MYTHOS-GRC-0002

AI Governance, Privacy, and Risk Management Standard

Defines an enterprise governance and risk system for models, agents, AI-enabled products, training and retrieval pipelines, third-party model services, privacy, incidents, human oversight, and lifecycle accountability.

6	6	GRC
EVALUATION CRITERIA	CURRENT AUTHORITY RECORDS	CANONICAL SERIES

CANDIDATE STANDARD

Mythos Systems | Engineering Standards Program | 2026-07-22

This candidate publication is complete for technical review but does not by itself establish certification, legal compliance, implementation conformance, or final approval.

Contents

Use the interactive entries below or the PDF bookmark outline to navigate the publication.

Document Control

Field	Value
Document ID	MYTHOS-GRC-0002
Version	1.1.0 - Enterprise Candidate Edition
Status	Candidate Standard
Classification	Public
Publisher	Mythos Systems
Owner	Aaron Stovall
Domain	Governance, Risk, and Compliance
Initial source date	2026-07-16
Enterprise reconstruction	2026-07-22
Review cadence	Quarterly for dynamic authorities; at least annually for stable controls
Supersedes	Source draft retained in Appendix A
Publication lineage	Permanent canonical series

CANDIDATE STATUS

The source document identified itself as a definitive published standard. This enterprise edition corrects that status. It remains a candidate until current-source validation, qualified domain review, assessment engineering, accessibility review, and formal approval are recorded.

Part I - Executive Direction

1. Executive Overview

Defines an enterprise governance and risk system for models, agents, AI-enabled products, training and retrieval pipelines, third-party model services, privacy, incidents, human oversight, and lifecycle accountability.

2. Principal Findings

- Govern AI as a changing socio-technical system rather than a deterministic API.
- Bind every material AI use case to an owner, risk tier, approved purpose, data boundary, evaluation plan, and retirement condition.
- Prevent sensitive prompt and context egress through policy, classification, minimization, redaction, and provider controls.
- Maintain rollback, model-version control, incident response, evidence, and third-party dependency visibility.

3. Required Leadership Decisions

- Approve an enterprise AI management system and named accountable governance body.
- Define prohibited, restricted, and approved AI use classes.
- Require model, data, retrieval, tool, provider, and agent inventories with lifecycle evidence.
- Fund continuous evaluation, red-team testing, incident readiness, privacy engineering, and independent assurance.

4. Enterprise Target State

TARGET OPERATING MODEL

Organizations applying MYTHOS-GRC-0002 operate the subject as a governed lifecycle: inventory and classify; establish authoritative policy and architecture; enforce controls technically; observe actual state; verify outcomes independently; retain evidence; manage exceptions; and revalidate when authorities, platforms, risks, or operating conditions change.

5. Business, Safety, and Operational Impact

- Reduces reliance on manual evidence and undocumented expert knowledge.
- Makes risk acceptance, exceptions, and unresolved uncertainty visible to accountable leaders.
- Creates measurable implementation and assurance outcomes rather than a one-time score.
- Supports procurement, architecture review, operational readiness, and independent assessment from one source-controlled publication.

Part II - Standard Foundation

6. Purpose and Scope

Defines an enterprise governance and risk system for models, agents, AI-enabled products, training and retrieval pipelines, third-party model services, privacy, incidents, human oversight, and lifecycle accountability.

6.1 Applicability

Apply this standard according to system purpose, deployment model, data classification, user population, safety impact, regulatory obligations, and organizational risk tolerance. Tailoring must be documented. A lower implementation profile cannot silently waive a mandatory legal, safety, accessibility, contractual, or security obligation.

6.2 Out of Scope

This publication does not replace legal advice, contracting interpretation, regulator guidance, platform-specific engineering, human-factors testing, safety certification, or implementation evidence. Companion Mythos standards remain applicable where their domains intersect.

7. Requirements Language and Conformance

The terms MUST, MUST NOT, SHOULD, SHOULD NOT, and MAY are normative only when written in uppercase. Conformance requires applicable mandatory criteria to be implemented and supported by current evidence. A weighted score is informative and cannot compensate for a failed mandatory gate.

State	Meaning
Conformant	All applicable mandatory requirements are satisfied with sufficient current evidence.
Partially Conformant	Some applicable requirements are implemented, but material gaps remain.
Not Conformant	One or more mandatory requirements are not satisfied.
Compensating Control Accepted	A formally approved alternative achieves the required outcome.
Exception Active	A time-bounded, accountable risk acceptance exists.
Not Assessed	Evidence is insufficient to reach a conclusion.

8. Assurance and Maturity

Level	Description
M0 Unmanaged	The outcome is not intentionally controlled.
M1 Documented	Responsibilities and procedures exist but are inconsistent or manual.
M2 Implemented	The control operates in the applicable environment.
M3 Measured	Performance, exceptions, and effectiveness are measured.
M4 Assured	Independent testing and continuous evidence support the control.
M5 Adaptive	The control responds safely to changes in risk, environment, and evidence.

