ENGINEERING STANDARDS LIBRARY / ARTIFICIAL INTELLIGENCE

Federated Learning & Privacy-Preserving ML Standard

The architecture of collaborative AI training has failed.

PERMANENT DOCUMENT ID

MYTHOS-AI-0015

PUBLICATION

Candidate Standard
Version 1.0.0-candidate

ASSURANCE PROFILE

Standard / Advanced
High Assurance

PERMANENT PUBLICATION IDENTITY

Federated Learning & Privacy-Preserving ML Standard

DOCUMENT ID
MYTHOS-AI-0015

VERSION
1.0.0-candidate

STATUS
Candidate Standard

PUBLISHER
Mythos Systems

PUBLICATION DATE
2026-07-24

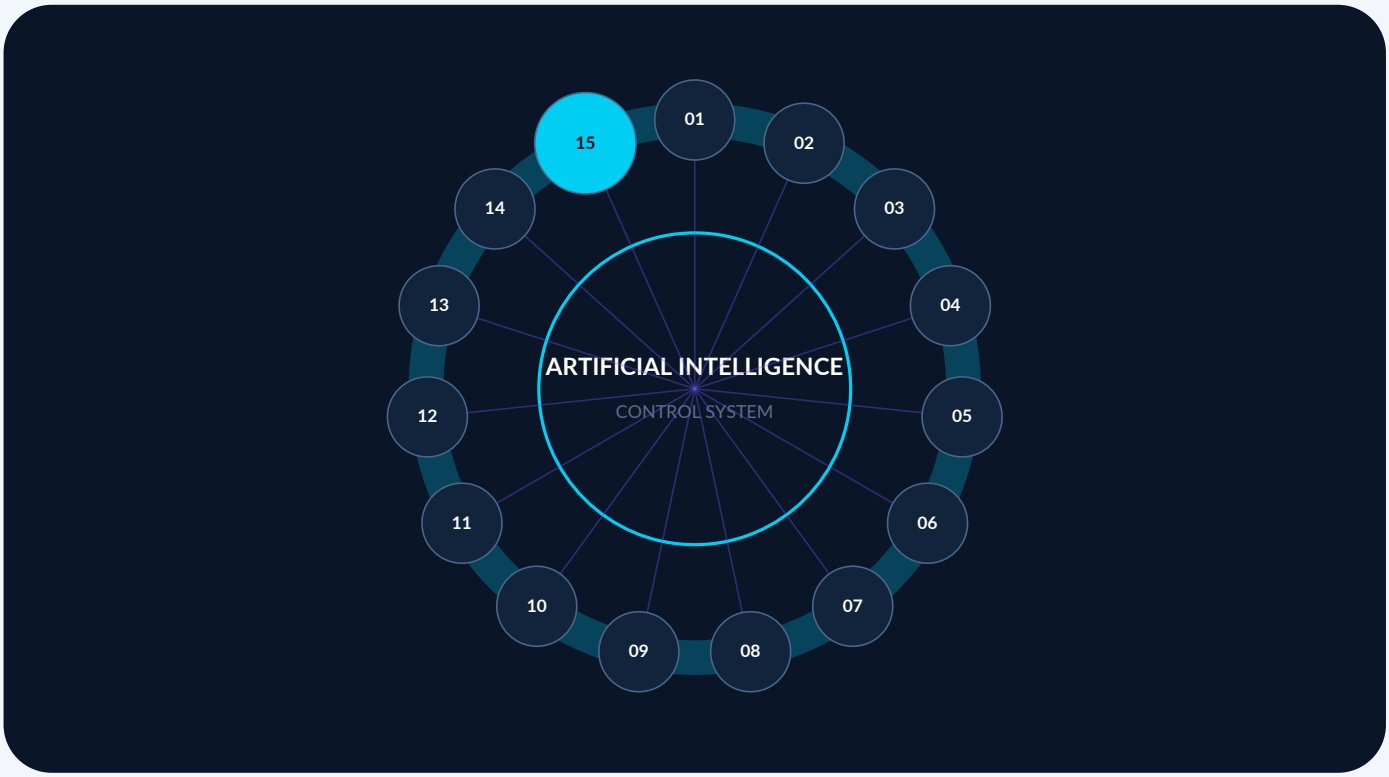
SCHEDULED REVIEW
2027-07-24

12	6	4	1
ATOMIC CRITERIA	ASSURANCE VIEWS	IMPLEMENTATION PROFILES	PERMANENT IDENTITY

Publication doctrine. This standard is a permanent library identity. Interpretation must preserve its recorded evidence, controlled exceptions, formal revision history, and explicit relationship to companion standards.

DOMAIN CONTROL SYSTEM

MYTHOS-AI-0015 inside the artificial intelligence and agentic systems constellation

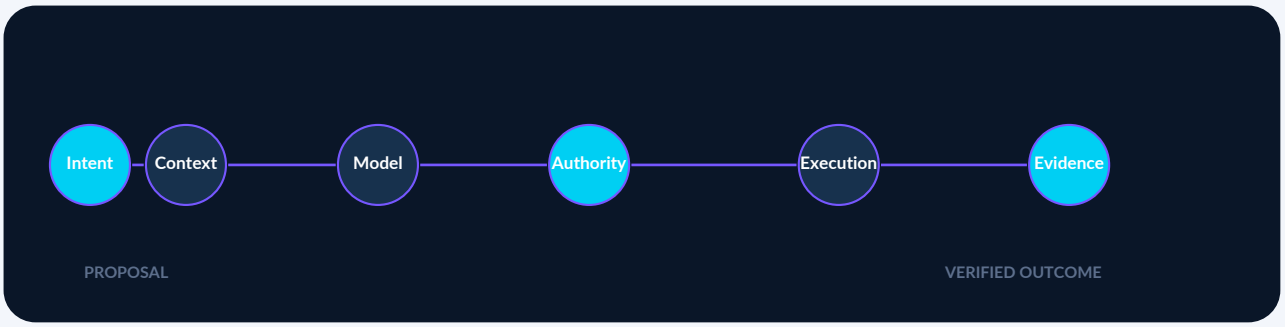


Connected assurance	Architecture, implementation, operations, evidence, and lifecycle are evaluated as one controlled system.
Specialization rule	Companion standards may add precision, but cannot silently weaken this publication.

Executive orientation

