SCIENTIFIC PAPER · AUTHOR MANUSCRIPT · 2026

CC-FEAI: A Compliance-Centric Federated and Explainable AI Architecture for Regulated Domains

Design, multi-objective trust optimization and controlled evaluation

Dr. Jami Kiran Kumar
Independent Global Technology Advisor
AI · Data · Cybersecurity · Cloud · Quantum
QuantumReadyCheck.com

Article type: Design-science and controlled-evaluation paper
Status: Author manuscript; not peer reviewed
Research origin: Derived from the author's doctoral thesis
Original contribution: CC-FEAI architecture and multi-objective trust formulation

Authored and authorised by Jami Kiran Kumar. This manuscript is publication-oriented but has not been accepted by or submitted through peer review at a named journal.

ABSTRACT

Abstract

Background. Centralized machine-learning architectures create privacy, sovereignty and concentration risks in regulated domains, while black-box predictions hinder human oversight and audit. Federated learning (FL) and explainable AI (XAI) address different parts of this problem but are commonly implemented without integrated lifecycle governance.

Objective. To specify and evaluate CC-FEAI, a compliance-centric federated and explainable AI architecture that integrates privacy preservation, explanation, security, regulatory evidence and model lifecycle management.

Method. A design-science approach was used to construct a layered architecture comprising local data nodes, FedAvg/FedAvgM training, SHAP/LIME explanations, compliance controls, differential privacy, robust aggregation and lifecycle monitoring. Controlled experiments used synthetic and enterprise-like structured and time-series datasets distributed across IID and non-IID nodes. Classification performance, explanation characteristics and compliance-readiness controls were assessed.

Results. In the reported configuration, FedAvgM achieved 95% accuracy, 94% precision, 92% recall and 93% F1. The integration generated local and global explanation artefacts and a traceable control-evidence structure. These results demonstrate controlled technical feasibility; they do not establish production effectiveness or legal compliance across domains.

Conclusion. CC-FEAI provides an integrated research architecture for optimizing predictive utility alongside privacy, explainability, robustness, compliance readiness and operational efficiency. External validation on real regulated datasets, human-centred explanation studies and formal privacy/security analysis remain necessary.

Keywords: federated learning; explainable AI; trustworthy AI; compliance by design; differential privacy; SHAP; LIME; EU AI Act; GDPR; NIS2; adversarial robustness.

Problem and research questions

AI systems used in healthcare, financial services, public administration and critical infrastructure must operate under simultaneous performance, privacy, security, transparency and governance constraints. Centralized training can conflict with data-localization and minimization requirements. High-performing nonlinear models can also be difficult to interpret, weakening informed reliance, contestability and post-event accountability.

Federated learning trains a shared model from decentralized data by aggregating local updates rather than raw records. Explainability methods such as SHAP and LIME provide feature-level accounts of model behaviour. Neither capability alone establishes trustworthy operation. FL introduces update leakage, poisoning, heterogeneity and coordination risks; explanations can be unstable, misleading or unsuitable for the stakeholder.

This study addresses the absence of a unified, lifecycle-oriented architecture in which privacy, explanation, security and regulatory evidence are co-designed. **RQ1:** How can FL, XAI and compliance controls be integrated into one architecture? **RQ2:** How can trust properties be represented in a multi-objective formulation? **RQ3:** What technical performance is observed in a controlled distributed evaluation? **RQ4:** What evidence and limitations must govern claims of deployment readiness?

The paper's contribution is CC-FEAI: a layered design linking local data processing, federated optimization, explanation generation, zero-trust and adversarial controls, regulatory mapping and continuous model lifecycle management.

Research stance. Compliance is not a model metric and cannot be "proven" by architecture alone. CC-FEAI operationalizes evidence and controls that support qualified legal, risk and assurance judgments.

Federation, explanation and AI governance

2.1 Federated learning

McMahan et al. introduced Federated Averaging as a communication-efficient approach to learning from decentralized, potentially unbalanced and non-IID data. Subsequent research has emphasized privacy leakage from gradients, secure aggregation, client drift, personalization and Byzantine robustness.

2.2 Explainable AI

LIME explains a prediction through a locally interpretable surrogate model. SHAP unifies additive feature-attribution methods using Shapley-value principles. Both are post-hoc methods; their usefulness depends on fidelity, stability, audience and decision context.

2.3 Regulatory and risk frameworks

The EU AI Act establishes risk-based obligations for actors and applicable systems, including requirements for certain high-risk systems. GDPR governs personal-data processing. NIS2 establishes cybersecurity risk-management obligations for entities in scope. NIST AI RMF organizes voluntary risk practice around Govern, Map, Measure and Manage.