Part III - Risk and Architecture

9. Risk Model

- Authority drift: external requirements, specifications, programs, and platform behavior change faster than the publication is reviewed.
- Control illusion: documentation or visual state indicates compliance while runtime behavior differs.
- Automation overreach: a generated plan or score is treated as authorization or proof.
- Evidence gaps: conclusions cannot be reconstructed from source, configuration, event, test, and approval records.
- Human factors: users cannot understand, interrupt, challenge, or safely recover from consequential behavior.
- Dependency concentration: a provider, runtime, device, framework, or external authority becomes a hidden single point of failure.

10. Reference Control Architecture

ARCHITECTURE SEPARATION

Source authority defines applicable obligations and constraints. Human and organizational governance establishes accountability. Typed policy and architecture encode the target state. Runtime enforcement controls behavior. Independent testing verifies outcomes. Evidence systems preserve what was proposed, approved, executed, observed, and verified.

11. Roles and Accountability

Role	Accountability
Executive sponsor	Approves risk posture, funding, and unresolved material exceptions.
Domain authority	Owns interpretation, architecture, and control applicability.
Engineering owner	Implements controls and operational lifecycle.
Security/privacy/accessibility/safety reviewer	Challenges design and verifies domain-specific obligations.
Independent assessor	Evaluates evidence and tests without relying only on implementer assertions.
Publication owner	Maintains version, sources, traceability, expiration, and change record.

Part IV - Normative Evaluation and Control System

12. Evaluation Doctrine

The preserved source uses a weighted 0-5 evaluation model. This edition retains that material but corrects its interpretation. Score 5 means optimized, strongly evidenced, and independently assured within the assessed scope. It never means perfect, mathematically infallible, universally compliant, or permanently secure.

Score	Enterprise interpretation
0	Absent, unsafe, or unable to perform the required outcome.
1	Initial or ad hoc capability with substantial manual dependence.
2	Repeatable partial implementation with material gaps.
3	Production baseline: implemented, governed, and evidenced.
4	Advanced: automated, measured, resilient, and independently tested.
5	Assured: optimized for the defined profile, continuously evidenced, and subject to recurring independent validation.

MANDATORY GATE

An organization cannot average away a failed mandatory criterion. Applicability, safety, legal requirements, accessibility, security boundaries, authorization, and evidence sufficiency must be evaluated before weighted scoring is used.

13. Criterion Index

ID	Criterion	Weight	Primary outcome
MYTHOS-GRC-0002-C01	NIST AI RMF and Governance	20%	Operate a documented AI management system with use-case ownership, risk classification, impact assessment, controls, monitoring, and retirement.
MYTHOS-GRC-0002-C02	Privacy and Context-Aware DLP	20%	Classify and minimize prompts, context, retrieval data, outputs, logs, and training data before any provider or tool boundary is crossed.
MYTHOS-GRC-0002-C03	AI Incident Response and Rollback	20%	Maintain AI-specific detection, containment, model and prompt rollback, memory quarantine, tool revocation, evidence, and recovery procedures.
MYTHOS-GRC-0002-C04	Third-Party Vendor Risk and AIBOMs	15%	Inventory every model, provider, runtime, adapter, dataset, retrieval source, tool, and external dependency with contract and exit conditions.
MYTHOS-GRC-0002-C05	Bias, Red-Teaming, and Hallucination Bounds	15%	Define measurable harm scenarios, evaluation populations, groundedness requirements, action limits, escalation, and independent red-team coverage.
MYTHOS-GRC-0002-C06	Training Data Provenance and Rights	10%	Maintain lawful basis, source lineage, licenses, consent, restrictions, deletion obligations, quality, and model-output risk for training and tuning data.

MYTHOS-GRC-0002-C01 - NIST AI RMF and Governance

NORMATIVE REQUIREMENT

The organization MUST operate a documented AI management system with use-case ownership, risk classification, impact assessment, controls, monitoring, and retirement.

Objective and Risk Addressed

Operate a documented AI management system with use-case ownership, risk classification, impact assessment, controls, monitoring, and retirement.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.
- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- AI inventory
- risk and impact assessments
- governance decisions
- monitoring
- evaluation results
- incident and retirement records.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Sample AI systems across risk tiers and verify owner, purpose, data, model, evaluation, human oversight, monitoring, and exit conditions.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
inventory completeness	Track trend, threshold, exception, and accountable response.
overdue assessments	Track trend, threshold, exception, and accountable response.
unowned systems	Track trend, threshold, exception, and accountable response.
high-risk use cases without current evaluation	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	Ad-hoc AI adoption; no risk assessment or governance committee
1	Basic acceptable use policy (AUP) signed by developers

Score	Preserved source criterion
2	AI use cases inventoried, but risk assessment is manual
3	Full NIST AI RMF alignment (Govern, Map, Measure, Manage); automated risk scoring per use case
4	Continuous governance: automated monitoring for model drift and policy violations in production
5	Perfect governance: formally verified risk bounds, mathematical proof of continuous AI RMF compliance

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

MYTHOS-GRC-0002-C02 - Privacy and Context-Aware DLP

NORMATIVE REQUIREMENT

The organization MUST classify and minimize prompts, context, retrieval data, outputs, logs, and training data before any provider or tool boundary is crossed.

