FORTEX: A Formal Framework for Optimizing the Explainability–Efficiency Trade-off in High-Stakes AI with Trust-Metric Federated Governance for Integrity and Accountability Across Silos

Mohan Siva Krishna Konakanchi
mohansivakrishna16@gmail.com

Abstract—**High-stakes AI systems in healthcare, finance, critical infrastructure, and public-sector decision support must balance three competing requirements: predictive performance, operational efficiency (latency, compute, and energy), and explainability suitable for audit and governance. Existing approaches often treat these dimensions independently or rely on post-hoc explanations that are unstable, costly, or poorly aligned with operational constraints. This paper proposes *FORTEX*, a practical formal framework for quantifying and optimizing the explainability–efficiency trade-off in high-stakes AI. FOR- TEX introduces lightweight, governance-oriented metrics for explanation quality (stability, succinctness, and domain concept alignment) and efficiency (latency, memory, and energy proxies), and combines them with task performance metrics into a unified optimization and selection process that yields policy-compliant operating points. To address real-world deployment where train- ing data are distributed across organizational silos, FORTEX includes a trust metric-based federated learning (FL) governance layer that ensures integrity and accountability by scoring client reliability through multi-signal trust indicators, applying robust aggregation with bounded influence, and producing auditable decision logs without centralizing sensitive data. Experiments using federated simulations over heterogeneous partitions demonstrate that FORTEX can produce Pareto-efficient model configurations that preserve utility while substantially improving explanation stability and resource efficiency. The trust-governed federated layer reduces degradation under noisy and adversarial clients and improves traceability for compliance. FORTEX provides a deployable blueprint for high-stakes AI teams to systematically manage explainability, efficiency, and performance under practical constraints.**

Index Terms—**High-stakes AI, explainability, efficiency, governance, federated learning, trust metrics, accountability, robust aggregation, model selection.**

I. Introduction

High-stakes AI applications increasingly shape consequential decisions: triage and risk scoring in healthcare, fraud and credit decisions in finance, anomaly detection in critical infrastructure, and adjudication support in public-sector workflows. In these settings, a model is not evaluated solely by predictive accuracy. Stakeholders require *explainability* to support review and accountability, and *efficiency* to meet operational budgets, latency constraints, and energy consumption limits. The challenge is that improving one dimension can degrade another. For example, complex models may improve predictive performance but increase inference costs and reduce transparency, while highly interpretable models may sacrifice accuracy. Post-hoc explainers can add computational overhead and may be unstable across small input changes, undermining trust.

In parallel, high-stakes deployments often span multiple institutions or business units, each holding sensitive data that cannot be centralized. Federated learning (FL) enables collaborative training without transferring raw data; however, standard FL does not guarantee integrity, robust behavior under unreliable clients, or accountability suitable for regulated environments. Governance requires auditable records of participation, update quality controls, and decision processes. This paper proposes *FORTEX*, a formal and practical frame- work for optimizing the explainability–efficiency trade-off in
high-stakes AI. FORTEX defines:

- measurable, lightweight explainability metrics suitable for governance and audit;
- operational efficiency metrics that reflect real deployment constraints;
- a selection and optimization pipeline that yields policy-compliant operating points;
- a trust metric-based federated governance layer for integrity and accountability across silos.

FORTEX is designed to be implementable without complex formulas or diagrams and to support governance workflows where decisions must be justified to non-technical oversight committees.

A. Motivating Challenges

Explainability is necessary but not free. Explanations that are too long, inconsistent, or computationally expensive are impractical. High-stakes teams need measurable indicators of explanation quality and cost.

Efficiency is increasingly constrained. Edge deployments, mobile clinical devices, and production pipelines with strict

service-level objectives require predictable latency and resource budgets. Efficiency must be treated as a first-class objective.

Data and responsibility are distributed. Multi-entity systems require federated training and shared governance. Without trust controls, FL can become a weak point in integrity and accountability.

B. Contributions

This paper offers three main contributions:

- **FORTEX metrics and objectives:** A lightweight metric suite that quantifies explainability and efficiency and integrates them with task performance for high-stakes decision support.
- **Optimization and operating-point selection:** A practical procedure to obtain Pareto-efficient candidates and choose policy-compliant models with transparent justification.
- **Trust-governed federated framework:** A trust metric-based FL governance layer that enforces integrity via robust aggregation and bounded influence, and provides accountability through auditable logs.

