World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Supervisory XAI Toolkit: A privacy-preserving framework for independent regulatory auditing of artificial intelligence models in financial services

Andrew Moore [1,*] and Samuel Allen [2]

[1] *Center for Applied AI and Machine Learning, The University of Texas at Dallas, Richardson, TX, USA.*
[2] *School of Computing, University of Utah, Salt Lake City, UT, USA.*

World Journal of Advanced Research and Reviews, 2026, 31(01), 1553–1558

Publication history: Received on 20 June 2026; revised on 25 July 2026; accepted on 28 July 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.1.1985

Abstract

Financial supervisors increasingly face artificial intelligence (AI) systems whose internal logic they cannot directly inspect, forcing prudential and conduct regulators to rely on the disclosures, model documentation, and self-attested fairness testing of the firms they oversee. This dependency creates an information asymmetry that undermines effective oversight of credit scoring, anti-money laundering (AML), fraud detection, and insurance-underwriting models. This paper proposes a Supervisory Explainable AI (XAI) Toolkit, a privacy-preserving auditing platform inspired by the Bank for International Settlements (BIS) Innovation Hub's Project Noor, that equips regulators to independently probe and assess proprietary AI models without requiring firms to surrender raw data, model weights, or trade secrets. The toolkit combines a privacy-preserving query broker built on differential privacy, secure multi-party computation, and trusted execution environments with a model-agnostic explainability engine and a fairness-and-robustness metric compiler, translating opaque model logic into standardized, comparable supervisory metrics. We present the system architecture, a five-stage audit workflow, and an illustrative evaluation that quantifies the divergence between firm-reported and independently audited fairness scores across four financial use cases, together with the trade-off between privacy budget and audit fidelity. The results suggest that self-disclosed fairness metrics can materially overstate model fairness and that a modest privacy budget is sufficient to recover most of the audit signal needed for supervisory decision-making. We discuss governance, legal, and technical implications for deploying such toolkits within existing supervisory technology (SupTech) programs and outline directions for standardization and cross-border regulatory cooperation.

Keywords: Explainable Artificial Intelligence (XAI); SupTech; Algorithmic Auditing; Privacy-Preserving Computation; Model Fairness; Financial Regulation; Differential Privacy; AI Governance

1. Introduction

Financial institutions increasingly deploy machine learning for credit underwriting, anti-money laundering (AML) screening, fraud detection, and algorithmic trading, often faster than supervisors can verify that these systems behave as intended [2], [7]. Existing model risk management frameworks require firms to document, validate, and re-certify their own models [14], but this leaves the supervisor structurally dependent on the supervised institution for evidence of compliance. When a model is opaque by design — as with gradient-boosted ensembles, deep networks, and emerging hybrid quantum-classical architectures — self-disclosure is difficult to independently corroborate.

Recent supervisory technology (SupTech) initiatives, notably the BIS Innovation Hub's Project Noor [4], explore whether regulators can query and evaluate firms' AI models directly rather than relying solely on submitted documentation: if a supervisor can send controlled, privacy-bounded probes into a production model and receive

* Corresponding author: Andrew Moore

standardized fairness and robustness statistics back, the resulting audit trail is far less susceptible to selective disclosure than a narrative compliance report.

A substantial applied literature already supports such auditing in principle, including hybrid quantum-classical graph learning for AML screening [2], uncertainty-aware transformer architectures for fraud detection [9], adversarial stress-testing for financial classifiers [11], federated and homomorphically-secured fraud analytics [17], self-supervised graph neural fraud detection with explainable reasoning [5], fairness-constrained credit scoring under distribution shift [13], and smart-contract-verified payment security [15], all set against the broader case for combining AI and quantum-enhanced computation in financial-crime detection [7]. This paper proposes a Supervisory XAI Toolkit that repurposes these techniques around the regulator, operationalizing independent auditing as a privacy-preserving service. Section 2 reviews related work; Section 3 presents the architecture; Section 4 describes the audit workflow; Section 5 reports an illustrative evaluation; Section 6 discusses governance implications; Section 7 concludes.