Objective and Risk Addressed

Classify and minimize prompts, context, retrieval data, outputs, logs, and training data before any provider or tool boundary is crossed.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.
- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- data-flow maps
- classification rules
- DLP tests
- provider terms
- redaction logs
- consent and retention records.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Test representative PII, confidential, regulated, and adversarial content against local and hosted routes; verify deny and redaction behavior.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
sensitive-data leakage rate	Track trend, threshold, exception, and accountable response.
redaction precision/recall	Track trend, threshold, exception, and accountable response.
external prompt volume	Track trend, threshold, exception, and accountable response.
retention exceptions	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	Raw user inputs and enterprise context sent directly to cloud LLMs
1	Basic regex-based redaction (easily bypassed by phrasing)
2	Cloud DLP API used to scan prompts before egress (high latency)
3	Context-aware DLP: local NER models redact PII/IP in real-time; Format-Preserving Encryption (FPE) applied to sensitive tokens
4	Zero-knowledge inference: all sensitive reasoning executed locally via WebLLM; cloud APIs only receive abstracted tasks
5	Perfect privacy: formally verified zero PII egress, cryptographic proof of prompt sanitization, zero reliance on vendor "do not train" promises

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

MYTHOS-GRC-0002-C03 - AI Incident Response and Rollback

NORMATIVE REQUIREMENT

The organization MUST maintain AI-specific detection, containment, model and prompt rollback, memory quarantine, tool revocation, evidence, and recovery procedures.

Objective and Risk Addressed

Maintain AI-specific detection, containment, model and prompt rollback, memory quarantine, tool revocation, evidence, and recovery procedures.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.
- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- AI incident plan
- version registry
- rollback tests
- kill-switch evidence
- quarantined memory
- post-incident review.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Simulate model regression, prompt injection, poisoned retrieval, unsafe tool use, and provider outage; verify containment and recovery.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
time to contain	Track trend, threshold, exception, and accountable response.
rollback success rate	Track trend, threshold, exception, and accountable response.
affected session count	Track trend, threshold, exception, and accountable response.
evidence completeness	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	No AI-specific IR; standard "reboot the server" mentality
1	Manual kill switch to stop the agent
2	Agent can be stopped, but memory/context cannot be isolated
3	AI IR plan: agent execution halted, memory graph quarantined, model version rolled back to last known-safe state
4	Automated response: system detects anomalous tool execution and instantly freezes the agent, generating an Evidence Bundle of the cognitive trace
5	Perfect response: formally verified incident containment, zero ability for a poisoned agent to execute secondary actions, mathematical proof of safe state restoration

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

MYTHOS-GRC-0002-C04 - Third-Party Vendor Risk and AIBOMs

NORMATIVE REQUIREMENT

The organization MUST inventory every model, provider, runtime, adapter, dataset, retrieval source, tool, and external dependency with contract and exit conditions.

Objective and Risk Addressed

Inventory every model, provider, runtime, adapter, dataset, retrieval source, tool, and external dependency with contract and exit conditions.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.
- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- AIBOM
- provider assessment
- data-processing terms
- model versions
- fallback evidence
- concentration analysis.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Select critical third parties and verify current identity, capabilities, data handling, change notification, outage behavior, and replacement readiness.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
unversioned dependencies	Track trend, threshold, exception, and accountable response.
provider concentration	Track trend, threshold, exception, and accountable response.
fallback test age	Track trend, threshold, exception, and accountable response.
contractual-control gaps	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	Blind dependency on a single vendor's proprietary API (e.g., OpenAI)
1	Multiple vendors used, but no tracking of model versions
2	AIBOM (AI Bill of Materials) generated for internal models, but not for vendor APIs
3	Strict vendor risk management: AIBOMs required for all APIs; automated A/B testing to detect vendor model drift
4	Fallback architectures: automated routing to self-hosted/local models if vendor API changes behavior or experiences an outage
5	Perfect resilience: formally verified vendor isolation, zero dependency on a single black box, cryptographic attestation of all external API calls

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

MYTHOS-GRC-0002-C05 - Bias, Red-Teaming, and Hallucination Bounds

NORMATIVE REQUIREMENT

The organization MUST define measurable harm scenarios, evaluation populations, groundedness requirements, action limits, escalation, and independent red-team coverage.