C. Paper Organization

Section II reviews related work. Section III defines requirements and formalizes the trade-off problem. Section IV introduces FORTEX metrics and selection. Section V presents trust-metric federated governance. Section VI describes methodology and experiments. Section VII presents results and discussion. Section VIII concludes.

II. Related Work

A. Explainable AI

Model-agnostic explainers such as LIME and SHAP provide local feature attributions, while broader literature argues for interpretable model design in high-stakes contexts. However, explainers vary in stability and runtime cost, and many evaluations focus on interpretability in isolation rather than under operational constraints.

B. Efficiency and Model Compression

Efficiency improvements come from pruning, quantization, knowledge distillation, and architectural choices. While these can reduce latency and memory, they may change model behavior, affecting both accuracy and interpretability. Few frameworks explicitly integrate explainability metrics into efficiency optimization.

C. Federated Learning, Robustness, and Privacy

Federated averaging introduced scalable decentralized training. Secure aggregation and differential privacy address privacy leakage. Robust aggregation methods tolerate Byzantine clients. Yet, regulated deployments also require auditable governance: why updates were trusted and how decisions were made, beyond pure statistical robustness.

D. Trust and Accountability

Interpretability and trust are intertwined. Trust often depends on stability, transparency, and traceability. In high-stakes systems, accountability requires auditable records and consistent governance policies. FORTEX treats trust and accountability as measurable properties integrated into the training and selection lifecycle.

III. Problem Definition and Requirements

A. High-Stakes AI Requirements

We assume a predictive task with significant consequences (e.g., clinical escalation, financial approval). The system must satisfy:

- **Performance:** predictive utility suitable for deployment thresholds.
- **Explainability:** stable, succinct, and domain-aligned explanations.
- **Efficiency:** acceptable latency, memory footprint, and energy proxies under deployment constraints.
- **Governance:** integrity against faulty/adversarial contributions and accountability through audit trails.

B. Operational Setting Across Silos

Training data are distributed across K silos. Each silo performs local training and shares updates. Constraints include privacy (no raw data sharing), heterogeneity (non-IID data), and intermittency (clients may drop out).

C. Trade-off as a Selection Problem

Rather than enforcing a single objective, FORTEX produces a set of candidate models and selects operating points based on policy. This matches real governance: committees set minimum performance thresholds and maximum acceptable costs, and require justification for trade-offs.

IV. FORTEX Framework

A. Overview

FORTEX consists of:

1) **Metric layer:** compute performance, explainability, and efficiency metrics.
2) **Candidate generation:** train or derive multiple candidate configurations (e.g., baseline, compressed, interpretable variants).
3) **Multi-objective selection:** identify Pareto-efficient candidates and choose policy-compliant operating points.
4) **Federated governance:** apply trust scoring, integrity controls, and accountability logs during federated training.

B. Explainability Metrics (Lightweight)

FORTEX uses three explanation quality metrics intended to be measurable, auditable, and comprehensible.

X1: Explanation stability. Evaluate how similar explanations remain under small input perturbations or reasonable preprocessing differences. Stability is critical in high-stakes settings because unstable explanations can mislead reviewers.

X2: Explanation succinctness. Quantify how concise an explanation is, e.g., how many features or concepts are needed to explain a decision. Excessively long explanations are difficult to review and operationalize.

X3: Domain concept alignment. Group features into domain-relevant concepts (e.g., vitals, claims history, sensor regimes). Measure whether explanations emphasize plausible concepts for the decision type. This encourages explanations that match domain mental models without requiring complex ontology machinery.

C. Efficiency Metrics (Deployment-Oriented)

FORTEX evaluates efficiency using lightweight, deployment-friendly measures.

E1: Inference latency. Measure average and tail latency on representative hardware or a proxy environment.

E2: Memory footprint. Approximate memory through model size and runtime allocations where feasible.

E3: Energy proxy. Use compute proxy measures (e.g., multiply-accumulate counts or runtime power estimates) rather than complex energy modeling.

D. Performance Metrics

Performance is task-specific and uses standard measures suitable for governance:

- discrimination and calibration proxies,
- sensitivity at fixed alert/approval rates,
- fairness indicators when relevant.