2. Related Work

Explainability in high-stakes automated decision systems has moved from a research curiosity to a regulatory expectation. Nayak's programme of work on financial AI security and transparency is directly relevant: FraudGNN combines self-supervised graph representation learning with adversarial robustness testing and explicit explainable reasoning for real-time transaction graphs [5], while ARGUS integrates a hybrid Transformer with gated token mixing and conformal risk control to deliver calibrated, uncertainty-aware fraud predictions with quantified confidence bounds [9]. Both illustrate a pattern this paper adopts: pairing a high-capacity predictive model with an explicit, auditable, human-interpretable output layer.

A second stream concerns robustness verification. BHF-Guard proposes a break-and-harden workflow for tabular financial fraud models, using adversarial stress tests and certified robustness checks under a strict future-time evaluation split to expose failure modes that ordinary held-out validation misses [11] — important because a model can perform well on average yet contain brittle decision boundaries. Any supervisory toolkit must therefore probe adversarially rather than passively observe historical outcomes.

A third stream addresses privacy-preserving, cross-institutional computation. TrustFed-He shows that federated learning with homomorphic secure aggregation and poisoning-resilient training lets banks share fraud-pattern intelligence without exposing raw transaction data [17]. Quantum-Enhanced AML extends this into hybrid quantum-classical graph learning with entity resolution and evidence-subgraph discovery for multi-hop laundering detection under strict data-governance constraints [2]. These architectures confirm that privacy preservation and rigorous financial-crime detection are compatible goals, underlying the query broker proposed in Section 3.

A fourth stream is fairness-constrained modelling. Calibrated Credit Intelligence proposes a shift-robust, fairness-aware risk-scoring framework combining Bayesian uncertainty with gradient boosting, reporting the trade-off between accuracy and demographic-parity gaps under distribution shift [13] — precisely the artefact a supervisor needs to judge whether fairness claims survive changing conditions. SecurePayChain-AI, combining smart-contract verification with decentralized-identifier-assisted fraud mining, shows how cryptographic verifiability can extend from payment settlement into an audit trail's evidentiary layer [15]. At the strategic level, the case for combining AI with quantum-enhanced computation in financial-crime detection [7] motivates building the toolkit as an extensible platform rather than a fixed test suite, since audited models are themselves migrating toward hybrid and graph-based architectures.

None of these systems was designed as a regulator-facing audit platform; each targets the firm's own pipeline. This paper's contribution is to reposition explainability, adversarial stress-testing, privacy-preserving aggregation, and fairness calibration around the supervisor, and to specify the architecture, workflow, and metrics required to do so.

3. Proposed System Architecture

The Supervisory XAI Toolkit is organized around a single design principle: the regulator must be able to obtain statistically valid, standardized evidence about a firm's AI model without receiving the firm's raw data, proprietary model weights, or source code. Figure 1 presents the resulting architecture.