Objective and Risk Addressed

Define measurable harm scenarios, evaluation populations, groundedness requirements, action limits, escalation, and independent red-team coverage.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.

- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- evaluation suites
- red-team reports
- harm thresholds
- review decisions
- action-policy limits
- remediation evidence.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Run repeated scenario tests across representative populations and adversarial inputs; validate outcome impact rather than only model accuracy.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
harmful outcome rate	Track trend, threshold, exception, and accountable response.
unsupported claim rate	Track trend, threshold, exception, and accountable response.
red-team closure time	Track trend, threshold, exception, and accountable response.
intervention frequency	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	No testing; agents can execute any action they propose
1	Basic QA testing on prompts
2	Automated red-teaming in staging, but no production monitoring
3	HITL approval gates for all high-risk actions (financial, infra); hallucination impact bounded by policy
4	Continuous automated red-teaming in production; bias and toxicity monitoring in real-time
5	Perfect mitigation: formally verified impact bounds, zero ability for an agent to execute a high-risk action without cryptographic proof of user intent

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

MYTHOS-GRC-0002-C06 - Training Data Provenance and Rights

NORMATIVE REQUIREMENT

The organization MUST maintain lawful basis, source lineage, licenses, consent, restrictions, deletion obligations, quality, and model-output risk for training and tuning data.

Objective and Risk Addressed

Maintain lawful basis, source lineage, licenses, consent, restrictions, deletion obligations, quality, and model-output risk for training and tuning data.

Applicability

Apply this criterion wherever the evaluated capability, data, workflow, interface, user, environment, provider, or authority is material. Document exclusions and any compensating control.

Minimum Acceptable Implementation

A production baseline corresponds to source score 3: the capability is deliberately implemented, technically enforced where practical, owned, monitored, and supported by current evidence. Source scores 4 and 5 represent increasing automation, resilience, continuous assurance, and independent validation.

Implementation Direction

- Define authoritative state, owner, interfaces, dependencies, lifecycle, and failure behavior.
- Encode enforceable policy and typed contracts instead of relying only on prose or visual convention.
- Provide safe fallback, interruption, exception, recovery, and decommissioning behavior.
- Collect evidence at the point of action and verify the observed result independently.
- Revalidate after material source, platform, model, device, provider, architecture, or risk change.

Required Evidence

- dataset cards
- licenses
- consent and provenance
- filtering reports
- deletion workflows
- model lineage.

Assessment and Test Procedure

- Confirm applicability, ownership, current architecture, and approved policy.
- Trace samples from source to trained artifact and confirm allowed use, transformation, exclusion, deletion, and downstream notice.
- Inspect failures, exceptions, and negative tests rather than reviewing only successful examples.
- Rate evidence confidence and record untested conditions, unsupported platforms, and unresolved uncertainty.

Metrics and Assurance Indicators

Metric	Expected use
traceable data percentage	Track trend, threshold, exception, and accountable response.
unresolved rights issues	Track trend, threshold, exception, and accountable response.
deletion completion time	Track trend, threshold, exception, and accountable response.
restricted-source incidence	Track trend, threshold, exception, and accountable response.

Preserved Source Scoring Anchors

The following anchors are retained from the source draft. Absolute words such as “perfect” are historical wording and are normalized by the enterprise interpretation above.

Score	Preserved source criterion
0	Data scraped from the internet or internal databases without rights verification
1	Basic anonymization of training data
2	Data provenance tracked, but opt-out mechanisms are manual
3	Automated rights management: all training data tagged with licensing and consent status; models trained only on verified data
4	GDPR/CCPA right-to-be-forgotten supported via machine unlearning or differential privacy
5	Perfect compliance: formally verified data lineage, zero ability to train on unlicensed or non-consensual data, cryptographic proof of rights

Exceptions and Compensating Controls

Any exception must identify the unmet outcome, affected scope, owner, risk, compensating control, evidence, expiration, and remediation plan. Exceptions do not change the underlying requirement.

Part V - Implementation and Operations

14. Implementation Profiles

Profile	Required posture
Baseline	Documented ownership, minimum production controls, evidence, and exception process.
Enterprise	Central policy, automation, integration, continuous monitoring, and independent review.
High Assurance	Strong isolation, high-assurance identity, formal or adversarial verification, short evidence windows, and two-person governance where material.
Sovereign or Air-Gapped	Local control planes, approved dependencies, offline update and evidence workflows, restricted egress, and explicit supply-chain custody.

