination (Layer IV) is activated when systemic thresholds in Layers II–III are exceeded.

Figure 1 illustrates the four-layer architecture of LIRAA in its governance context. The vertical arrangement reflects the escalation logic of the framework: Layer I anchors all governance activity in sovereign institutional authority; Layer II provides real-time algorithmic accountability and convergence monitoring through MCRI; Layer III operationalizes risk-tiered escalation through ASEM, dynamically adjusting oversight intensity based on DIL, CS, and SRI inputs; and Layer IV extends governance beyond individual institutions to address systemic contagion and cross-authority coordination. Each layer is explicitly mapped to the design requirements (DR1–DR8) identified in Stage 2 of the DSR process, ensuring full traceability between problem diagnosis, design logic, and governance mechanism. Experts assessed the framework against all four layers using the structured instrument described above, with evaluation

findings reported in the subsequent Results section.

Stage 6: Communication

Findings are communicated as a governance artifact comprising: (i) a framework diagram, (ii) design principles, (iii) layer-specific mechanisms, (iv) evaluation evidence summary, and (v) policy implications. This structured communication format ensures both academic rigor and practical usability for policy stakeholders.

This study establishes four validity dimensions consistent with contemporary DSR standards:

1) Construct validity is ensured through explicit alignment between identified governance risks and artifact design principles.
2) Internal validity is achieved through coherent layering and traceable escalation pathways.
3) External validity is bounded through clearly stated institutional assumptions and boundary conditions.
4) Practical validity is supported through structured expert evaluation and scenario-based stress testing.

By explicitly articulating and operationalizing these validity dimensions, the study strengthens methodological transparency, analytical rigor, and the defensibility of its governance contributions, thereby enhancing the credibility and practical applicability of the proposed framework in addressing emerging risks from agentic AI in central bank.

RESULTS AND DISCUSSIONS

RESULTS

LIRAA Governance Architecture

The primary result of this study is the development of the Layered Institutional Resilience Architecture for Agentic AI (LIRAA), a multi layer governance framework designed to supervise agentic artificial intelligence in central bank environments. The framework conceptualizes AI governance not as a single compliance mechanism, but as a layered institutional resilience system integrating authority control, algorithmic accountability, supervisory escalation, and systemic risk containment.

The LIRAA architecture is built on the premise that agentic AI introduces autonomous reasoning, multi-step planning, and adaptive decision-making capabilities that may alter traditional governance boundaries. As such, governance must shift from model-centric oversight to institutional architecture design, ensuring that autonomy remains embedded within enforceable supervisory boundaries.

The framework consists of four interdependent governance layers:

- Layer I: Institutional Authority Anchor, which preserves sovereign decision-making authority through formal governance mandates, decision-rights allocation, and escalation sovereignty mechanisms.
- Layer II: Algorithmic Accountability and Model Integrity, which ensures transparency, traceability, and detection of model convergence risks across institutions.
- Layer III: Adaptive Supervisory Escalation Matrix (ASEM), which operationalizes graduated autonomy through risk-sensitive escalation thresholds.
- Layer IV: Systemic Risk Containment and Cross Authority Coordination, which mitigates macro-level contagion risks through coordination protocols and system-wide stress monitoring.

Together, these layers establish a defensive governance architecture that ensures agentic AI systems remain

aligned with institutional mandates while maintaining operational flexibility.

Design Mechanisms: ASEM and MCRI

To operationalize governance principles, the LIRAA framework introduces two core mechanisms.

The first mechanism, the Autonomy–Supervision Escalation Matrix (ASEM), replaces binary human-in-the-loop control with a graduated autonomy model. Decision-making authority is dynamically adjusted based on three parameters: decision impact level, confidence score, and systemic relevance. This enables real-time escalation from automated execution to human oversight or institutional committee review when risk thresholds are exceeded.