Defines privacy-preserving and federated machine-learning requirements for distributed training, aggregation, privacy budgets, poisoning resistance, participation governance, and evidence.

WHY THIS STANDARD EXISTS	The architecture of collaborative AI training has failed.
OPERATIONAL SCOPE	This standard covers the evaluation of: - Federated Learning (FL) topologies (Cross-silo, Cross-device) - Secure Multi-Party Computation (SMPC) and secure aggregation protocols - Differential Privacy (DP-SGD) and privacy budget (epsilon/delta) tracking - Byzantine-robust aggregation (Krum, Trimmed Mean, robust averaging) - Gradient inversion / model inversion attack defenses - Cross-silo model versioning and reconciliation
PRIMARY AUDIENCE	- AI researchers and ML Ops engineers building multi-party training pipelines - Security architects evaluating data leakage in collaborative ML - Compliance and privacy officers overseeing GDPR/HIPAA AI initiatives - Edge and mobile engineers building on-device model training - Standards bodies seeking a reference privacy-preserving ML methodology



Assurance principle. Model capability, data quality, identity, authority, operational evidence, and human accountability are treated as a connected system. A high aggregate score cannot compensate for a failed mandatory safety or authority boundary.

Contents

Evaluation criterion index

This standard contains 12 atomic criteria. Each criterion is independently testable and requires retained evidence.