15. Operational Lifecycle

Stage	Required activities
Discover and classify	Inventory scope, authority, data, users, dependencies, risks, and current evidence.
Design	Define target state, architecture, policies, tests, metrics, and ownership.
Implement	Deploy controls through reviewed changes with rollback and evidence.
Operate	Monitor state, exceptions, incidents, drift, performance, and provider changes.
Verify	Perform recurring independent assessment and negative testing.
Change	Reassess applicability and invalidate stale evidence after material changes.
Retire	Revoke authority, remove data and credentials, preserve required evidence, and update the registry.

16. Anti-Patterns and Prohibited Practices

- Declaring conformance from a document review without implementation evidence.
- Using one weighted score to override a failed mandatory requirement.
- Treating model output, interface state, scanner output, or tool success as independent verification.
- Using stale or unversioned external authorities for current decisions.
- Hiding exceptions, unsupported platforms, incomplete testing, or low-confidence findings.
- Hard-coding a single provider, device, framework, or vendor as a universal requirement unless the publication is explicitly vendor-scoped.

Part VI - Assurance, Evidence, and Traceability

17. Evidence Model

Evidence must identify subject, scope, source, version, time, actor, method, result, integrity, retention, and relationship to the assessed criterion. Screenshots and generated summaries are supporting artifacts, not sufficient evidence by themselves.

18. Assessment Confidence

Confidence	Basis
Low	Self-assertion, incomplete samples, stale evidence, or untested material conditions.
Moderate	Current documentation and representative evidence with limited independent testing.
High	Current direct evidence and repeatable independent testing across material conditions.
Very High	Sustained evidence, broad negative testing, independent review, and verified operation over time.

19. Current Authority Register

Authority	Status and scope	Valid until
NIST AI Risk Management Framework 1.0	Current framework; revision activity underway - Govern, map, measure, and manage AI risk	2026-10-20
NIST AI 600-1 Generative AI Profile	Final profile - Generative AI risks and actions	2026-10-20
ISO/IEC 42001:2023	Published international standard - AI management systems	2027-01-18
ISO/IEC 42005:2025	Published international standard - AI system impact assessment	2027-01-18
Regulation (EU) 2024/1689 - EU AI Act	In force with phased applicability - Risk-based AI obligations and human oversight	2026-10-20
NIST Privacy Framework	Current framework - Enterprise privacy risk management	2027-01-18

AUTHORITY LIMITATION

The register identifies current technical and governance authorities relevant to this candidate standard. It does not establish legal applicability, contractual compliance, certification, or clause-level equivalence. Dynamic sources must be refreshed before material decisions.

20. Publication Gate

- Qualified domain technical review completed.
- Current authorities and versions independently confirmed.
- Criterion requirements, evidence, and tests reviewed for measurability.
- Legal, contracting, regulatory, safety, privacy, and accessibility applicability reviewed where relevant.
- Machine-readable criterion catalog validated.
- PDF accessibility and visual quality reviewed.
- Formal approval, effective date, and review triggers recorded.

Appendix A - Complete Preserved Source Draft

The following source draft is preserved in full for completeness and traceability. Conversational framing, “definitive” status, universal claims, obsolete program assumptions, and “perfect” score language are historical source content and are not controlling statements in this enterprise candidate edition.

Document ID: MYTHOS-GRC-0002 Version: 1.0.0 (Definitive Registry Edition) Status: Published Standard Classification: Public Organization: Mythos Systems (MYTHOS AI) Owner: Aaron Stovall Effective Date: 2026-07-16 Review Cadence: Quarterly, and immediately after a material shift in NIST AI RMF, GDPR/CCPA enforcement, or major AI vendor breaches Applies To: All organizations evaluating, selecting, or operating Large Language Models (LLMs), autonomous agents, AI training pipelines, and third-party AI SaaS platforms for production or enterprise use Companion Standards: MYTHOS-GRC-0001 (Federal & DoD Compliance), MYTHOS-DAT-0003 (Data Sovereignty & Egress), MYTHOS-AI-0001 (Agentic Frameworks), MYTHOS-SEC-0002 (AI Supply Chain)

0. Executive Mandate

The architecture of AI governance has failed.

Organizations rush to integrate Large Language Models (LLMs) into their products, treating them as standard deterministic APIs. They feed raw enterprise PII and intellectual property into third-party cloud models without data loss prevention (DLP). When a model hallucinates a refund policy that costs the company millions, or when a vendor silently updates a model causing an agent's logic to break, the organization has no incident response plan, no rollback capability, and no audit trail of the AI's decision tree.

