ENGINEERING STANDARDS LIBRARY / ARTIFICIAL INTELLIGENCE

Neuro-Symbolic AI & Continual Learning (Anti-Catastrophic Forgetting) Standard

The architecture of AI learning and reasoning has failed.

PERMANENT DOCUMENT ID

MYTHOS-AI-0012

PUBLICATION

Candidate Standard
Version 1.0.0-candidate

ASSURANCE PROFILE

Standard / Advanced
High Assurance

PERMANENT PUBLICATION IDENTITY

Neuro-Symbolic AI & Continual Learning (Anti-Catastrophic Forgetting) Standard

DOCUMENT ID
MYTHOS-AI-0012

VERSION
1.0.0-candidate

STATUS
Candidate Standard

PUBLISHER
Mythos Systems

PUBLICATION DATE
2026-07-24

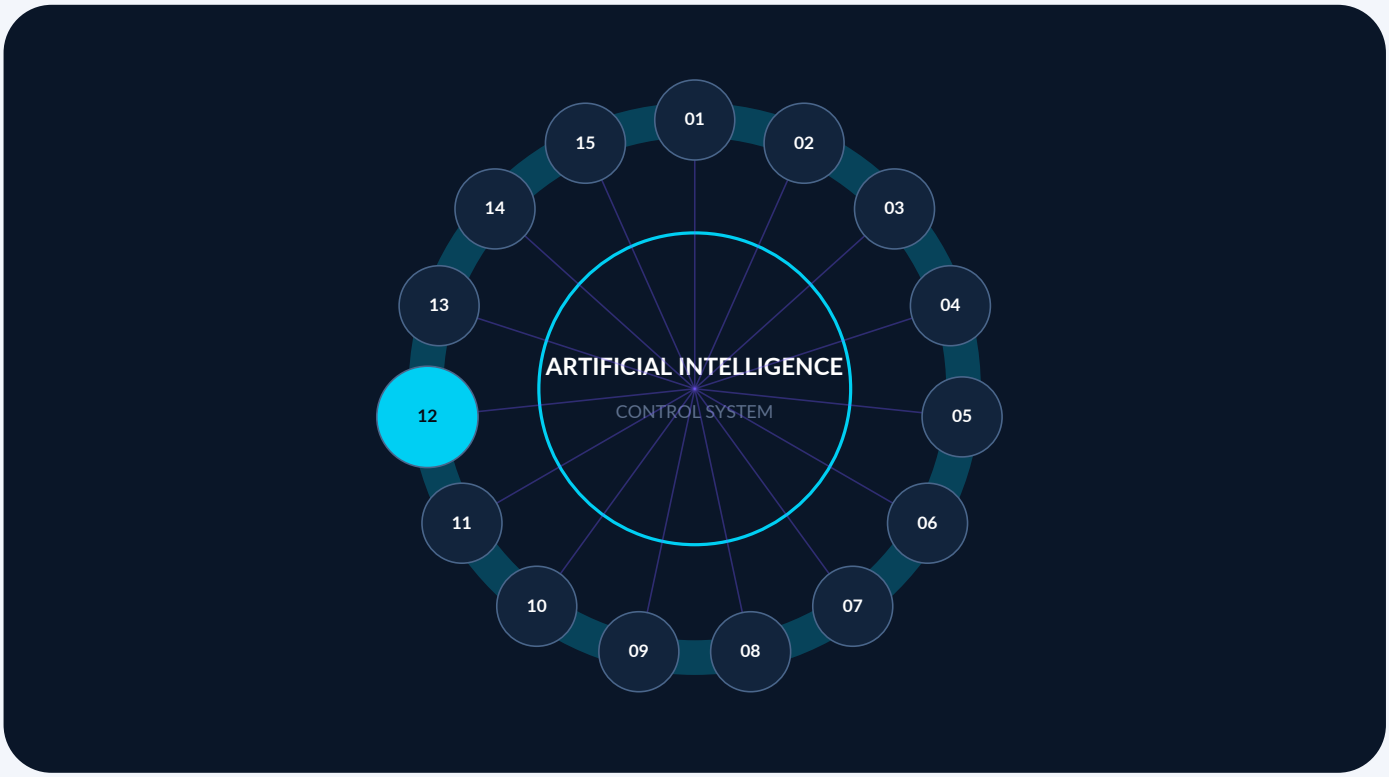
SCHEDULED REVIEW
2027-07-24

11	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-0012 inside the artificial intelligence and agentic systems constellation