ID	EVALUATION CRITERION
4.1.1	Raw Data Containment
4.2.1	Gradient Privacy
4.3.1	Memorization Prevention
4.4.1	Malicious Update Detection
4.5.1	Reconstruction Resistance
4.6.1	Network Reliability
5.1.1	MFLS-Privacy
5.1.2	MFLS-Byzantine
6.1.1	Gradient Inversion (Reconstruction)
6.1.2	Model Poisoning (Byzantine Attack)
6.1.3	Privacy Budget (Epsilon) Exhaustion
6.1.4	Free-Riding

MYTHOS-AI-0015-4.1.1

4.1.1 Raw Data Containment

Normative source requirement

Does the architecture strictly prohibit the centralization of raw training data?

Score	Criteria
0	All data pooled in a central data lake for training
1	Data is pseudonymized centrally, but raw PII is still transmitted
2	On-device preprocessing, but raw features still sent to the cloud
3	True Federated Learning: only model weights/gradients leave the local boundary
4	Local data sovereignty enforced cryptographically; server cannot query raw data schemas
5	Perfect sovereignty: formally verified zero raw data egress, cryptographic attestation of local data boundaries

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-4.2.1

4.2.1 Gradient Privacy

Normative source requirement

How are model updates transmitted and aggregated by the central server?

Score	Criteria
0	Plaintext gradients sent to the server
1	TLS encryption in transit, but server decrypts and sees gradients
2	Basic homomorphic encryption, but prohibitively slow for production
3	Secure Multi-Party Computation (SMPC) secure aggregation: server only sees the masked aggregate sum
4	Peer-to-peer aggregation with zero central server trust (decentralized FL)
5	Perfect privacy: formally verified SMPC bounds, zero server knowledge of individual client updates, sub-linear communication overhead

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Policy configuration, identity or trust records, authorization decisions, revocation evidence, and adversarial test results.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-4.3.1

4.3.1 Memorization Prevention

Normative source requirement

Is the model mathematically prevented from memorizing individual training records?

Score	Criteria
0	No noise added; standard SGD training
1	Basic noise injection, but no formal privacy tracking
2	DP-SGD implemented, but epsilon (privacy budget) is unbounded or too high
3	Formally bounded DP-SGD with tracked epsilon/delta across rounds
4	Adaptive noise scaling based on real-time gradient sensitivity
5	Perfect prevention: formally verified zero memorization, mathematically proven privacy bounds, optimal utility-to-privacy ratio

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Policy configuration, identity or trust records, authorization decisions, revocation evidence, and adversarial test results.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-4.4.1

4.4.1 Malicious Update Detection

Normative source requirement

Can the aggregation server survive a compromised client submitting poisoned gradients?

Score	Criteria
0	Naive averaging (FedAvg); single malicious client destroys the global model
1	Basic gradient clipping, but vulnerable to coordinated attacks
2	Simple outlier detection based on gradient norm magnitude
3	Byzantine-robust aggregation (e.g., Krum, Trimmed Mean, robust averaging) filtering malicious updates
4	Coordinate-wise median or geometric median aggregation resistant to up to 49% malicious clients
5	Perfect resilience: formally verified Byzantine fault tolerance, zero ability to inject backdoors, cryptographic proof of update validity

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Context manifests, retrieval traces, source citations, freshness records, conflict decisions, and memory provenance.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-4.5.1

4.5.1 Reconstruction Resistance

Normative source requirement

Are clients protected from server-side gradient reconstruction attacks?

Score	Criteria
0	No defenses; server can easily reconstruct raw images/text from gradients
1	Basic gradient compression applied
2	Client-side differential privacy noise applied to gradients
3	Client-side gradient perturbation and secure aggregation combined
4	Adversarial training: clients train local models to be resistant to inversion attacks
5	Perfect defense: formally verified reconstruction impossibility, mathematical proof that gradients reveal zero information about local data

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Policy configuration, identity or trust records, authorization decisions, revocation evidence, and adversarial test results.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-4.6.1

4.6.1 Network Reliability

Normative source requirement

How does the system handle the realities of a distributed, unreliable network?