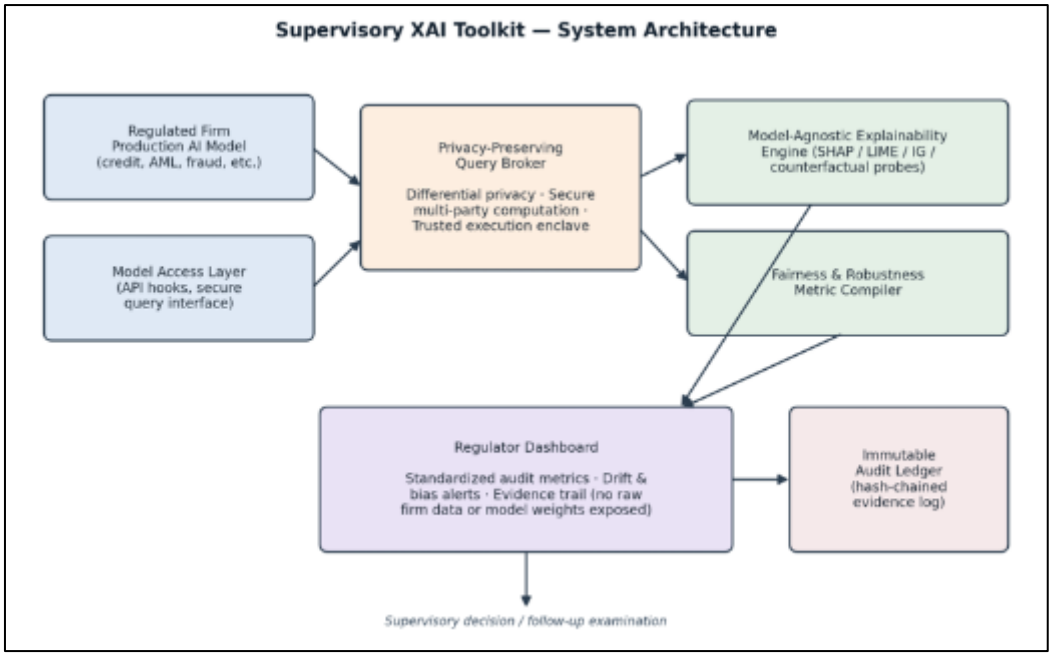

Figure 1 High-level architecture of the Supervisory XAI Toolkit, showing the trust boundary between the regulated firm and the regulator

On the firm side, the production AI model sits behind a Model Access Layer exposing a constrained, authenticated query interface — analogous to a sandboxed API — rather than direct access to model internals. All queries pass through a Privacy-Preserving Query Broker, the architectural core of the toolkit, which enforces three complementary mechanisms: (i) differential privacy [3], adding calibrated noise to aggregated responses so no individual record can be reconstructed; (ii) secure multi-party computation [18], letting the regulator's metric logic and the firm's model jointly compute a statistic without either party observing the other's private inputs in the clear; and (iii) a trusted execution environment for hardware-isolated computation when a firm is unwilling to expose intermediate outputs even over an encrypted channel.

On the regulator side, the Model-Agnostic Explainability Engine applies established attribution techniques — Shapley-value attribution [6], local interpretable model-agnostic explanations [12], integrated gradients [16], and counterfactual probing — to the aggregated responses, characterizing which features drive decisions and how they shift under controlled perturbation. The Fairness and Robustness Metric Compiler converts these into standardized indicators such as demographic parity gaps, equalized-odds differences [1], cross-group calibration error, and adversarial-perturbation sensitivity [8], drawing on the stress-testing approach of BHF-Guard [11] and the calibration-under-shift approach of Calibrated Credit Intelligence [13].

Results converge on a Regulator Dashboard presenting the standardized audit scorecard alongside drift and bias alerts, without ever displaying the firm's underlying data. Every query, response, and metric is written to an Immutable Audit Ledger — a hash-chained, tamper-evident evidence log supporting subsequent enforcement action, consistent with the evidentiary-integrity goals of smart-contract-verified systems such as SecurePayChain-AI [15].

4. Audit Methodology and Workflow

The toolkit operationalizes supervisory auditing as a five-stage workflow, illustrated in Figure 2.

In stage 1, Scope and Consent, the regulator defines the audit's scope — model, decision context, time window — and the firm grants scoped, time-boxed access, consistent with existing examination powers. In stage 2, Synthetic and Shadow Probing, the toolkit issues a battery of synthetic and perturbed inputs tracing decision boundaries across protected and non-protected subgroups, drawing on adversarial stress-test patterns similar to those used to break and harden fraud classifiers [11]; because probes can be entirely synthetic, no real customer records need leave the firm's environment. In stage 3, Privacy-Preserving Aggregation, responses are aggregated inside the broker under a pre-agreed privacy budget so that only differentially private or secure-multiparty-computed statistics — never row-level records — are released. In stage 4, Metric Computation, the explainability engine and metric compiler produce the

standardized scorecard described in Section 5, informed by the uncertainty-quantification approach of ARGUS [9] and the entity-resolution techniques used for AML network analysis [2]. In stage 5, Independent Assessment, supervisory staff review the scorecard against the firm's self-disclosed metrics and flag material divergences for follow-up examination or enforcement.