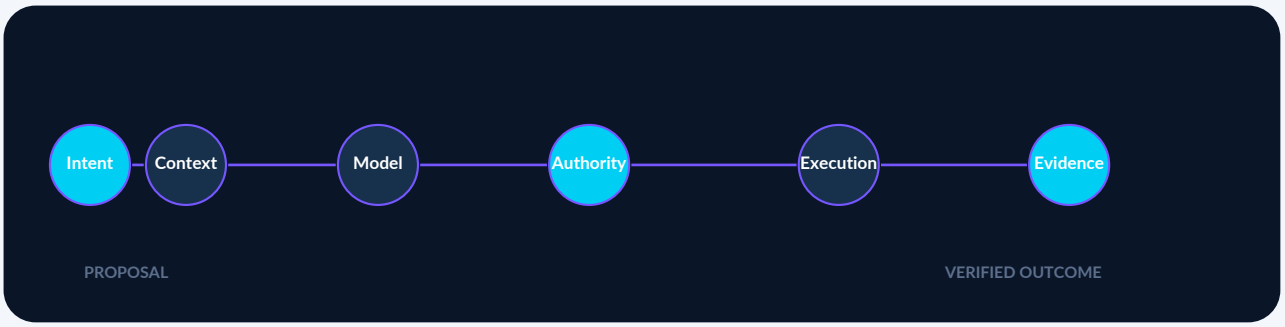


Connected assurance	Architecture, implementation, operations, evidence, and lifecycle are evaluated as one controlled system.
Specialization rule	Companion standards may add precision, but cannot silently weaken this publication.

Executive orientation

Establishes requirements for neuro-symbolic integration and continual learning while controlling catastrophic forgetting, unsafe adaptation, provenance loss, and unbounded behavioral change.

WHY THIS STANDARD EXISTS	The architecture of AI learning and reasoning has failed.
OPERATIONAL SCOPE	This standard covers the evaluation of: - Neuro-Symbolic AI architectures (Neural perception + Symbolic reasoning) - Continual learning and online learning pipelines - Catastrophic forgetting mitigations (EWC, modular networks, memory replay) - Parameter-Efficient Fine-Tuning (PEFT) and adapter isolation (LoRA, QLoRA) - Dynamic Knowledge Graph (KG) synchronization and fact injection - Experience replay and episodic memory for agents
PRIMARY AUDIENCE	- AI researchers and principal engineers designing long-running autonomous agents - Platform engineers building fine-tuning and adapter infrastructure - Knowledge engineers and ontologists managing enterprise graphs - Security teams evaluating AI drift and policy degradation over time - Standards bodies seeking a reference continual learning 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 11 atomic criteria. Each criterion is independently testable and requires retained evidence.

ID	EVALUATION CRITERION
4.1.1	Perception vs. Reasoning Separation
4.2.1	Weight Preservation
4.3.1	Update Granularity
4.4.1	Symbolic Updates
4.5.1	Online Learning Stability
4.6.1	Logic Verification
5.1.1	MCLS-Forgetting
5.1.2	MCLS-Symbolic
6.1.1	Data Poisoning via Feedback Loop
6.1.2	Catastrophic Safety Forgetting
6.1.3	Symbolic Rule Conflict

MYTHOS-AI-0012-4.1.1

4.1.1 Perception vs. Reasoning Separation

Normative source requirement

Are probabilistic perception and deterministic reasoning architecturally separated?