Score	Criteria
0	System halts if a single client drops mid-round
1	Fixed timeout; dropped clients cause partial aggregation failures
2	Asynchronous FL: server aggregates updates as they arrive, but risks stale gradients
3	Semi-synchronous FL with buffer management; handles high client churn gracefully
4	Automated version reconciliation: clients re-sync global weights seamlessly after network partitions
5	Perfect reliability: formally verified convergence bounds despite node churn, zero stale-gradient degradation, deterministic global state convergence

---

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records.
ASSESSMENT METHOD	MATURITY SIGNAL
Run a version-pinned workload with representative inputs, record distributions rather than a single score, repeat under failure and load, and compare reproduced results with the stated acceptance threshold.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-5.1.1

5.1.1 MFLS-Privacy

Normative source requirement

• Gradient Inversion Attack: 100+ tests executing state-of-the-art reconstruction algorithms on captured gradients, verifying the defenses prevent raw data extraction.

• Membership Inference Attack: 50+ tests attempting to determine if a specific record was in a client's training set, verifying DP-SGD bounds hold.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Policy configuration, identity or trust records, authorization decisions, revocation evidence, and adversarial test results. Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Trace the decision path from identity and policy through execution, attempt bypass and privilege-escalation scenarios, verify revocation behavior, and inspect the resulting evidence record.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-5.1.2

5.1.2 MFLS-Byzantine

Normative source requirement

• Poisoning Injection: 100+ tests simulating compromised clients sending backdoored gradients, verifying the robust aggregation filters the attack without degrading global accuracy.

• Colluding Attack: 50+ tests simulating 30% of clients colluding to bias the model, verifying the system detects and quarantines the cohort.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Reproducible benchmark results, workload definitions, raw outputs, and variance records. Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Trace the decision path from identity and policy through execution, attempt bypass and privilege-escalation scenarios, verify revocation behavior, and inspect the resulting evidence record.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-6.1.1

6.1.1 Gradient Inversion (Reconstruction)

Normative source requirement

Severity: Critical Description: A malicious central server (or network sniffer) intercepts a client's plaintext gradient update. By optimizing for synthetic data that produces the same gradient, the attacker reconstructs the client's raw training images or text.

Required Controls: 1. Secure Aggregation (SMPC) ensuring the server only ever sees the cryptographic sum of all clients. 2. Client-side DP-SGD noise injection.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Inspect the architecture and configured control, execute normal and adverse scenarios, verify measurable outcomes, sample retained evidence, and confirm that exceptions are time-bound and approved.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-6.1.2

6.1.2 Model Poisoning (Byzantine Attack)

Normative source requirement

Severity: Critical Description: A compromised client deliberately sends large, malformed gradients to the server. When averaged into the global model, this corrupts the weights, rendering the model useless or embedding a hidden backdoor (e.g., "ignore safety rule X").

Required Controls: 1. Byzantine-robust aggregation algorithms (e.g., Trimmed Mean). 2. Server-side anomaly detection on gradient norms.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Context manifests, retrieval traces, source citations, freshness records, conflict decisions, and memory provenance. Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Trace the decision path from identity and policy through execution, attempt bypass and privilege-escalation scenarios, verify revocation behavior, and inspect the resulting evidence record.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-6.1.3

6.1.3 Privacy Budget (Epsilon) Exhaustion

Normative source requirement

Severity: High Description: Over thousands of federated rounds, the cumulative epsilon (privacy loss) exceeds safe bounds. An attacker can use membership inference attacks to prove a specific individual's data was used in training.

Required Controls: 1. Strict, centralized tracking of epsilon across all training rounds. 2. Automatic halt of training when the privacy budget is exhausted.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Policy configuration, identity or trust records, authorization decisions, revocation evidence, and adversarial test results. Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Trace the decision path from identity and policy through execution, attempt bypass and privilege-escalation scenarios, verify revocation behavior, and inspect the resulting evidence record.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

MYTHOS-AI-0015-6.1.4

6.1.4 Free-Riding

Normative source requirement

Severity: Medium Description: A lazy or malicious client contributes zero compute, sending empty or copied gradients to receive the updated global model without doing any work.