E. Composite Scoring and Policy Constraints

FORTEX does not collapse everything into a single opaque score by default. Instead, it uses:

- **hard constraints:** minimum performance, maximum latency, minimum stability;
- **transparent weighted scoring:** for tie-breaking and prioritization within compliant candidates.

This improves explainability of the selection process itself.

F. Operating Points

FORTEX outputs three operating points for governance:

- **High Performance:** maximize predictive utility while meeting minimum governance thresholds.
- **Balanced:** near-maximum utility with improved explainability and efficiency.
- **High Explainability–Efficiency:** prioritize interpretability and resource constraints in the highest-risk deployments.

TABLE I
FORTEX OPERATING POINT PROFILES (NARROW SUMMARY)

Profile	Explainability	Efficiency
High Perf.	Meets min	Meets min
Balanced	High	High
High X–E	Best	Best

V. TRUST-METRIC FEDERATED GOVERNANCE FOR INTEGRITY AND ACCOUNTABILITY

A. Motivation

In multi-silo settings, federated learning must handle unreliable clients and provide auditable training decisions. Robust aggregation helps, but governance demands *traceability* and *accountability*. FORTEX introduces trust metrics as a structured governance instrument.

B. Trust Signals

Each client receives a trust score per round derived from multiple signals. Signals are selected to be explainable to auditors.

T1: Cohort update consistency. Compare compressed update sketches to detect extreme deviations.

T2: Historical reliability. Track past anomaly frequency, stability, and compliance with protocol versions.

T3: Utility validation. Assess update impact on a policy-approved validation set or via aggregated client-reported validation summaries.

T4: Robust aggregation rank. Integrate robust selection output as an explicit signal rather than hidden logic.

T5: Provenance metadata. Use non-sensitive provenance artifacts: model version, preprocessing signature, training steps, and optional attestation token.

C. Trust-Weighted Aggregation with Bounded Influence

FORTEX applies:

- robust pre-filtering to remove extreme outliers,
- trust-weighted aggregation to reduce influence of low-trust updates,
- caps on maximum influence so no client dominates.

This design supports fairness across institutions and reduces risk from both faults and centralization of influence.

D. Accountability Logs

Each round produces a non-sensitive log containing:

- participating clients (hashed identifiers),
- trust score summaries and component signals,
- decisions (accepted, down-weighted, quarantined),
- aggregation configuration version.

This supports incident review and regulatory audits without exposing raw data.

E. Interaction with Explainability–Efficiency Selection

Trust health influences selection policy:

- when trust is low (high uncertainty), select more interpretable and stable candidates,
- when trust is high, allow more performance-focused candidates under compliance thresholds.

This aligns governance risk with model operating points.

VI. Methodology

A. Candidate Generation

FORTEX generates multiple candidates by varying:

- model class (e.g., baseline deep model vs constrained interpretable variant),
- compression level (pruning/quantization/distillation),
- explanation method configuration (e.g., top-k attribution size),
- calibration and thresholding for operational constraints.

B. Federated Training Protocol

Each round:

1) select eligible clients via policy checks;
2) train locally for a fixed number of steps;
3) optionally apply privacy controls (secure aggregation / differential privacy);
4) compute trust scores from permitted signals;
5) aggregate with robust filtering + trust weighting + influence caps;
6) evaluate candidates on performance, explainability, and efficiency metrics;
7) update operating point recommendations and logs.

C. Experimental Setup

We evaluate FORTEX using federated simulations on standard tabular and image proxy tasks representative of high-stakes decision support. Data are partitioned non-IID across K clients. We simulate:

- **Noisy clients:** preprocessing errors and label noise;
- **Adversarial clients:** crafted updates attempting to degrade performance or destabilize explanations;
- **Intermittent clients:** dropout patterns.

Baselines:

- centralized training (upper bound, when available),
- FedAvg,
- robust-only FL (robust aggregation without trust scoring and accountability).

D. Evaluation

Metrics include:

- task utility: AUC and thresholded sensitivity proxies,
- explainability: stability, succinctness, concept alignment,
- efficiency: latency and memory proxies,
- governance: degradation under faults and audit completeness.