Score	Criteria
0	Monolithic LLM attempts both perception and logical deduction (guaranteed hallucinations)
1	LLM uses basic tool-calling to query a database, but reasoning is still neural
2	LLM generates structured queries (SPARQL/SQL), but engine results are not fed back to constrain the LLM
3	Neuro-Symbolic: LLM parses intent -> Symbolic engine executes logic -> LLM formats deterministic output
4	LLM is strictly forbidden from declaring facts; all facts must be resolved by the symbolic engine
5	Perfect separation: formally verified symbolic grounding, zero unconstrained neural reasoning allowed

---

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-0012-4.2.1

4.2.1 Weight Preservation

Normative source requirement

How does the system prevent the learning of new information from degrading old capabilities?

Score	Criteria
0	Full parameter fine-tuning; old tasks degrade instantly when new tasks are learned
1	Periodic mixing of old datasets (experience replay) during training, but offline only
2	Basic parameter freezing (e.g., freezing lower transformer layers)
3	Elastic Weight Consolidation (EWC) or Synaptic Intelligence mathematically bounding weight drift
4	Dynamic, per-task sub-networks activated via routing (Modular Networks)
5	Perfect prevention: formally verified zero-interference learning, mathematically proven retention of all prior capabilities

---

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-0012-4.3.1

4.3.1 Update Granularity

Normative source requirement

How are new domains, skills, or facts learned without global retraining?

Score	Criteria
0	Global weight updates for every new skill
1	Manual creation of separate model instances per task
2	Basic Parameter-Efficient Fine-Tuning (PEFT) like LoRA, but manually loaded/unloaded
3	Dynamic adapter routing: system loads/unloads isolated LoRA adapters per task in real-time
4	Automated adapter generation and lifecycle management (versioning, A/B testing, rollback)
5	Perfect isolation: formally verified adapter boundaries, zero cross-contamination of adapter weights, automated capability profiling

---

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-0012-4.4.1

4.4.1 Symbolic Updates

Normative source requirement

Can the deterministic reasoning engine learn new facts in real-time without retraining?

Score	Criteria
0	Symbolic rules are hardcoded and static
1	Rules require manual restart to update
2	Periodic batch updates to the knowledge graph
3	Real-time, transactional updates to the Knowledge Graph (ACID compliant)
4	Contradiction resolution: new facts automatically flag and quarantine conflicting old facts
5	Perfect synchronization: formally verified logical consistency, automated fact provenance tracking, zero rule conflicts

---

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-0012-4.5.1

4.5.1 Online Learning Stability

Normative source requirement

Does the agent leverage past experiences to guide future learning without storing raw gradients?

Score	Criteria
0	Agent learns in isolation; no memory of past successes/failures
1	Agent stores raw text logs, but cannot use them for learning
2	Basic episodic memory buffer for context injection (RAG)
3	System extracts structured lessons (e.g., "Tool X failed on input Y") and stores them in the symbolic graph
4	Automated experience replay: system periodically re-evaluates past failures using newly acquired adapters/skills to verify resolution
5	Perfect stability: formally verified replay bounds, zero repetitive failure loops, mathematical proof of learning 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. 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-0012-4.6.1

4.6.1 Logic Verification

Normative source requirement

How is the agent's evolving logic verified to ensure it hasn't drifted into unsafe behavior?

Score	Criteria
0	No verification; agent behavior changes unpredictably over time
1	Manual human review of agent outputs periodically
2	Standard evaluation suite run after major retraining
3	Automated safety bounds checking against the symbolic logic engine after every update
4	Continuous formal verification: system mathematically proves the agent still adheres to safety constraints before deploying new adapters
5	Perfect verification: formally verified invariant preservation, zero ability to deploy logically conflicting adapters, real-time drift detection

---

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-0012-5.1.1

5.1.1 MCLS-Forgetting

Normative source requirement

• Sequential Task Injection: 100+ tests teaching the agent a new skill, followed immediately by an evaluation on an older skill, verifying accuracy does not degrade.