Required Controls: 1. Proof-of-Work or cryptographic attestation requiring clients to prove they executed the local training step. 2. Verification of gradient variance.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Model and dataset cards, lineage, licenses, evaluation reports, release approvals, and rollback artifacts.
ASSESSMENT METHOD	MATURITY SIGNAL
Inspect the architecture and configured control, execute normal and adverse scenarios, verify measurable outcomes, sample retained evidence, and confirm that exceptions are time-bound and approved.	0 absent | 1 ad hoc | 2 repeatable | 3 controlled | 4 measured | 5 continuously verified

Decision rule. A criterion is not satisfied by documentation alone when the claim concerns runtime behavior. Reproduced evidence, responsible ownership, known limitations, and controlled exceptions are required.

Current authority and freshness register

These sources inform interpretation but do not convert this candidate publication into certification, legal compliance, or implementation evidence. Rolling sources must be version-pinned at adoption time.

AUTHORITY	VERSION / FRESHNESS	APPLICABILITY LIMIT
NIST - Artificial Intelligence Risk Management Framework (AI RMF 1.0)	NIST AI 100-1; current framework, revision underway Expires 2026-10-24	Voluntary risk-management framework; does not establish product certification or implementation effectiveness.
NIST - Generative Artificial Intelligence Profile	NIST AI 600-1, July 2024 Expires 2027-01-24	Profile guidance requires tailoring to the organization, use case, and risk tolerance.
NIST - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations	NIST AI 100-2e2025 Expires 2027-01-24	Taxonomy and terminology do not substitute for system-specific red-team evidence.
OWASP - Top 10 for LLM Applications 2025	2025 edition Expires 2026-10-24	Community risk guidance; prioritization and mitigations require application-specific validation.
OWASP - Top 10 for Agentic Applications 2026	2026 edition Expires 2026-10-24	Emerging community guidance for agentic systems; not a certification framework.
OpenTelemetry - Semantic Conventions	1.43.0 rolling specification; GenAI conventions maintained separately Expires 2026-08-24	Several GenAI and agent conventions remain under active development and require version pinning.
C2PA - Content Credentials Technical Specification	2.4 rolling publication Expires 2026-10-24	Provenance establishes signed assertions and tamper evidence, not the truth or quality of the underlying content.

Source lineage appendix

The following text preserves the substantive source draft in a normalized public form. Where it conflicts with the permanent publication identity, candidate status, current authority register, or evidence requirements above, the controlled publication layer governs.

0. Executive Mandate

The architecture of collaborative AI training has failed.

Organizations attempt to build shared intelligence by pooling raw, sensitive data into a monolithic central data lake. This creates a honeypot of immense value, violating data sovereignty regulations (GDPR, HIPAA, CMMC), exposing intellectual property, and guaranteeing that a single breach compromises all participating entities. When organizations refuse to share raw data, they resort to training isolated, weaker models.

This standard exists because:

1. Data Gravity Requires Distributed Training: The most valuable data (medical records, proprietary code, battlefield telemetry) is heavily restricted and cannot be centralized. Production multi-party AI MUST utilize Federated Learning (FL), bringing the compute to the data rather than the data to the compute.

2. Raw Gradients are Leaked Data: Sending model weight updates back to a central server is not inherently private. An attacker (or a curious server admin) can execute a gradient inversion (model inversion) attack to reconstruct the raw training data from the gradients. Systems MUST utilize Secure Multi-Party Computation (SMPC) or Homomorphic Encryption (HE) to aggregate gradients without the server ever seeing them.

3. Differential Privacy Must Be Provable: Adding noise to gradients is essential to prevent memorization, but too much noise destroys model accuracy. Systems MUST implement formally bounded Differential Privacy (DP-SGD), tracking the privacy budget (epsilon) to mathematically guarantee that no individual data point can be extracted.

4. Federated Networks are Byzantine: A single compromised client in a federated network can submit poisoned gradients, backdooring the global model. Federated aggregation MUST be Byzantine-robust, capable of detecting and rejecting malicious updates without halting convergence.