The second mechanism, the Model Convergence Risk Indicator (MCRI), is designed to detect systemic vulnerabilities arising from correlated AI outputs across institutions. MCRI evaluates similarity in model outputs, shared infrastructure dependencies, and cross-institutional decision synchronization. High convergence levels trigger supervisory alerts and escalation via ASEM protocols, preventing algorithmic monoculture and synchronized policy amplification. Together, ASEM and MCRI operationalize the transition from static governance rules to adaptive, risk-responsive institutional control systems.

Demonstration of Framework Functionality

The applicability of LIRAA is demonstrated through a stylized stress scenario in which an agentic AI supervisory system identifies emerging credit overheating and simultaneously generates macroprudential tightening recommendations across multiple financial institutions.

In a non-governed environment, such synchronized outputs could lead to procyclical contraction and systemic amplification effects. However, under LIRAA governance:

- Layer I ensures final macroprudential authority remains with the central bank board or monetary committee.
- Layer II activates MCRI alerts, detecting high correlation in AI-generated recommendations.
- Layer III (ASEM) escalates decision-making from automated recommendation to human supervisory validation.
- Layer IV triggers cross-authority coordination protocols to prevent simultaneous tightening across institutions.

This demonstration shows how LIRAA transforms autonomous AI outputs into controlled institutional decision flows, reducing the risk of algorithmic amplification during systemic stress conditions.

Evaluation Results

The framework was evaluated using structured expert assessment involving central bank officials, regulatory specialists, AI risk experts, and legal governance scholars. Evaluation focused on institutional feasibility, risk mitigation effectiveness, escalation clarity, systemic containment robustness, and legal accountability coherence.

Results indicate strong overall performance:

- Institutional feasibility received high scores, indicating practical implementability within central bank structures.
- Risk mitigation adequacy was rated highest, reflecting strong coverage of systemic and model-related risks.

- Systemic containment mechanisms were evaluated positively, particularly in relation to cross-institutional coordination.
- Escalation clarity received moderate scores, indicating the need for further refinement of threshold calibration.
- Legal accountability coherence was the lowest-rated dimension, reflecting ongoing ambiguity in statutory alignment for AI-mediated decision systems.

DISCUSSION

Agentic artificial intelligence fundamentally shifts supervisory technology from decision support systems toward semi-autonomous orchestration mechanisms. This transition requires regulators and central banks to reconceptualize governance design as a financial stability instrument, rather than a purely technical compliance layer. Recent evidence and policy assessments indicate that AI-related vulnerabilities can propagate through multiple channels including market functioning, operational resilience, and stress transmission via correlated models and shared infrastructures making macroprudential ready governance essential from the outset (Aldasoro et al. 2025). In this context, AI governance becomes not an operational add-on but a core component of systemic risk management architecture.

A first implication concerns institutional authority and mandate design. As central banks expand AI use across their core functions—monetary policy, payment system policy, financial stability policy, and in some jurisdictions banking supervision—governance must explicitly preserve decision sovereignty under conditions of increasing automation. Notably, among these functions, payment system policy has been most significantly affected by the growing deployment of AI. Stocktaking evidence shows accelerating and diversifying AI adoption in central bank functions, reinforcing the need for governance charters that preserve institutional legitimacy alongside technological advancement (Araujo, Doerr, and Gambacorta 2024). Accordingly, mandate clarity and "sovereign override" principles should be formally codified in supervisory and policy workflows. This includes defining which decisions remain exclusively human, under what conditions agentic systems may execute actions, and when mandatory rollback or suspension mechanisms must be activated.

A second implication relates to data governance and institutional interoperability. Data access technologies and "new governance" approaches increasingly embed regulatory objectives into institutional processes, improving supervisory granularity but simultaneously increasing exposure to concentration risk, fragmented data lineage, and cross-institutional dependency structures (McNulty, Miglionico, and Milne 2023). Without coordinated interoperability governance, these developments may unintentionally amplify systemic fragility. In Indonesia, this challenge is particularly relevant given Bank Indonesia's strategic direction toward payment system ecosystem integration (Indonesia 2023) and OJK's IT risk governance framework for financial institutions (Financial Services Authority (OJK) 2016). The implication is that data governance must evolve from compliance orientation toward systemic risk-aware infrastructure governance.