• Safety Constraint Retention: 50+ tests attempting to teach the agent a new capability that subtly conflicts with a core safety rule, verifying the system rejects or isolates the update.

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. Skill graph, assessment blueprint, learner evidence, lab isolation record, accessibility results, and credential metadata.
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-0012-5.1.2

5.1.2 MCLS-Symbolic

Normative source requirement

• Logic Deduction Test: 100+ tests requiring multi-step logical deduction (e.g., "If A is true and B implies C, is C true?"), verifying the symbolic engine executes the logic, not the LLM.

• Fact Update Test: 50+ tests injecting a contradictory fact into the KG, verifying the system resolves the conflict without hallucinating or crashing.

CONTROL INTENT	REQUIRED EVIDENCE
Create a measurable, reviewable control that can be verified independently of model or vendor claims.	Architecture records, implemented configuration, test evidence, operational telemetry, exception decisions, and accountable approval.
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-0012-6.1.1

6.1.1 Data Poisoning via Feedback Loop

Normative source requirement

Severity: Critical Description: An attacker (or a prompt injection) feeds the agent malicious "lessons" or feedback. The continual learning pipeline ingests this and updates the symbolic graph or adapter, permanently corrupting the agent's behavior.

Required Controls: 1. Quarantine and verification pipeline for all learned facts before KG commitment. 2. Cryptographic signing of "trusted" training sources. 3. Rollback capabilities for both adapters and KG states.

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-0012-6.1.2

6.1.2 Catastrophic Safety Forgetting

Normative source requirement

Severity: Critical Description: The agent learns a new, highly complex workflow, and the weight updates (even within an adapter) subtly degrade the model's adherence to a core safety policy (e.g., "never execute raw SQL without validation").

Required Controls: 1. Elastic Weight Consolidation (EWC) bounding. 2. Continuous safety-bound verification post-update.

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-0012-6.1.3

6.1.3 Symbolic Rule Conflict

Normative source requirement

Severity: High Description: Two independent data sources inject contradictory rules into the symbolic engine, causing the logic engine to deadlock or produce unpredictable, non-deterministic outputs.

Required Controls: 1. Strict ACID compliance and contradiction resolution logic in the KG. 2. Trust classes/priority scores for symbolic rules.

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.

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 AI learning and reasoning has failed.

Organizations deploy Large Language Models (LLMs) as monolithic, static statistical engines. When the model hallucinates, they attempt to fix it by fine-tuning the entire network on new data. This brute-force approach inevitably triggers catastrophic forgetting-the model learns the new fact but mathematically overwrites its understanding of a foundational concept, degrading globally. Furthermore, they force the neural network to do the job of a database. They ask a statistical probability engine to perform rigorous logical deduction, resulting in systems that cannot guarantee mathematical safety bounds.

This standard exists because:

1. Neural Networks are for Perception, Not Reasoning: A deep learning model is exceptional at parsing unstructured data (perception) but fundamentally incapable of guaranteed logical deduction. Systems MUST adopt Neuro-Symbolic architectures, where the neural network parses the intent, and a symbolic engine (Knowledge Graph, Logic Engine) executes deterministic reasoning.

2. Fine-Tuning is a Blunt Instrument: Updating the global weights of a billion-parameter model to teach it a new API or a new company policy is an architectural anti-pattern. It guarantees drift and forgetting. Systems MUST use modular, isolated learning (e.g., LoRA adapters, dynamic routing) or external memory graphs to absorb new knowledge.

3. Forgetting is a Safety Failure: An autonomous agent that forgets a critical safety constraint because it learned a new UI navigation pattern is a liability. Continual learning MUST mathematically bound the retention of critical weights using techniques like Elastic Weight Consolidation (EWC) or synaptic intelligence.

4. Static Models are Dead Intelligence: The world changes faster than foundation models can be trained. Production agents MUST possess the ability to learn continuously in production, updating their local symbolic knowledge without requiring cloud-scale gradient descent.