TABLE II
Evaluation Summary (Minimal Columns)

Aspect	Measures
Performance	AUC, sensitivity proxy, calibration proxy
Explainability	Stability, succinctness, concept alignment
Efficiency	Latency, memory proxy, energy proxy
Governance	Fault degradation, trust health, audit logs

VII. Results and Discussion

A. Trade-off Behavior and Pareto Efficiency

Across tasks and partitions, FORTEX identifies candidates that dominate naive baselines: models that maintain near-maximum performance while substantially improving explanation stability and reducing latency via compression and operating-point tuning. As expected, aggressive compression can reduce performance and may also change explanation patterns. FORTEX mitigates this by measuring explanation stability and concept alignment rather than assuming interpretability is preserved.

Balanced operating points frequently provide a strong practical compromise: explanation stability improves materially and latency decreases, with only minor reductions in utility compared to high-performance configurations. High explainability–efficiency configurations further improve stability and cost but may reduce utility; these are appropriate for the most risk-sensitive or resource-limited deployments.

B. Effect of Trust-Governed Federated Training

FedAvg degrades under noisy and adversarial clients, sometimes exhibiting unstable explanation behavior even when performance remains acceptable. Robust-only aggregation reduces the impact of extreme outliers but may not address persistent medium-strength faults and offers limited governance traceability. FORTEX's trust scoring improves resilience by consistently down-weighting unreliable clients based on combined signals and by limiting influence through caps.

Importantly, audit logs improve accountability: governance teams can identify participation patterns, anomalous rounds, and remediation triggers. This is crucial in regulated settings where model changes must be justified.

C. Explainability Quality as a Governance Artifact

FORTEX treats explainability as part of governance, not merely user interface. Stability and concept alignment improve confidence that explanations can be used in audits and incident review. Succinct explanations reduce reviewer burden. These indicators are not perfect proxies for human understanding, but they provide measurable, monitorable signals that can be attached to model releases.

D. Efficiency as a First-Class Constraint

In many deployments, latency and memory constraints are non-negotiable. FORTEX enables selection under hard constraints, preventing governance from approving models that cannot meet operational budgets. This reduces operational failure modes where an "accurate" model is unusable at runtime.

E. Limitations

FORTEX does not guarantee fairness or safety automatically; it provides a structured selection framework that makes trade-offs explicit and auditable. The quality of concept alignment depends on the appropriateness of concept groupings and domain curation. Privacy-preserving modes reduce observability for trust scoring; FORTEX accommodates this by shifting emphasis to robust rank signals and provenance metadata, but some trust resolution is inevitably reduced.

VIII. Conclusion

This paper introduced *FORTEX*, a formal and practi- cal framework for optimizing the explainability–efficiency trade-off in high-stakes AI. FORTEX defines lightweight, governance-oriented metrics for explainability and efficiency, integrates them with performance to select policy-compliant operating points, and extends the framework to multi-silo deployments with a trust metric-based federated governance layer that enforces integrity and accountability. Experimental simulations under heterogeneity and client faults show that FORTEX produces Pareto-efficient configurations, improves explanation stability and resource efficiency with minimal utility loss, and reduces vulnerability to noisy and adversarial updates while improving auditability.

Future work includes prospective deployment studies, integration of richer fairness and uncertainty auditing, and stronger privacy-preserving trust signals under secure aggregation constraints.

Acknowledgment

The author thanks the research community for foundational advances in federated learning, privacy, robust aggregation, model compression, and interpretable machine learning.

References

[1] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in *Proc. ACM CCS*, 2015.
[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD*, 2016.
[3] M. Abadi *et al.*, “Deep learning with differential privacy,” in *Proc. ACM CCS*, 2016.
[4] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
[5] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. NeurIPS*, 2017.
[6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agu¨era y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. AISTATS*, 2017.
[7] K. Bonawitz *et al.*, “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. ACM CCS*, 2017.
[8] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in *Proc. NeurIPS*, 2017.
[9] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proc. ICML*, 2017.
[10] R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning: A client level perspective,” *arXiv preprint arXiv:1712.07557*, 2017.
[11] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
[12] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” in *Proc. ICLR*, 2016.
[13] A. G. Howard *et al.*, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
[14] B. Jacob *et al.*, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in *Proc. CVPR*, 2018.
[15] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” *arXiv preprint arXiv:1812.06127*, 2018.
[16] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, “The hidden vulnerability of distributed learning in Byzantine-robust federated learning,” in *Proc. ICML*, 2018.
[17] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, 2019.
[18] P. Kairouz *et al.*, “Advances and open problems in federated learning,” *arXiv preprint arXiv:1912.04977*, 2019.