Third, the findings highlight the need to reposition model governance from validation toward systemic resilience management. Existing AI governance approaches are largely

centered on model accuracy and ex-post validation, but empirical evidence shows that AI systems can reshape systemic risk dynamics by amplifying feedback loops, synchronizing decision patterns, and increasing interconnected exposures during stress periods (Aldasoro et al. 2025). Policy institutions similarly warn that model opacity, third-party dependencies, and correlated AI usage can amplify systemic vulnerability, particularly as generative and agentic systems scale. This study therefore implies that model governance must institutionalize four complementary functions: (i) comprehensive model inventory and traceability, (ii) drift and stress testing under macro scenarios, (iii) concentration monitoring across vendors and model families, and (iv) explainability and audit logging proportional to decision impact. These mechanisms collectively shift governance from reactive validation toward proactive systemic resilience monitoring.

Fourth, a critical implication concerns the design of adaptive supervisory escalation mechanisms for agentic systems. Traditional human-in-the-loop models rely on static intervention rules, which are insufficient in environments where AI systems perform multi-step reasoning and autonomous planning. Instead, central banks should adopt risk-tiered escalation structures that dynamically adjust oversight intensity based on confidence levels, systemic relevance, and potential harm thresholds. This includes formalizing conditions for mandatory human review, supervisory escalation, and kill-switch activation. This approach aligns with emerging research on autonomous agents and emphasizes the need for auditability in multi-step decision processes within high-stakes institutional environments (Wang et al. 2024). In this sense, escalation governance becomes a core component of AI-enabled supervisory architecture rather than an exceptional intervention mechanism.

Finally, the study highlights the necessity of cross-authority systemic containment frameworks to address algorithmic contagion risks. As AI adoption expands across financial institutions, supervisory authorities increasingly face the risk of synchronized model behavior and correlated policy responses. This creates the potential for "algorithmic contagion," where shared analytical infrastructures amplify systemic stress rather than mitigating it. Policy discussions increasingly emphasize the importance of improving monitoring capacity through coordinated supervision, shared analytics, and scenario-based stress testing across authorities. Building on this, the LIRAA framework implies the need for institutionalizing (i) joint AI risk registers, (ii) interoperability standards for supervisory signal exchange, and (iii) coordinated stress-testing exercises that explicitly evaluate correlated model outputs and synchronized institutional responses. These mechanisms extend macroprudential oversight into the domain of AI-driven decision infrastructures and reinforce system-wide resilience.

CONCLUSION AND SUGGESTION

This study develops the Layered Institutional Resilience Architecture for Agentic AI (LIRAA) as a governance artifact for supervising autonomous artificial intelligence within central bank systems. Moving beyond traditional model-centric approaches, LIRAA conceptualizes AI governance as a multi-layer institutional resilience framework integrating authority anchoring,

algorithmic accountability, adaptive escalation, and systemic risk containment.

The findings show that agentic AI significantly transforms supervisory infrastructures through autonomous reasoning, multi step planning, and distributed decision-making capabilities. While these features enhance analytical efficiency and responsiveness, they also introduce new governance risks, including model convergence, automation bias, correlated institutional behavior, and algorithmic contagion. Existing literature largely emphasizes technical performance and supervisory digitization but lacks integrated institutional architectures capable of governing autonomous AI systems in macroprudential environments. LIRAA addresses this gap by embedding governance mechanisms directly into institutional design.