Stream	Strength	Residual gap
Federated learning	Reduced raw-data centralization	Update leakage, poisoning and accountability
Post-hoc XAI	Local and global model insight	Fidelity, stability and human usefulness
Privacy-enhancing technology	Formal or protocol protection	Utility, cost and implementation risk
Regulatory mapping	Structured obligation coverage	Can remain disconnected from executable controls

Research gap. Prior components are often evaluated independently; CC-FEAI studies their architectural integration and makes the trade-offs explicit.

Design-science methodology

Stage	Activity	Output
Problem identification	Analyze centralization, opacity and compliance-after-design failures	Trust requirements
Objective definition	Specify privacy, explanation, compliance, robustness and performance goals	Design criteria
Artefact construction	Integrate FL, XAI, governance and lifecycle components	CC-FEAI architecture
Demonstration	Implement distributed controlled environment	Executable workflow
Evaluation	Measure classification, explanation and control-readiness outputs	Results and limitations
Communication	Translate findings into scientific and executive publications	Reproducible claims

Evaluation questions

- Does the architecture complete distributed training and aggregation across heterogeneous nodes?
- How do FedAvg and FedAvgM perform under the selected IID/non-IID conditions?
- Can the pipeline generate decision-level and model-level explanation artefacts?
- Can obligations be traced to versioned controls, owners, tests and records?
- What communication, explanation and generalizability constraints remain?

Epistemic boundary. The evaluation is an artefact demonstration using controlled data. It supports internal validity for the tested configuration, not external validity across all regulated applications.

CC-FEAI system architecture

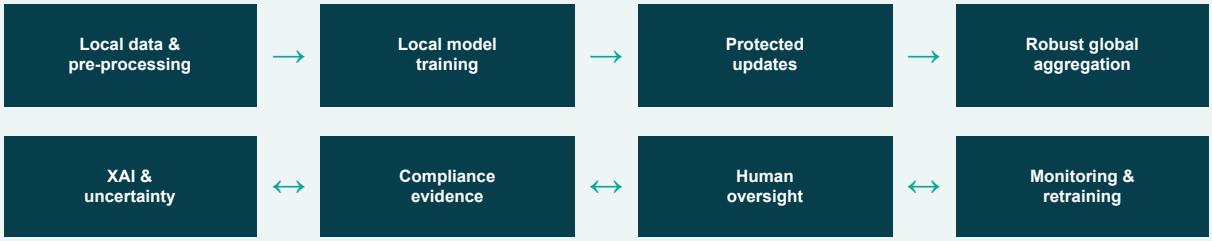

Figure 1. CC-FEAI couples distributed learning to explanation, control evidence and lifecycle governance.

Component	Primary function	Key risk
Client node	Local data processing and optimization	Compromise, bias, poor data quality
Aggregator	Weighted or momentum-based model update	Single point of control or manipulation
Explainability engine	Local/global attribution and reason artefacts	Unfaithful or unstable explanation
Governance engine	Classification, policy, logs and evidence	Checklist without effective control
Lifecycle manager	Version, drift, incident and retraining	Undetected change in operating context

Design proposition 1. Trustworthy operation requires bidirectional coupling: governance constrains training and deployment, while technical telemetry continuously updates governance evidence.

Global learning under local constraints

$$F(w) = \Sigma k\ (nk / n)\ Fk(w) \quad \text{and} \quad wt+1 = \Sigma k\ (nk / n)\ wk,t+1$$

FedAvg aggregates locally optimized models in proportion to client sample counts. FedAvgM adds server momentum to improve convergence stability in some heterogeneous settings.

Privacy and security extensions

- **Secure aggregation:** protects individual client updates during combination.
- **Differential privacy:** clips sensitivity and adds calibrated noise, parameterized by a privacy budget (ε, δ).
- **Client authentication and attestation:** restrict participation to authorized environments.
- **Byzantine-resilient aggregation:** limits malicious or corrupted update influence.
- **Anomaly detection:** monitors data, update and output deviations.

$$Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D') \in S] + \delta$$

Trade-off	Effect	Required report
Privacy noise	May reduce utility and slow convergence	Privacy accountant and performance curve
More local epochs	Less communication, greater client drift	Rounds, divergence and subgroup error
Robust aggregation	Attack resistance, possible benign-update loss	Attack/false-rejection evaluation
Client sampling	Efficiency, participation imbalance	Coverage and representation analysis

Explanation as a measured output

CC-FEAI generates local explanations for individual decisions and aggregated patterns for model review. SHAP estimates additive feature contributions; LIME fits a sparse local surrogate around a selected prediction.

$$f(x) \approx \varphi_0 + \Sigma_i \varphi_i \quad \text{where } \varphi_i \text{ represents feature contribution}$$

Property	Operational definition	Test
Fidelity	Explanation reflects model behaviour in its scope	Surrogate agreement / perturbation test
Stability	Similar cases produce appropriately similar explanations	Repeated sampling and neighbourhood variation
Comprehensibility	Intended user accurately understands it	User study and comprehension task
Actionability	Explanation supports a valid decision or review	Decision-quality experiment
Completeness	Material uncertainty and limitations are included	Expert audit

Design proposition 2. Explanation quality must be evaluated separately from predictive accuracy; an accurate model can produce unusable explanations, and a persuasive explanation can be unfaithful.