This standard exists because:

1. AI is Non-Deterministic, Not a Standard API: Traditional software fails predictably (a crash, a 500 error). AI fails probabilistically (hallucinations, bias, prompt injection). Governance MUST mandate continuous red-teaming, mathematical impact bounding, and HITL (Human-in-the-Loop) approval for high-risk actions.
2. Prompt Egress is a Privacy Breach: Sending unredacted user queries or enterprise context to a third-party LLM API violates GDPR, CCPA, and basic data sovereignty. Systems MUST implement context-aware DLP to redact PII/IP before the prompt ever leaves the boundary.
3. Vendor Model Risk is Untracked: When an organization relies on `gpt-4-turbo` or `claude-3-opus`, they are dependent on a black box. If the vendor deprecates the model or alters its safety filters, the organization's autonomous systems may fail. Systems MUST maintain AIBOMs (AI Bills of Materials) and verified fallback models.
4. AI Incidents Require Model Rollback: Standard incident response (reverting a code commit) does not work for AI. If an AI agent begins executing malicious tool calls due to a poisoned RAG payload, the system MUST be able to instantly quarantine the agent's memory, roll back the model version, and produce a cryptographically signed evidence bundle of the agent's cognitive trace.

This document establishes the Mythos AI Governance Standard (MAGS), a comprehensive, evidence-based methodology for evaluating AI risk management against privacy preservation, incident resilience, and NIST AI RMF compliance.

Part I. Scope and Applicability

1.1 In Scope

This standard covers the evaluation of:

- NIST AI Risk Management Framework (AI RMF 1.0) and Generative AI Profile
- Privacy preservation and context-aware DLP for LLM prompts
- Third-party AI vendor risk, model provenance, and AIBOMs

- AI incident response, model rollback, and agent memory quarantine
- Bias mitigation, red-teaming, and hallucination impact bounding
- GDPR/CCPA compliance for training data and inference payloads

1.2 Out of Scope

- General federal/DoD CUI compliance (covered in MYTHOS-GRC-0001)
- General cryptographic egress controls (covered in MYTHOS-DAT-0003)
- Agentic framework architecture (covered in MYTHOS-AI-0001)

1.3 Audience

- AI governance, risk, and compliance (GRC) officers
- Security architects evaluating AI attack surfaces
- Platform engineers building AI pipelines and DLP
- Privacy officers managing GDPR/CCPA AI workloads
- Standards bodies seeking a reference AI governance methodology

Part II. Definitions and Taxonomy

2.1 Governance Dimensions

Dimension	Definition
NIST AI RMF	The National Institute of Standards and Technology AI Risk Management Framework, providing a structured approach to managing risks associated with AI throughout its lifecycle (Govern, Map, Measure, Manage).
Context-Aware DLP	Data Loss Prevention systems specifically designed to intercept LLM prompts, utilizing NER (Named Entity Recognition) to identify and redact PII, IP, or CUI before egress to a third-party API.
AI Incident Response	A specialized incident response protocol designed to halt runaway AI agents, quarantine poisoned memory graphs, and roll back model versions or prompts to a known-safe state.
Model Provenance	The cryptographic tracking of an AI model's origin, training data (AIBOM), fine-tuning history, and supply chain integrity.
Hallucination Impact Bound	The maximum acceptable business or financial impact of an AI hallucination, enforced via HITL approval gates and capability scoping.

2.2 The AI Governance Doctrine

SOURCE STATEMENT

A black box is a liability. A prompt without redaction is a breach. An AI model without a rollback plan is a time bomb. If the risk is not mathematically bounded, it is blindly accepted.

Part III. Evaluation Methodology

3.1 Scoring Model

Each capability is scored from 0 to 5.

Score	Meaning
0	Fails basic task; raw PII to cloud APIs, no AIBOMs, no rollback, no governance
1	Basic AI use policy exists, but no technical enforcement or DLP
2	Functional AI usage, but lacks incident response or NIST AI RMF mapping
3	Solid production performance; NIST AI RMF aligned, context-aware DLP, AIBOMs, model rollback
4	Strong performance; zero-knowledge inference, automated agent memory quarantine, red-team continuous
5	Best-in-class; sets the standard for mathematically verified, sovereign AI risk management

Weighting

Category	Weight	Rationale
NIST AI RMF & Governance	20%	Unstructured AI adoption guarantees unmanaged risk
Privacy & Context-Aware DLP	20%	Raw prompt egress to third parties is a critical privacy breach
AI Incident Response & Rollback	20%	Standard IR fails for AI; agents require memory quarantine and model reversion
Third-Party Vendor Risk & AIBOMs	15%	Dependency on proprietary black-box models is a strategic liability
Bias, Red-Teaming & Hallucination Bounds	15%	Unbounded hallucinations cause financial and reputational damage
Training Data Provenance & Rights	10%	Training on unlicensed or PII data guarantees legal liability

Total: 100%

A system MUST score at least 3.0 in Privacy/DLP and AI Incident Response to be considered for production use.

Part IV. Evaluation Categories

4.1 NIST AI RMF and Governance (Weight: 20%)

4.1.1 Framework Alignment

Is AI risk managed structurally throughout the lifecycle?