Theoretically, this study contributes to algorithmic accountability, macroprudential governance, and supervisory digital transformation literature. It extends systemic risk theory by introducing the concept of algorithmic contagion governance, where correlated AI outputs across institutions may amplify systemic vulnerability. By integrating hybrid human AI orchestration with layered institutional controls, the study positions governance architecture not computational capability as the central determinant of financial system stability in the era of agentic AI.

From a policy perspective, LIRAA offers central banks and financial regulators a structured blueprint for supervising autonomous AI while preserving institutional legitimacy and sovereign decision authority. The layered design ensures clear allocation of decision rights, calibrated escalation pathways, and cross-authority coordination mechanisms to mitigate systemic spillover risks. This shifts governance from a compliance-oriented function toward a core component of macroprudential policy design.

However, the study remains conceptual. Empirical validation across jurisdictions, as well as simulation-based stress testing of correlated AI deployments, is required to strengthen external validity. In addition, operationalization of mechanisms such as the Model Convergence Risk Indicator (MCRI) and Adaptive Supervisory Escalation Matrix (ASEM) remains an important area for future research.

The applicability of LIRAA is bounded by institutional conditions, including the assumption of advisory or semi-autonomous AI roles, formalized governance structures, and clear legal accountability frameworks typical of central bank environments. Implementation may therefore vary across jurisdictions depending on regulatory maturity and digital infrastructure readiness.

Governing agentic AI in central banks is not, at its core, a technical challenge. Waiting for something to go wrong and then auditing the model is not a governance strategy; it is a gamble. Resilience, in this context, has to be built into institutional design from the start, before autonomous systems are embedded deeply enough to make course correction costly or slow. LIRAA offers a structured path toward that kind of ex ante readiness. It does not claim to solve every problem that agentic AI will eventually pose, and the authors are candid about the limits of what a framework alone can accomplish. What it does provide is a coherent architecture for keeping autonomous systems accountable to the core mandates of central banks, including monetary

policy, payment system policy, and financial stability, as those systems take on more consequential roles in an increasingly algorithmically mediated financial order.

REFERENCES

Akoka, J., Comyn-Wattiau, I., Prat, N., & Storey, V. C. (2023). Knowledge contributions in design science research: Paths of knowledge types. *Decision Support Systems*, *166*, 113898. https://doi.org/10.1016/j.dss.2022.113898

Auer, R., Cornelli, G., & Frost, J. (2022). *Artificial Intelligence in Central Bank* (Issue 1044).

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial Intelligence, firm growth, and product innovation. *Journal of Financial Economics*, *151*(January), 103745. https://doi.org/10.1016/j.jfineco.2024.103916

Bagherifam, N., Naghdi, S., Ahmadian, V., Fazlzadeh, A., & Shishehgarkhaneh, M. B. (2025). Digital regulatory governance: The role of RegTech and SupTech in transforming financial oversight and administrative capacity. *International Journal of Financial Studies*, *13*(4), 217. https://doi.org/10.3390/ijfs13040217

Bank for International Settlements. (2025). *Financial stability implications of artificial intelligence - Executive summary*. https://www.bis.org/fsi/fsisummaries/exsum_23904.htm

Bank Indonesia. (2019). *Blueprint sistem pembayaran Indonesia 2025*.

Bank Indonesia. (2021). *Regulation No. 23/11/PBI/2021 on National Payment System Standards*.

Bholat, D. (2020). *Big Data and Central Banks*.

Brendel, A. B., Lembcke, T.-B., Muntermann, J., & Kolbe, L. M. (2021). Toward replication study types for design science research. *Journal of Information Technology*, *36*(3), 198–215. https://doi.org/10.1177/02683962211006429

Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, *55*(3), Article 64. https://doi.org/10.1145/3502289

Chen, Y. (2022). Bank interconnectedness and financial stability: The role of bank capital. *Journal of Financial Stability*, *61*, 101019. https://doi.org/10.1016/j.jfs.2022.101019

Chen, Z., Pelger, M., & Zhu, J. (2023). Deep learning in asset pricing. *Management Science*, *70*(2). https://doi.org/10.1287/mnsc.2022.4417