This document establishes the Mythos Federated Learning Standard (MFLS), a comprehensive, evidence-based methodology for evaluating multi-party ML pipelines against cryptographic privacy, Byzantine resilience, and data sovereignty.

Part I. Scope and Applicability

1.1 In Scope

This standard covers the evaluation of:

- Federated Learning (FL) topologies (Cross-silo, Cross-device)
- Secure Multi-Party Computation (SMPC) and secure aggregation protocols
- Differential Privacy (DP-SGD) and privacy budget (epsilon/delta) tracking
- Byzantine-robust aggregation (Krum, Trimmed Mean, robust averaging)
- Gradient inversion / model inversion attack defenses
- Cross-silo model versioning and reconciliation

1.2 Out of Scope

- General continual learning / anti-forgetting (covered in MYTHOS-AI-0012)
- Fully Homomorphic Encryption (FHE) for inference (covered in MYTHOS-SEC-0010)
- General data egress controls (covered in MYTHOS-DAT-0003)

1.3 Audience

- AI researchers and ML Ops engineers building multi-party training pipelines
- Security architects evaluating data leakage in collaborative ML
- Compliance and privacy officers overseeing GDPR/HIPAA AI initiatives
- Edge and mobile engineers building on-device model training
- Standards bodies seeking a reference privacy-preserving ML methodology

Part II. Definitions and Taxonomy

2.1 Federated & Privacy Dimensions

Dimension	Definition
Federated Learning (FL)	A machine learning paradigm where the model is trained across multiple decentralized edge devices or servers holding local data samples, without ever exchanging the raw data.
Secure Aggregation (SMPC)	A cryptographic protocol that allows a central server to compute the sum of multiple clients' encrypted gradients without ever seeing any individual client's plaintext gradient.
Differential Privacy (DP)	A mathematical framework that provides guarantees about the privacy of individuals in a dataset by adding calibrated noise to the training process (e.g., DP-SGD), bounding the impact of any single record.
Gradient Inversion	An attack where an adversary reconstructs the raw training data from the model gradients sent by a client, by optimizing for input images that produce the same gradients.
Byzantine Client	A compromised or malicious participant in a federated network that submits incorrect or poisoned model updates to degrade or backdoor the global model.

2.2 The Federated Learning Doctrine

> A central data lake is a breach waiting to happen. A shared gradient is a leaked prompt. A federated network without Byzantine defense is a backdoor. If the privacy budget is not bounded, the data is not private.

Part III. Evaluation Methodology

3.1 Scoring Model

Each capability is scored from 0 to 5.

Score	Meaning
0	Fails basic task; centralized raw data pooling, no privacy controls
1	Basic data anonymization (easily reversed), no federated topology
2	Functional FL, but raw gradients exposed to the central server
3	Solid production performance; SMPC secure aggregation, DP-SGD, basic Byzantine defense
4	Strong performance; formally bounded epsilon, gradient inversion defenses, robust outlier detection
5	Best-in-class; sets the standard for mathematically proven, zero-knowledge multi-party ML

Weighting

Category	Weight	Rationale
Data Sovereignty & Egress Elimination	20%	The core premise of FL is that raw data never moves
Secure Aggregation (SMPC/HE)	20%	Plaintext gradients can be inverted to reveal raw data
Differential Privacy (DP-SGD)	15%	Without bounded noise, models memorize and leak training data
Byzantine Resilience & Poisoning Defense	15%	One malicious client can backdoor the global model
Gradient Inversion Defenses	15%	Even encrypted pipelines must defend against reconstruction attacks
Operational Lifecycle & Reconciliation	15%	Federated networks must handle dropped clients and version drift

Total: 100%

A system MUST score at least 3.0 in Data Sovereignty and Secure Aggregation to be considered for production use.

Part IV. Evaluation Categories

4.1 Data Sovereignty and Egress Elimination (Weight: 20%)

4.1.1 Raw Data Containment

Does the architecture strictly prohibit the centralization of raw training data?