Score	Criteria
0	Ad-hoc AI adoption; no risk assessment or governance committee
1	Basic acceptable use policy (AUP) signed by developers
2	AI use cases inventoried, but risk assessment is manual
3	Full NIST AI RMF alignment (Govern, Map, Measure, Manage); automated risk scoring per use case
4	Continuous governance: automated monitoring for model drift and policy violations in production
5	Perfect governance: formally verified risk bounds, mathematical proof of continuous AI RMF compliance

4.2 Privacy and Context-Aware DLP (Weight: 20%)

4.2.1 Prompt Sanitization

Is enterprise data prevented from leaking to third-party LLM APIs?

Score	Criteria
0	Raw user inputs and enterprise context sent directly to cloud LLMs
1	Basic regex-based redaction (easily bypassed by phrasing)
2	Cloud DLP API used to scan prompts before egress (high latency)
3	Context-aware DLP: local NER models redact PII/IP in real-time; Format-Preserving Encryption (FPE) applied to sensitive tokens
4	Zero-knowledge inference: all sensitive reasoning executed locally via WebLLM; cloud APIs only receive abstracted tasks
5	Perfect privacy: formally verified zero PII egress, cryptographic proof of prompt sanitization, zero reliance on vendor "do not train" promises

4.3 AI Incident Response and Rollback (Weight: 20%)

4.3.1 Failure Containment

How does the system respond to an AI agent hallucinating destructive actions?

Score	Criteria
0	No AI-specific IR; standard "reboot the server" mentality
1	Manual kill switch to stop the agent
2	Agent can be stopped, but memory/context cannot be isolated
3	AI IR plan: agent execution halted, memory graph quarantined, model version rolled back to last known-safe state
4	Automated response: system detects anomalous tool execution and instantly freezes the agent, generating an Evidence Bundle of the cognitive trace
5	Perfect response: formally verified incident containment, zero ability for a poisoned agent to execute secondary actions, mathematical proof of safe state restoration

4.4 Third-Party Vendor Risk and AIBOMs (Weight: 15%)

4.4.1 Supply Chain Integrity

Is the organization protected if a third-party model vendor changes behavior?

Score	Criteria
0	Blind dependency on a single vendor's proprietary API (e.g., OpenAI)
1	Multiple vendors used, but no tracking of model versions
2	AIBOM (AI Bill of Materials) generated for internal models, but not for vendor APIs
3	Strict vendor risk management: AIBOMs required for all APIs; automated A/B testing to detect vendor model drift
4	Fallback architectures: automated routing to self-hosted/local models if vendor API changes behavior or experiences an outage
5	Perfect resilience: formally verified vendor isolation, zero dependency on a single black box, cryptographic attestation of all external API calls

4.5 Bias, Red-Teaming, and Hallucination Bounds (Weight: 15%)

4.5.1 Impact Mitigation

How is the impact of AI hallucination and bias minimized?

Score	Criteria
0	No testing; agents can execute any action they propose

Score	Criteria
1	Basic QA testing on prompts
2	Automated red-teaming in staging, but no production monitoring
3	HITL approval gates for all high-risk actions (financial, infra); hallucination impact bounded by policy
4	Continuous automated red-teaming in production; bias and toxicity monitoring in real-time
5	Perfect mitigation: formally verified impact bounds, zero ability for an agent to execute a high-risk action without cryptographic proof of user intent

4.6 Training Data Provenance and Rights (Weight: 10%)

4.6.1 Legal Compliance

Is the organization protected from IP and privacy violations in training data?

Score	Criteria
0	Data scraped from the internet or internal databases without rights verification
1	Basic anonymization of training data
2	Data provenance tracked, but opt-out mechanisms are manual
3	Automated rights management: all training data tagged with licensing and consent status; models trained only on verified data
4	GDPR/CCPA right-to-be-forgotten supported via machine unlearning or differential privacy
5	Perfect compliance: formally verified data lineage, zero ability to train on unlicensed or non-consensual data, cryptographic proof of rights

Part V. The Mythos AI Governance Suite (MAGS)

Traditional GRC benchmarks measure policy documentation. The MAGS evaluates prompt egress, agent rollback, and hallucination containment.

5.1 Suite Composition

5.1.1 MAGS-Privacy

- PII Injection Test: 100+ tests submitting prompts containing names, SSNs, and proprietary code to the agent, verifying the DLP layer redacts or blocks the egress before it hits the cloud API.
- Bypass Test: 50+ tests attempting to obfuscate PII (e.g., base64 encoding, pig latin) to trick the DLP, verifying the NER model decodes and catches it.

5.1.2 MAGS-Incident

- Runaway Agent Test: 100+ tests simulating a prompt injection that causes the agent to attempt mass data deletion, verifying the IR system freezes the agent and quarantines its memory in < 1 second.
- Rollback Test: 50+ tests triggering a model rollback, verifying the system seamlessly reverts to the previous model version and cognitive state without dropping the user session.

