ENGINEERING STANDARDS LIBRARY / ARTIFICIAL INTELLIGENCE

AI Avatar, Persona, & Provenance Standard

The evaluation of AI avatars and personas has failed.

PERMANENT DOCUMENT ID

MYTHOS-AI-0010

PUBLICATION

Candidate Standard
Version 1.0.0-candidate

ASSURANCE PROFILE

Standard / Advanced
High Assurance

PERMANENT PUBLICATION IDENTITY

AI Avatar, Persona, & Provenance Standard

DOCUMENT ID
MYTHOS-AI-0010

VERSION
1.0.0-candidate

STATUS
Candidate Standard

PUBLISHER
Mythos Systems

PUBLICATION DATE
2026-07-24

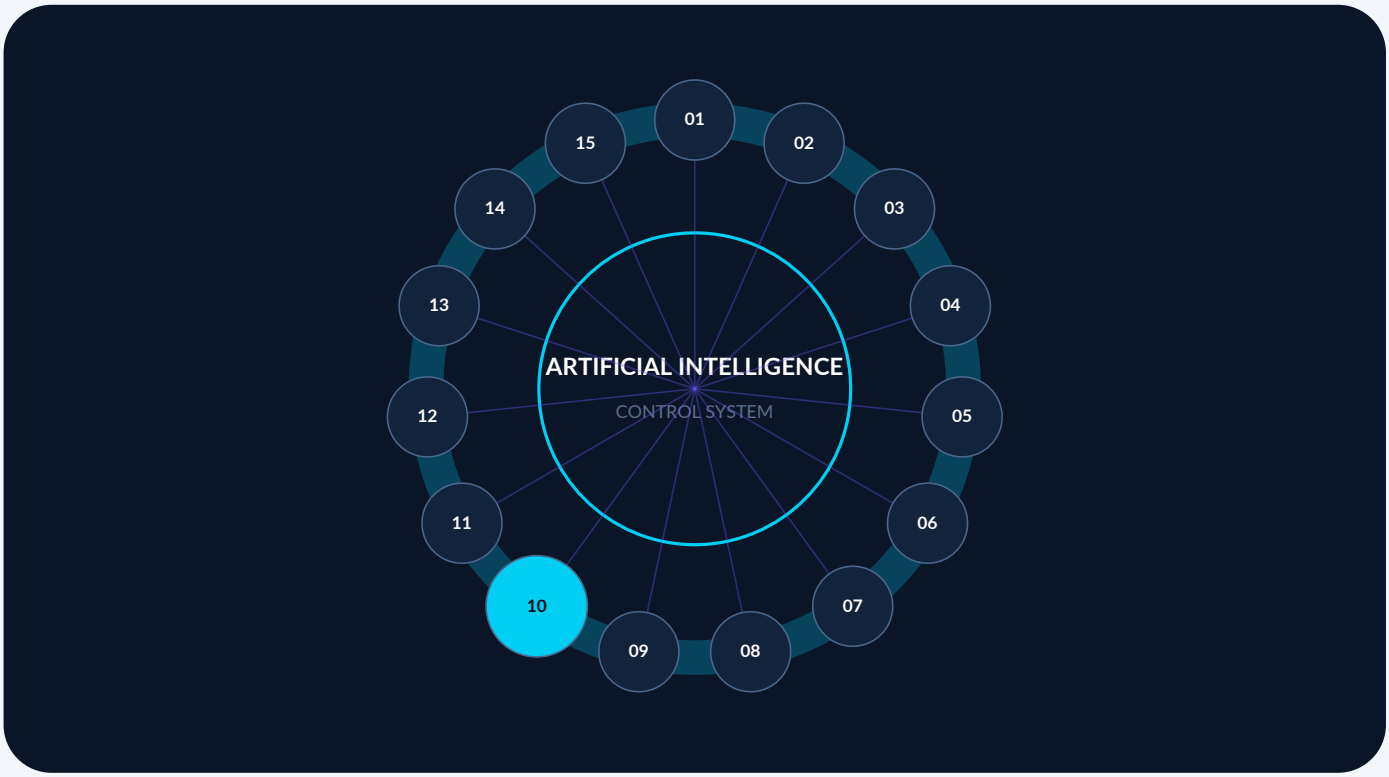
SCHEDULED REVIEW
2027-07-24

21	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-0010 inside the artificial intelligence and agentic systems constellation