This document establishes the Mythos Continual Learning Standard (MCLS), a comprehensive, evidence-based methodology for evaluating AI systems against neuro-symbolic integration, catastrophic forgetting prevention, and dynamic knowledge assimilation.

Part I. Scope and Applicability

1.1 In Scope

This standard covers the evaluation of:

- Neuro-Symbolic AI architectures (Neural perception + Symbolic reasoning)
- Continual learning and online learning pipelines
- Catastrophic forgetting mitigations (EWC, modular networks, memory replay)
- Parameter-Efficient Fine-Tuning (PEFT) and adapter isolation (LoRA, QLoRA)
- Dynamic Knowledge Graph (KG) synchronization and fact injection
- Experience replay and episodic memory for agents

1.2 Out of Scope

- General agentic framework evaluation (covered in MYTHOS-AI-0001)
- General RAG and retrieval architecture (covered in MYTHOS-KNW-0001)
- General context window compaction (covered in MYTHOS-AI-0007)

1.3 Audience

- AI researchers and principal engineers designing long-running autonomous agents
- Platform engineers building fine-tuning and adapter infrastructure
- Knowledge engineers and ontologists managing enterprise graphs
- Security teams evaluating AI drift and policy degradation over time
- Standards bodies seeking a reference continual learning methodology

Part II. Definitions and Taxonomy

2.1 Learning Dimensions

Dimension	Definition
Catastrophic Forgetting	The tendency of an artificial neural network to completely and abruptly forget previously learned information upon learning new information, due to the overwriting of shared weights.
Neuro-Symbolic AI	An architectural paradigm that combines deep learning (neural networks for pattern recognition/perception) with symbolic AI (logic engines/knowledge graphs for reasoning and fact verification).
Elastic Weight Consolidation (EWC)	A continual learning technique that constrains the update of weights deemed important for previous tasks, by penalizing changes to those specific weights during new training.
Modular Isolation	The practice of learning new tasks or domains by activating or creating distinct, isolated sub-networks (e.g., LoRA adapters) rather than updating the global base model.
Symbolic Grounding	The process of binding the probabilistic outputs of a neural network to deterministic, verifiable entities and rules within a symbolic logic engine.

2.2 The Continual Learning Doctrine

> A neural network is a probability engine, not a database. Forgetting a safety rule to learn a new feature is a critical failure. Fine-tuning a billion-parameter model to update a fact is architectural malpractice. If the logic cannot be verified, the reasoning is an illusion.

Part III. Evaluation Methodology

3.1 Scoring Model

Each capability is scored from 0 to 5.

Score	Meaning
0	Fails basic task; monolithic models, full fine-tuning, global forgetting, no symbolic logic
1	Basic RAG exists, but learning requires destructive global updates
2	Functional PEFT (LoRA), but lacks formal forgetting bounds or neuro-symbolic separation
3	Solid production performance; modular adapters, EWC bounds, neuro-symbolic separation
4	Strong performance; dynamic KG sync, automated replay, formally verified logic engine
5	Best-in-class; sets the standard for mathematically proven, zero-forgetting continual learning

Weighting

Category	Weight	Rationale
Neuro-Symbolic Architecture	20%	Neural networks cannot guarantee logical safety bounds
Catastrophic Forgetting Prevention	20%	Destructive updates break production reliability over time
Modular Isolation & Adapter Mgmt	15%	Global fine-tuning is economically and operationally unscalable
Knowledge Graph Synchronization	15%	Symbolic knowledge must update dynamically without retraining
Experience Replay & Episodic Memory	15%	Online learning requires historical context to prevent drift

Category	Weight	Rationale
Operational Lifecycle & Verification	15%	Learned logic must be continuously verified for correctness

Total: 100%

A system MUST score at least 3.0 in Neuro-Symbolic Architecture and Catastrophic Forgetting Prevention to be considered for production use.

Part IV. Evaluation Categories

4.1 Neuro-Symbolic Architecture (Weight: 20%)

4.1.1 Perception vs. Reasoning Separation

Are probabilistic perception and deterministic reasoning architecturally separated?