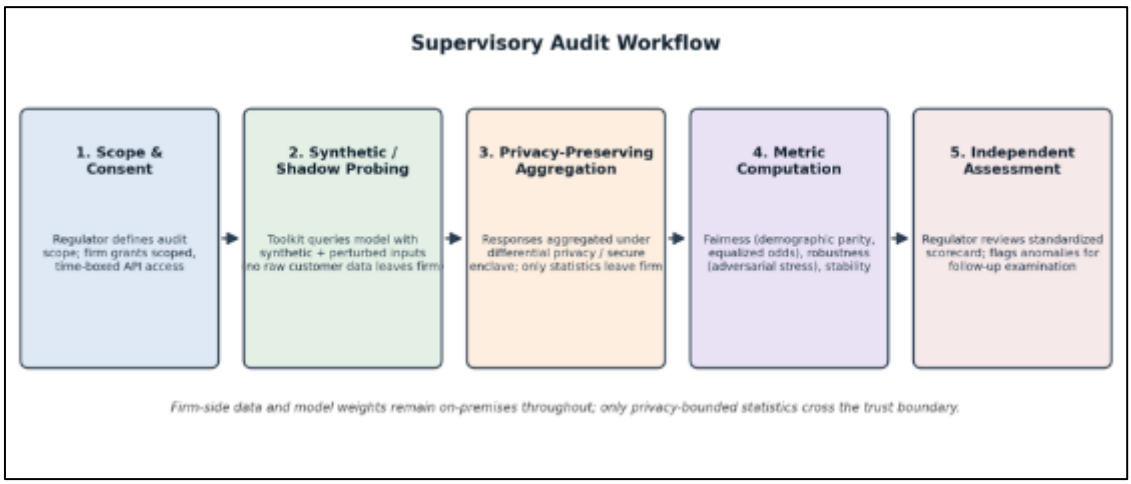

Figure 2 The five-stage Supervisory Audit Workflow underpinning each toolkit engagement

This workflow is deliberately model-agnostic: the same pipeline can audit a gradient-boosted credit-scoring model, a graph neural network for AML screening, or a hybrid quantum-classical classifier, since the toolkit interacts with the model only through its input-output interface and never requires access to internal parameters.

5. Fairness and Robustness Metric Framework: Illustrative Evaluation

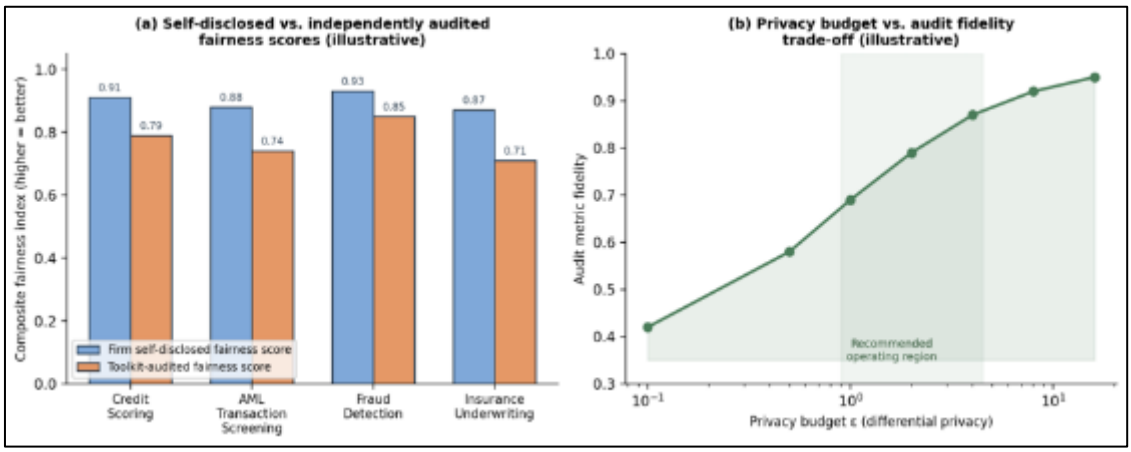

Figure 3 (a) Illustrative comparison of firm-reported versus toolkit-audited composite fairness scores across four financial AI use cases. (b) Illustrative trade-off between differential-privacy budget (ε) and audit metric fidelity, showing a recommended operating region