Score	Criteria
0	All data pooled in a central data lake for training
1	Data is pseudonymized centrally, but raw PII is still transmitted
2	On-device preprocessing, but raw features still sent to the cloud
3	True Federated Learning: only model weights/gradients leave the local boundary
4	Local data sovereignty enforced cryptographically; server cannot query raw data schemas
5	Perfect sovereignty: formally verified zero raw data egress, cryptographic attestation of local data boundaries

4.2 Secure Aggregation (SMPC/HE) (Weight: 20%)

4.2.1 Gradient Privacy

How are model updates transmitted and aggregated by the central server?

Score	Criteria
0	Plaintext gradients sent to the server
1	TLS encryption in transit, but server decrypts and sees gradients
2	Basic homomorphic encryption, but prohibitively slow for production
3	Secure Multi-Party Computation (SMPC) secure aggregation: server only sees the masked aggregate sum
4	Peer-to-peer aggregation with zero central server trust (decentralized FL)
5	Perfect privacy: formally verified SMPC bounds, zero server knowledge of individual client updates, sub-linear communication overhead

4.3 Differential Privacy (DP-SGD) (Weight: 15%)

4.3.1 Memorization Prevention

Is the model mathematically prevented from memorizing individual training records?

Score	Criteria
0	No noise added; standard SGD training
1	Basic noise injection, but no formal privacy tracking
2	DP-SGD implemented, but epsilon (privacy budget) is unbounded or too high
3	Formally bounded DP-SGD with tracked epsilon/delta across rounds
4	Adaptive noise scaling based on real-time gradient sensitivity
5	Perfect prevention: formally verified zero memorization, mathematically proven privacy bounds, optimal utility-to-privacy ratio

4.4 Byzantine Resilience and Poisoning Defense (Weight: 15%)

4.4.1 Malicious Update Detection

Can the aggregation server survive a compromised client submitting poisoned gradients?

Score	Criteria
0	Naive averaging (FedAvg); single malicious client destroys the global model
1	Basic gradient clipping, but vulnerable to coordinated attacks
2	Simple outlier detection based on gradient norm magnitude
3	Byzantine-robust aggregation (e.g., Krum, Trimmed Mean, robust averaging) filtering malicious updates
4	Coordinate-wise median or geometric median aggregation resistant to up to 49% malicious clients
5	Perfect resilience: formally verified Byzantine fault tolerance, zero ability to inject backdoors, cryptographic proof of update validity

4.5 Gradient Inversion Defenses (Weight: 15%)

4.5.1 Reconstruction Resistance

Are clients protected from server-side gradient reconstruction attacks?

Score	Criteria
0	No defenses; server can easily reconstruct raw images/text from gradients
1	Basic gradient compression applied
2	Client-side differential privacy noise applied to gradients
3	Client-side gradient perturbation and secure aggregation combined
4	Adversarial training: clients train local models to be resistant to inversion attacks
5	Perfect defense: formally verified reconstruction impossibility, mathematical proof that gradients reveal zero information about local data

4.6 Operational Lifecycle and Reconciliation (Weight: 15%)

4.6.1 Network Reliability

How does the system handle the realities of a distributed, unreliable network?

Score	Criteria
0	System halts if a single client drops mid-round
1	Fixed timeout; dropped clients cause partial aggregation failures
2	Asynchronous FL: server aggregates updates as they arrive, but risks stale gradients
3	Semi-synchronous FL with buffer management; handles high client churn gracefully
4	Automated version reconciliation: clients re-sync global weights seamlessly after network partitions
5	Perfect reliability: formally verified convergence bounds despite node churn, zero stale-gradient degradation, deterministic global state convergence

Part V. The Mythos Federated Learning Suite (MFLS)

Traditional ML benchmarks measure model accuracy. The MFLS evaluates privacy leakage, Byzantine resilience, and data sovereignty.