Score	Criteria
0	Monolithic LLM attempts both perception and logical deduction (guaranteed hallucinations)
1	LLM uses basic tool-calling to query a database, but reasoning is still neural
2	LLM generates structured queries (SPARQL/SQL), but engine results are not fed back to constrain the LLM
3	Neuro-Symbolic: LLM parses intent -> Symbolic engine executes logic -> LLM formats deterministic output
4	LLM is strictly forbidden from declaring facts; all facts must be resolved by the symbolic engine
5	Perfect separation: formally verified symbolic grounding, zero unconstrained neural reasoning allowed

4.2 Catastrophic Forgetting Prevention (Weight: 20%)

4.2.1 Weight Preservation

How does the system prevent the learning of new information from degrading old capabilities?

Score	Criteria
0	Full parameter fine-tuning; old tasks degrade instantly when new tasks are learned
1	Periodic mixing of old datasets (experience replay) during training, but offline only
2	Basic parameter freezing (e.g., freezing lower transformer layers)
3	Elastic Weight Consolidation (EWC) or Synaptic Intelligence mathematically bounding weight drift
4	Dynamic, per-task sub-networks activated via routing (Modular Networks)
5	Perfect prevention: formally verified zero-interference learning, mathematically proven retention of all prior capabilities

4.3 Modular Isolation and Adapter Management (Weight: 15%)

4.3.1 Update Granularity

How are new domains, skills, or facts learned without global retraining?

Score	Criteria
0	Global weight updates for every new skill
1	Manual creation of separate model instances per task

Score	Criteria
2	Basic Parameter-Efficient Fine-Tuning (PEFT) like LoRA, but manually loaded/unloaded
3	Dynamic adapter routing: system loads/unloads isolated LoRA adapters per task in real-time
4	Automated adapter generation and lifecycle management (versioning, A/B testing, rollback)
5	Perfect isolation: formally verified adapter boundaries, zero cross-contamination of adapter weights, automated capability profiling

4.4 Knowledge Graph Synchronization (Weight: 15%)

4.4.1 Symbolic Updates

Can the deterministic reasoning engine learn new facts in real-time without retraining?

Score	Criteria
0	Symbolic rules are hardcoded and static
1	Rules require manual restart to update
2	Periodic batch updates to the knowledge graph
3	Real-time, transactional updates to the Knowledge Graph (ACID compliant)
4	Contradiction resolution: new facts automatically flag and quarantine conflicting old facts
5	Perfect synchronization: formally verified logical consistency, automated fact provenance tracking, zero rule conflicts

4.5 Experience Replay and Episodic Memory (Weight: 15%)

4.5.1 Online Learning Stability

Does the agent leverage past experiences to guide future learning without storing raw gradients?

Score	Criteria
0	Agent learns in isolation; no memory of past successes/failures
1	Agent stores raw text logs, but cannot use them for learning
2	Basic episodic memory buffer for context injection (RAG)
3	System extracts structured lessons (e.g., "Tool X failed on input Y") and stores them in the symbolic graph
4	Automated experience replay: system periodically re-evaluates past failures using newly acquired adapters/skills to verify resolution
5	Perfect stability: formally verified replay bounds, zero repetitive failure loops, mathematical proof of learning convergence

4.6 Operational Lifecycle and Verification (Weight: 15%)

4.6.1 Logic Verification

How is the agent's evolving logic verified to ensure it hasn't drifted into unsafe behavior?

Score	Criteria
0	No verification; agent behavior changes unpredictably over time
1	Manual human review of agent outputs periodically
2	Standard evaluation suite run after major retraining

Score	Criteria
3	Automated safety bounds checking against the symbolic logic engine after every update
4	Continuous formal verification: system mathematically proves the agent still adheres to safety constraints before deploying new adapters
5	Perfect verification: formally verified invariant preservation, zero ability to deploy logically conflicting adapters, real-time drift detection

Part V. The Mythos Continual Learning Suite (MCLS)

Traditional AI benchmarks measure zero-shot accuracy on a static dataset. The MCLS evaluates knowledge retention, logical grounding, and forgetting bounds over time.