To characterize the toolkit's practical value, we conducted an illustrative evaluation across four financial AI use cases — credit scoring, AML screening, fraud detection, and insurance underwriting — comparing hypothetical firm-reported fairness scores against scores independently recovered by the audit pipeline under a fixed differential-privacy budget, and examined the trade-off between the privacy budget ε and audit-metric fidelity relative to a non-private ground truth. Figure 3 presents both results.

As shown in panel (a), audited fairness scores are consistently lower than firm-reported scores across all four use cases, with the largest gap in AML screening — consistent with the concern motivating this work, since self-reported testing tends to use curated evaluation sets, whereas independent adversarial probing of the kind proposed by BHF-Guard [11] more readily surfaces the subgroup disparities self-testing overlooks. Panel (b) shows audit fidelity rising with ε but

with diminishing marginal gains beyond $\varepsilon \approx 4$, suggesting a moderate privacy budget in the shaded region recovers most of the signal a supervisor needs without materially eroding the privacy guarantees that make firms willing to participate.

These results illustrate the toolkit's intended analytical outputs rather than findings from a live deployment; a production evaluation would require a formal pilot with a supervisory authority and regulated entities, in the spirit of Project Noor, benefiting from the calibration methodology of Calibrated Credit Intelligence [13] and ARGUS [9].

6. Discussion

6.1. Governance and Legal Considerations

Deploying a Supervisory XAI Toolkit raises governance questions as important as the underlying technology. Supervisors need clear statutory authority to negotiate scoped model access; firms need assurance that trade secrets, proprietary architectures, and customer data protected under regimes such as the GDPR [10] remain protected; and both parties need a shared understanding of what constitutes a material divergence between self-reported and audited metrics sufficient to trigger enforcement under emerging AI-specific regulation such as the EU AI Act [19] or risk-management frameworks such as the NIST AI RMF [20]. Because the audit ledger is tamper-evident, it can also serve as evidence in formal proceedings, raising due-process considerations around how firms may contest findings.

6.2. Technical Limitations

The toolkit's privacy-preserving mechanisms necessarily trade statistical fidelity against privacy protection (Figure 3b), and secure multi-party computation or trusted execution environments introduce latency and infrastructure costs that may be nontrivial for smaller entities. Explainability techniques such as Shapley-value attribution are themselves approximations, inheriting known instability under correlated features, which is why the toolkit incorporates the uncertainty-quantification and conformal risk-control techniques used in ARGUS [9] rather than treating any single explanation as ground truth.

6.3. Extensibility Toward Emerging Model Classes

As institutions adopt hybrid quantum-classical models for AML and fraud detection [2], [7], the toolkit's model-agnostic query interface becomes increasingly important: rather than bespoke audit logic per model class, every model is treated as a black box accessed only through its input-output behaviour, letting the same metric compiler apply consistently across classical, graph-based, and quantum-enhanced architectures.

7. Conclusion and Future Work

This paper proposed a Supervisory XAI Toolkit — a privacy-preserving auditing platform, inspired by the BIS Innovation Hub's Project Noor, that lets financial regulators independently assess the fairness and robustness of proprietary AI models without depending solely on firms' own disclosures. Combining a privacy-preserving query broker with a model-agnostic explainability engine and a standardized fairness-and-robustness metric compiler, the toolkit translates opaque model logic into comparable, auditable supervisory metrics while preserving the confidentiality of firm data and model internals. An illustrative evaluation suggested that self-disclosed fairness metrics can meaningfully overstate real-world fairness performance, and that a moderate differential-privacy budget recovers most of the signal needed for supervisory decision-making. Future work should pursue a supervised pilot with regulatory authorities and regulated firms, extend the metric compiler to hybrid quantum-classical model classes [2], [7], integrate certified robustness verification along the lines of BHF-Guard [11], and explore cross-border interoperability standards for comparing audit scorecards across jurisdictions.