Federated complication

Local feature distributions may differ. A global explanation can hide node-specific behaviour, while local explanations may disclose sensitive attributes. CC-FEAI therefore requires per-node analysis, aggregation controls and privacy review for explanation artefacts.

Control-evidence mapping

CC-FEAI maps applicable requirements to technical and organizational controls. The mapping is contextual and maintained across versions.

Lifecycle	Control objective	Evidence artefact
Purpose and classification	Define use, actors, risks and lawful boundaries	Use-case record, applicability and impact assessment
Data	Quality, minimization, provenance and rights	Data sheet, lineage, processing record and tests
Development	Performance, privacy, robustness and explanation	Experiment, privacy and threat reports
Validation	Independent TEVV against thresholds	Validation protocol and signed results
Deployment	Human authority, access and rollback	Operating procedure and authorization
Monitoring	Drift, incidents and material change	Telemetry, alerts, incident and change logs

> **Design proposition 3.** Compliance readiness improves when evidence is generated as a by-product of controlled lifecycle events rather than reconstructed for periodic audit.

EU AI Act, GDPR and NIS2 scope and duties differ. The architecture can support aligned controls, but qualified interpretation remains necessary. In particular, regulatory classification must precede control claims.

8. UNIFIED TRUST OBJECTIVE

Multi-objective optimization

$$J(w)=\lambda_1 F(w)+\lambda_2 P(w)+\lambda_3 X(w)+\lambda_4 C(w)+\lambda_5 R(w)+\lambda_6 O(w)$$

F represents predictive loss; P privacy risk; X opacity/explanation loss; C compliance gap; R robustness risk; and O operational cost. λ weights express governed preferences or constraints.

Term	Candidate metric	Constraint example
F(w)	Cross-entropy, calibration, subgroup error	Minimum recall for safety-critical class
P(w)	Privacy budget and empirical leakage	Maximum ε and attack-success threshold
X(w)	Infidelity, instability, comprehension failure	Minimum stability and user accuracy
C(w)	Unmet control/evidence obligations	No unresolved critical control gap
R(w)	Worst-case and poisoning degradation	Maximum performance loss under attack
O(w)	Communication, latency, compute, human effort	Operational service-level ceiling

Design proposition 4. Trust weights are governance parameters because they allocate protection, burden and residual risk among stakeholders; they should not be selected by the model-development team alone.

A constrained formulation may be preferable in high-stakes systems: optimize utility subject to hard privacy, safety, oversight and legal constraints, rather than permitting compensation between incompatible values.

Controlled evaluation protocol

Element	Configuration reported in thesis
Environment	Multi-node distributed, hybrid edge/cloud
Frameworks	TensorFlow / PyTorch
Algorithms	FedAvg and FedAvgM; centralized baseline
Data	Synthetic and enterprise-like structured, time-series and sensitive-regulatory data
Distribution	IID and non-IID partitions
Explainability	SHAP and LIME
Performance	Accuracy, precision, recall, F1 and AUC-ROC
Trust assessment	Interpretability, feature stability, decision transparency and control readiness

Threats to validity

- **Construct:** qualitative “high” explanation ratings require reproducible quantitative definitions.
- **Internal:** algorithm, hyperparameter and data-generation details must be fixed for replication.
- **External:** synthetic data do not reproduce full clinical, financial or administrative complexity.
- **Conclusion:** a single configuration cannot establish universal superiority.

Replication package recommended: code, seeds, data generator, client partitions, hyperparameters, compute environment, privacy budget, attack scenarios, explanation tests and raw result tables.

Controlled performance and interpretation

Metric	Reported FedAvgM result	Interpretation
Accuracy	95%	Strong overall correctness in tested configuration
Precision	94%	High positive predictive value
Recall	92%	Some false negatives remain
F1	93%	Balanced precision–recall performance

The thesis reports that FedAvgM produced the strongest classification results among evaluated approaches and that non-IID data modestly reduced accuracy while providing a more realistic robustness challenge. SHAP identified high-impact features and supported decision traceability; LIME provided complementary local views.