5.1 Suite Composition

5.1.1 MFLS-Privacy

• Gradient Inversion Attack: 100+ tests executing state-of-the-art reconstruction algorithms on captured gradients, verifying the defenses prevent raw data extraction.

• Membership Inference Attack: 50+ tests attempting to determine if a specific record was in a client's training set, verifying DP-SGD bounds hold.

5.1.2 MFLS-Byzantine

• Poisoning Injection: 100+ tests simulating compromised clients sending backdoored gradients, verifying the robust aggregation filters the attack without degrading global accuracy.

• Colluding Attack: 50+ tests simulating 30% of clients colluding to bias the model, verifying the system detects and quarantines the cohort.

5.2 Execution Protocol

```
1. Deploy federated learning architecture with SMPC and DP-SGD
2. Execute a multi-round training protocol across distributed nodes
3. Run MFLS suite (automated inversion attacks + Byzantine injection)
4. Score each category 0-5
5. Check for raw data egress or unencrypted gradient transmission (instant disqualification)
6. If all thresholds met: approve for production
```

Part VI. Security Threat Model for Federated Learning

6.1 Threat Catalog

6.1.1 Gradient Inversion (Reconstruction)

Severity: Critical Description: A malicious central server (or network sniffer) intercepts a client's plaintext gradient update. By optimizing for synthetic data that produces the same gradient, the attacker reconstructs the client's raw training images or text.

Required Controls:

1. Secure Aggregation (SMPC) ensuring the server only ever sees the cryptographic sum of all clients.

2. Client-side DP-SGD noise injection.

6.1.2 Model Poisoning (Byzantine Attack)

Severity: Critical Description: A compromised client deliberately sends large, malformed gradients to the server. When averaged into the global model, this corrupts the weights, rendering the model useless or embedding a hidden backdoor (e.g., "ignore safety rule X").

Required Controls:

1. Byzantine-robust aggregation algorithms (e.g., Trimmed Mean).

2. Server-side anomaly detection on gradient norms.

6.1.3 Privacy Budget (Epsilon) Exhaustion

Severity: High Description: Over thousands of federated rounds, the cumulative epsilon (privacy loss) exceeds safe bounds. An attacker can use membership inference attacks to prove a specific individual's data was used in training.

Required Controls:

1. Strict, centralized tracking of epsilon across all training rounds.

2. Automatic halt of training when the privacy budget is exhausted.

6.1.4 Free-Riding

Severity: Medium Description: A lazy or malicious client contributes zero compute, sending empty or copied gradients to receive the updated global model without doing any work.

Required Controls:

1. Proof-of-Work or cryptographic attestation requiring clients to prove they executed the local training step.

2. Verification of gradient variance.

Part VII. The Mythos Federated Learning Standard

7.1 The Mythos Federated Doctrine

```
A central data lake is a breach waiting to happen.
```

```
A shared gradient is a leaked prompt.

A federated network without Byzantine defense is a backdoor.
If the privacy budget is not bounded, the data is not private.

Models may be shared.
Data must be absolute.
```

7.2 Selection Rules

1. No multi-party ML architecture enters production without passing MFLS.
2. No architecture is approved if it centralizes raw data for model training.
3. No architecture is approved if the server receives plaintext client gradients.
4. No architecture is approved without Byzantine-robust aggregation.
5. No architecture is approved without formally bounded Differential Privacy (DP-SGD).

7.3 The Prime Directive for Federated Learning

> Keep the data, share the weights. > Mask the gradient, do not trust the server. > Bound the epsilon, do not leak the record. > Filter the Byzantine, do not average the poison.

```
Academic benchmarks measure model accuracy.
Applied benchmarks measure secure aggregation overhead.
Security benchmarks measure gradient inversion resistance.
Operational benchmarks measure Byzantine resilience.

> Intelligence may be collaborative.
> Privacy must be absolute.
```

End of Standard MYTHOS-AI-0015

End of controlled publication

MYTHOS-AI-0015 remains a Candidate Standard until technical review, security and privacy review, accessibility review, current-source validation, and formal approval are recorded.