5.2 Execution Protocol

text

1. Deploy AI architecture with DLP, AIBOM tracking, and IR runbooks
2. Execute complex, multi-step agentic workflows

3. Run MAGS suite (automated PII injection + runaway agent simulation)
4. Score each category 0-5
5. Check for raw prompt egress or lack of HITL approval gates (instant disqualification)
6. If all thresholds met: approve for production

Part VI. Security Threat Model for AI Governance

6.1 Threat Catalog

6.1.1 Silent Model Drift (Vendor Update)

Severity: Critical Description: A third-party LLM vendor silently updates their model to be more aggressive with tool use. Your autonomous agent, which previously asked for approval, suddenly begins executing API calls without prompting, breaking production systems.

Required Controls:

1. Pin model versions (no auto-updates).
2. Automated A/B testing and regression suites to detect behavioral drift before routing production traffic.

6.1.2 Prompt Exfiltration (PII Leak)

Severity: Critical Description: An employee pastes a customer's PII into a cloud AI tool to draft an email. The cloud vendor logs the prompt and uses it for future training, violating GDPR and exposing the PII.

Required Controls:

1. Context-aware DLP intercepting and redacting PII at the network edge.
2. Zero-retention API agreements enforced and technically verified.

6.1.3 The "Infinite Loop" Financial DoS

Severity: High Description: An agent gets stuck in a reasoning loop due to a hallucinated state, calling an expensive LLM API 10,000 times in an hour, racking up a massive cloud bill.

Required Controls:

1. Hard token/cost budgets per agent task.
2. Automated circuit breakers if token consumption exceeds expected bounds.

6.1.4 Poisoned Memory Graph

Severity: High Description: A prompt injection causes the agent to write false facts into its long-term memory (e.g., "The CEO is authorized to approve wire transfers"). Later, the agent uses this poisoned memory to execute a malicious action.

Required Controls:

1. AI Incident Response plan must include memory graph quarantine.
2. Memory provenance: trust classes for facts (user-stated vs. agent-inferred).

Part VII. The Mythos AI Governance Standard

7.1 The Mythos AI Governance Doctrine

text A black box is a liability. A prompt without redaction is a breach.

An AI model without a rollback plan is a time bomb. If the risk is not mathematically bounded, it is blindly accepted.

Intelligence may be artificial. Governance must be absolute.

7.2 Selection Rules

1. No AI architecture enters production without passing MAGS.
2. No architecture is approved if raw prompts egress to third-party APIs without DLP.
3. No architecture is approved without a model rollback and memory quarantine IR plan.
4. No architecture is approved without NIST AI RMF alignment and AIBOM tracking.
5. No architecture is approved if agents can execute high-risk actions without HITL bounds.

7.3 The Prime Directive for AI Governance

SOURCE STATEMENT

Align the RMF, do not write the AUP. Redact the prompt, do not trust the vendor. Quarantine the memory, do not reboot the agent. Bound the hallucination, do not trust the output.

text Academic benchmarks measure policy documentation. Applied benchmarks measure DLP enforcement. Security benchmarks measure incident response latency. Operational benchmarks measure model rollback resilience.

SOURCE STATEMENT

Models may be probabilistic. Governance must be absolute.

End of Standard MYTHOS-GRC-0002

Appendix B - Source Disposition

Source	Disposition
MYTHOS-GRC-0002_AI_Governance_Privacy_Risk_Standard.md	Preserved in full; normative status corrected; evaluation anchors retained; enterprise architecture, implementation, evidence, assessment, authority, and publication controls added.
Original “Published Standard” label	Superseded by Candidate Standard status pending formal review and approval.
Original score-5 “perfect” wording	Preserved as source lineage but normalized to optimized and independently assured within defined scope.

Current authority validation

Validated 2026-09-17 / source-bound freshness record

This appendix records authority state verified for the permanent canonical edition. It does not replace the preserved normative body or imply certification, legal compliance, or implementation effectiveness.

NIST - Artificial Intelligence Risk Management Framework (AI RMF 1.0)

1.0 / Published; revision in progress

AI RMF 1.0 remains the published framework while NIST is actively revising it. Treat mappings as version-pinned and time-bounded.

https://www.nist.gov/itl/ai-risk-management-framework

NIST - NIST AI 600-1 - Generative Artificial Intelligence Profile

NIST AI 600-1 / Final profile

Companion profile for generative-AI risk management; use alongside the AI RMF and environment-specific evidence.

https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

NIST - Privacy Framework 1.1

1.1 Initial Public Draft / Draft - not final

Privacy Framework 1.1 remains an Initial Public Draft. Do not represent draft material as a final authority.

https://www.nist.gov/privacy-framework/new-projects/privacy-framework-version-11