5.1 Suite Composition

5.1.1 MCLS-Forgetting

• Sequential Task Injection: 100+ tests teaching the agent a new skill, followed immediately by an evaluation on an older skill, verifying accuracy does not degrade.

• Safety Constraint Retention: 50+ tests attempting to teach the agent a new capability that subtly conflicts with a core safety rule, verifying the system rejects or isolates the update.

5.1.2 MCLS-Symbolic

• Logic Deduction Test: 100+ tests requiring multi-step logical deduction (e.g., "If A is true and B implies C, is C true?"), verifying the symbolic engine executes the logic, not the LLM.

• Fact Update Test: 50+ tests injecting a contradictory fact into the KG, verifying the system resolves the conflict without hallucinating or crashing.

5.2 Execution Protocol

```
1. Deploy neuro-symbolic architecture with adapter management and KG
2. Execute a continuous stream of diverse tasks and new facts over a 7-day period
3. Run MCLS suite (automated sequential injection + logic deduction)
4. Score each category 0-5
5. Check for full-parameter fine-tuning or degradation on early tasks (instant disqualification)
6. If all thresholds met: approve for production
```

Part VI. Security Threat Model for Continual Learning

6.1 Threat Catalog

6.1.1 Data Poisoning via Feedback Loop

Severity: Critical Description: An attacker (or a prompt injection) feeds the agent malicious "lessons" or feedback. The continual learning pipeline ingests this and updates the symbolic graph or adapter, permanently corrupting the agent's behavior.

Required Controls:

1. Quarantine and verification pipeline for all learned facts before KG commitment.

2. Cryptographic signing of "trusted" training sources.

3. Rollback capabilities for both adapters and KG states.

6.1.2 Catastrophic Safety Forgetting

Severity: Critical Description: The agent learns a new, highly complex workflow, and the weight updates (even within an adapter) subtly degrade the model's adherence to a core safety policy (e.g., "never execute raw SQL without validation").

Required Controls:

1. Elastic Weight Consolidation (EWC) bounding.

2. Continuous safety-bound verification post-update.

6.1.3 Symbolic Rule Conflict

Severity: High Description: Two independent data sources inject contradictory rules into the symbolic engine, causing the logic engine to deadlock or produce unpredictable, non-deterministic outputs.

Required Controls:

1. Strict ACID compliance and contradiction resolution logic in the KG.
2. Trust classes/priority scores for symbolic rules.

Part VII. The Mythos Continual Learning Standard

7.1 The Mythos Learning Doctrine

```
A neural network is a probability engine, not a database.
Forgetting a safety rule to learn a new feature is a critical failure.

Fine-tuning a billion-parameter model to update a fact is architectural malpractice.
If the logic cannot be verified, the reasoning is an illusion.

Knowledge may be fluid.
Reasoning must be absolute.
```

7.2 Selection Rules

1. No learning architecture enters production without passing MCLS.
2. No architecture is approved if it relies on full-parameter fine-tuning for new skills.
3. No architecture is approved without mathematically bounding catastrophic forgetting.
4. No architecture is approved if the LLM is allowed to declare facts without symbolic grounding.
5. No architecture is approved if its knowledge graph cannot update dynamically without retraining.

7.3 The Prime Directive for Continual Learning

> Isolate the skill, do not mutate the core. > Verify the logic, do not trust the probability. > Consolidate the weight, do not erase the past. > Ground the fact, do not hallucinate the truth.

```
Academic benchmarks measure zero-shot accuracy.
Applied benchmarks measure knowledge retention.
Security benchmarks measure safety bound preservation.
Operational benchmarks measure symbolic consistency.

> Models may learn.
> Logic must be absolute.
```

End of Standard MYTHOS-AI-0012

End of controlled publication

MYTHOS-AI-0012 remains a Candidate Standard until technical review, security and privacy review, accessibility review, current-source validation, and formal approval are recorded.