Compliance-readiness outputs

The architecture produced control mappings for privacy, transparency, risk classification, security and logging. These demonstrate artefact coverage, not a legal certification.

Finding 1. Integrated execution of FL and XAI is technically feasible in the controlled setting.

Finding 2. FedAvgM achieved strong reported classification performance, but comparative significance and cross-domain generalization require additional experiments.

Finding 3. Lifecycle evidence can be structured around architectural events, but its regulatory sufficiency remains contextual.

Contributions, trade-offs and implications

11.1 Scientific contribution

CC-FEAI's differentiator is integration: it treats accuracy, privacy, explanation, compliance readiness and robustness as interacting properties rather than sequential workstreams.

11.2 Technical implication

Reducing raw-data movement changes rather than removes the attack surface. Secure aggregation, privacy accounting, client integrity and robust aggregation must be measured alongside model utility.

11.3 Organizational implication

Federation distributes operational responsibility. Each node requires accountable ownership, local data governance, security operations and participation rules.

11.4 Regulatory implication

Compliance by design should be understood as continuous evidence engineering. The same control may support multiple frameworks, but mappings must preserve differences in scope, actor and legal effect.

11.5 Human implication

Explanations create value only when they improve calibrated reliance, review or contestability. Human-centred studies are therefore essential.

> **Author's synthesis.** Trustworthiness is an emergent system property. It cannot be inferred from any single algorithm, privacy technique, explanation method or compliance checklist.

Research boundaries and next studies

Limitations

- Controlled synthetic and enterprise-like data limit external validity.
- Dataset generation, model architecture and hyperparameter detail require publication-grade replication artefacts.
- Communication overhead and client dropout at large scale remain insufficiently characterized.
- Explanation evaluation requires quantitative fidelity/stability metrics and human studies.
- Multimodal data and domain-specific workflows were not evaluated.
- Formal privacy guarantees and empirical attack resistance require joint reporting.
- Compliance mappings do not constitute legal determinations.

Future research

1. Prospective pilots using governed healthcare, finance and public-sector datasets.
2. Cross-silo evaluations with realistic node failures and non-IID drift.
3. Comparison of FedAvgM with adaptive and personalized FL methods.
4. Privacy–utility frontiers across multiple ε/δ budgets.
5. Poisoning, inference and Byzantine attack benchmarks.
6. Stakeholder-specific explanation comprehension and decision-quality trials.
7. Automated evidence graphs linked to lifecycle events.
8. Energy and carbon accounting for distributed training.

Priority experiment: A preregistered multi-institution study comparing centralized, federated and CC-FEAI variants using the same task, governance boundary and outcome measures.

Conclusion and declarations

CC-FEAI addresses a central challenge in regulated AI: how to improve shared models without centralizing sensitive data, while maintaining explanations, security controls, human authority and regulatory evidence.

The controlled evaluation reports strong FedAvgM classification metrics and demonstrates that FL, XAI and a compliance-centric lifecycle can operate as a cohesive artefact. The evidence supports technical feasibility and a research programme; it does not justify automatic compliance or unrestricted real-world deployment.

The multi-objective trust model reframes system quality. Accuracy remains essential, but privacy leakage, explanation failure, control gaps, adversarial risk and operational cost also determine whether an AI system is fit for consequential use.

> **Conclusion.** Trustworthy AI in regulated domains should be designed as a continuously governed distributed system—not evaluated as a model plus a retrospective checklist.

Declarations

Funding: No external funding declared. **Conflicts of interest:** None declared. **Research origin:** Derived from the author's doctoral thesis. **Data:** Controlled synthetic and enterprise-like datasets as described; a public replication package is recommended. **Ethics:** No claim of human-subject experimentation is made. **Peer review:** This author manuscript has not undergone peer review.

Authorship

Dr. Jami Kiran Kumar conceived the CC-FEAI framework, developed the compliance-centric and multi-objective trust model, conducted the thesis evaluation and authored this scientific manuscript. The document is fully authored and authorised by the named author.

References

1. European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*.
2. European Parliament and Council. (2022). Directive (EU) 2022/2555 (NIS2 Directive). *Official Journal of the European Union*.
3. European Parliament and Council. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Official Journal of the European Union*.
4. Kumar, J. K. (2026). *Federated & Explainable AI: A Compliance-Centric Framework for Trustworthy AI in Regulated Domains*. Doctoral thesis / author manuscript.
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
6. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of AISTATS, PMLR 54*, 1273–1282.
7. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of KDD '16*.

Correspondence: Dr. Jami Kiran Kumar · Independent Global Technology Advisor · QuantumReadyCheck.com

This author manuscript is publication-ready in structure but requires venue-specific formatting, independent peer review and—where requested—a complete replication package before formal scholarly publication.