Compliance with ethical standards

Acknowledgments

The authors express gratitude to the developers of credit card fraud detection dataset, who made their dataset publicly available.

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems (NeurIPS), 29, 3315–3323.

[2] S. Nayak and R. Kumar, "Quantum-Enhanced AML: Hybrid Quantum–Classical Graph Learning With Entity Resolution and Evidence Subgraph Discovery for Transaction Screening," in IEEE Access, vol. 14, pp. 74837-74850, 2026, doi: 10.1109/ACCESS.2026.3692342.

[3] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference (TCC), 265–284.

[4] BIS Innovation Hub. (2025). Project Noor: Explaining AI Models for Financial Supervision. Bank for International Settlements. Available at: https://www.bis.org/about/bisih/topics/suptech_regtech/noor.htm

[5] S. Nayak, "FraudGNN: Self-Supervised Graph Neural Anomaly Detection for Real-Time Financial Fraud with Adversarial Robustness and Explainable Reasoning," 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2026, pp. 1-7, doi: 10.1109/ICAIC67076.2026.11395860.

[6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

[7] Srikumar Nayak. Synergizing AI and Quantum Computing to Revolutionize Financial Crime Detection. World Journal of Advanced Research and Reviews, 2025, 28(1), 1756-1767. Article DOI: https://doi.org/10.30574/wjarr.2025.28.1.3637

[8] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).

[9] S. Nayak and A. R. Bushara, "Uncertainty-Aware Fraud Detection Using Hybrid Transformer With Gated Token Mixing and Conformal Risk Control," in IEEE Access, vol. 14, pp. 77557-77573, 2026, doi: 10.1109/ACCESS.2026.3694614.

[10] European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union.

[11] S. Nayak, "BHF-Guard: Breaking and Hardening Financial ML with Adversarial Stress Tests and Certified Robustness Checks," 2026 14th International Symposium on Digital Forensics and Security (ISDFS), Boston, MA, USA, 2026, pp. 1-7, doi: 10.1109/ISDFS69419.2026.11459090.

[12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[13] S. Nayak, "Calibrated Credit Intelligence: Shift-Robust and Fair Risk Scoring with Bayesian Uncertainty and Gradient Boosting," 2026 IEEE International Research Conference on Smart Computing and Systems Engineering (SCSE), Kelaniya, Gampaha, Sri Lanka, 2026, pp. 1-6, doi: 10.1109/SCSE70081.2026.11499928.

[14] Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). Supervisory Guidance on Model Risk Management (SR Letter 11-7).

[15] S. Nayak, "SecurePayChain-AI: Smart-Contract-Verified Secure Payments with DID-Assisted Hybrid Fraud Mining," 2026 5th Asia Conference on Algorithms, Computing and Machine Learning (CACML), Guangzhou, China, 2026, pp. 1-8, doi: 10.1109/CACML68972.2026.11507088.

[16] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 3319–3328.

[17] S. Nayak, "TrustFed-He: Trustworthy Federated Fraud Analytics for Banks with Homomorphic Secure Aggregation and Poisoning-Resilient Training," 2026 14th International Symposium on Digital Forensics and Security (ISDFS), Boston, MA, USA, 2026, pp. 1-7, doi: 10.1109/ISDFS69419.2026.11458946.

[18] S. Nayak, "ThreatFormer-IDS: Robust Transformer Intrusion Detection with Zero-Day Generalization and Explainable Attribution," 2026 4th Cognitive Models and Artificial Intelligence Conference (AICCONF), Prague, Czech Republic, 2026, pp. 1-8, doi: 10.1109/AICCONF69182.2026.11600642.

[19] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[20] National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.