Damaris, R., Rosadi, S. D., & Bratadana, I. M. D. (2025). Data governance for Artificial Intelligence implementation in the financial sector: An Indonesian perspective. *Journal of Central Bank Law and Institutions*, *4*(3). https://doi.org/10.21098/jcli.v4i3.430

Daníelsson, J., Macrae, R., & Uthemann, A. (2022). Artificial Intelligence and systemic risk. *Journal of Banking \& Finance*, *140*(July), 106290. https://doi.org/10.1016/j.jbankfin.2021.106290

Das, S., Stanton, R., & Wallace, N. (2023). Algorithmic fairness. *Annual Review of Financial Economics*, *15*, 565–593. https://doi.org/10.1146/annurev-

financial-110921-125930

de Haan, J., Eijffinger, S., & Waller, C. (2020). *The European Central Bank: Credibility, transparency, and governance*. Oxford University Press. https://doi.org/10.1093/oso/9780190628211.001.0001

Goodhart, C., & Lastra, R. (2018). Populism and central bank independence. *Open Economies Review*, *29*, 49–68. https://doi.org/10.1007/s11079-017-9447-y

Gregor, S., & Zwikael, O. (2024). Design science research and the co-creation of project management knowledge. *International Journal of Project Management*, *42*(3), 102584. https://doi.org/10.1016/j.ijproman.2024.102584

Hevner, A. R., Parsons, J., Brendel, A. B., Lukyanenko, R., Tiefenbeck, V., Tremblay, M. C., & vom Brocke, J. (2024). Transparency in design science research. *Decision Support Systems*, *182*, 114236. https://doi.org/10.1016/j.dss.2024.114236

International Monetary Fund. (2024). Artificial Intelligence and the future of financial supervision. *September 6, 2024*. https://www.imf.org/en/news/articles/2024/09/06/sp090624-artificial-intelligence-and-its-impact-on-financial-markets-and-financial-stability

Kerry, C. F. (2020). *Protecting privacy in an AI-driven world*. https://www.brookings.edu/articles/protecting-privacy-in-an-ai-driven-world/

Larsen, K. R., Lukyanenko, R., Mueller, R. M., Storey, V. C., Parsons, J., VanderMeer, D., & Hovorka, D. S. (2025). Validity in Design Science. *MIS Quarterly*, *49*(4), 1267–1294. https://doi.org/10.25300/MISQ/2024/18064

Matheus, R., Janssen, M., & Janowski, T. (2021). Design principles for creating digital transparency in government. *Government Information Quarterly*, *38*(1), 101550. https://doi.org/10.1016/j.giq.2020.101550

McNulty, D., Miglionico, A., & Milne, A. (2023). Data access technologies and the "new governance" techniques of financial regulation. *Journal of Financial Regulation*, *9*(2), 225–248. https://doi.org/10.1093/jfr/fjad008

Mo, H. (2025). (Generative) AI in financial Economics. *Review of Behavioral Economics*, *23*(4), 509–587. https://doi.org/10.1080/14765284.2025.2569006

Nguyen, H., & Pham, X. (2024). AI and organizational governance in financial services. *Information Systems Research*, *35*(1). https://doi.org/10.1287/isre.2023.1147

POJK No. 19 of 2024 concerning the Implementation of Information Technology and Digital Risk Management 2024, (2024). https://jdih.ojk.go.id/

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Toward human-AI hybrids. *MIS Quarterly*, *43*(1). https://doi.org/10.25300/MISQ/2019/14117

Tuunanen, T., Winter, R., & vom Brocke, J. (2024). Dealing with complexity in design science research: A methodology using design echelons. *MIS Quarterly*,

48(2), 427–458. https://doi.org/10.25300/MISQ/2023/16700

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*, 186345. https://doi.org/10.1007/s11704-024-40231-1