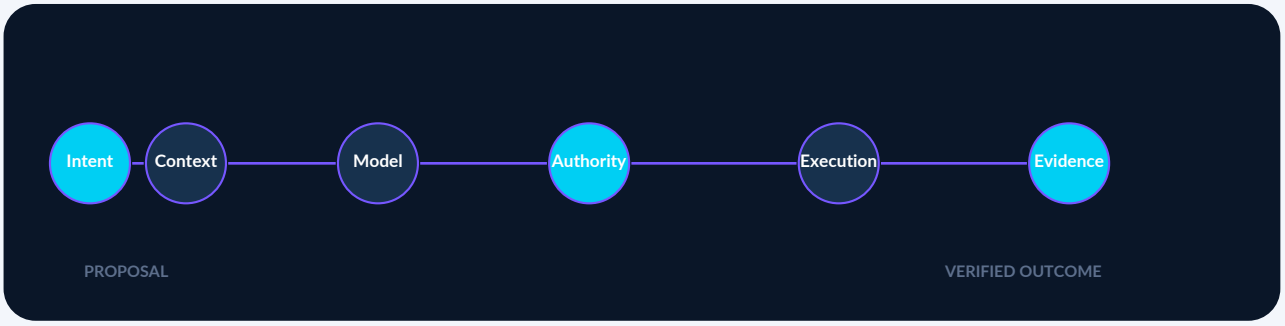


Connected assurance	Architecture, implementation, operations, evidence, and lifecycle are evaluated as one controlled system.
Specialization rule	Companion standards may add precision, but cannot silently weaken this publication.

Executive orientation

Establishes requirements for AI presence, persona, voice, visual embodiment, consent, provenance, continuity, disclosure, impersonation resistance, and user control.

WHY THIS STANDARD EXISTS	The evaluation of AI avatars and personas has failed.
OPERATIONAL SCOPE	This standard covers the evaluation of: - Avatar rendering architectures (VRM, Live2D, 3D GLB, 2D canvas, minimal orb) - Avatar state machines and deterministic event-driven state transitions - Persona definitions and communication behavior governance - Voice and avatar independence (configurable separation) - Lip-sync and viseme generation standards - Avatar package manifests and marketplace distribution - AI-generated media provenance (C2PA, Sigstore, watermarking) - Voice cloning governance and deepfake prevention - Accessibility for avatar-driven interfaces (WCAG, reduced motion) - Performance profiles and rendering tiers - Transparency requirements (distinguishing presentation from execution)
PRIMARY AUDIENCE	- UX designers and frontend engineers building avatar interfaces - Principal engineers selecting avatar rendering and persona systems - Security teams evaluating deepfake and impersonation risks - AI governance, risk, and compliance teams - Standards bodies seeking a reference avatar and provenance 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 21 atomic criteria. Each criterion is independently testable and requires retained evidence.

ID	EVALUATION CRITERION
4.1.1	Event-Driven State
4.1.2	Transparency
4.1.3	Approval Pressure
4.2.1	Independence
4.2.2	Persona Authority Isolation
4.3.1	Generated Media Provenance
4.3.2	Voice Cloning Governance
4.4.1	Lip-Sync Accuracy
4.4.2	Renderer Performance
4.5.1	Reduced Motion Compliance
4.6.1	Avatar Footprint
4.7.1	Avatar Package Security
4.8.1	Consent and Revocation
5.1.1	MABS-State
5.1.2	MABS-Provenance
5.1.3	MABS-Accessibility
6.1.1	Approval Pressure via Animation
6.1.2	Persona Authority Escalation
6.1.3	Deepfake Impersonation
6.1.4	Avatar Package Supply Chain
6.1.5	State Hallucination

MYTHOS-AI-0010-4.1.1

4.1.1 Event-Driven State

Normative source requirement

Is the avatar state driven by authoritative system events, not model-generated text?

Score	Criteria
0	Avatar state is inferred from conversational text (model says "executing" so avatar shows "Executing")
1	Some events drive state, but model output can override
2	Basic event subscription exists, but state transitions are not typed
3	Typed state machine subscribed to authoritative events (the evidence and history service/the human control plane)
4	Full deterministic state: avatar cannot transition state without an authoritative event; model output is ignored for state purposes
5	Perfect deterministic governance: typed state machine, cryptographic event signing, and zero model-inferred state transitions

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-0010-4.1.2

4.1.2 Transparency

Normative source requirement

Does the avatar explicitly distinguish presentation from execution?

Score	Criteria
0	Avatar implies it is the intelligence and the authority
1	Basic disclaimers in documentation
2	Avatar displays active model and agent name
3	Avatar explicitly states: "the multimodal interaction service is presenting; a model is reasoning; the deterministic execution service is executing"
4	Avatar displays: active model, agent, execution state, verification status, and local/remote processing indicator
5	Perfect transparency: real-time display of model identity, agent role, execution boundary, evidence status, and cost; verified by transparency evaluation suite

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-0010-4.1.3

4.1.3 Approval Pressure

Normative source requirement

Does the avatar use animation or urgency to pressure users into approving actions?

Score	Criteria
0	Avatar uses aggressive animation, flashing, or urgency to pressure approval
1	Avatar displays standard approval prompts but with subtle pressure cues
2	Avatar is neutral during approval but does not explicitly calm the user
3	Avatar explicitly states risk, target, and blast radius without pressure
4	Avatar shifts to amber/warning state, explains risk in plain language, and waits patiently
5	Perfect approval governance: no pressure cues, explicit risk communication, hold-to-click approval, and dual-control for destructive actions

---

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-0010-4.2.1

4.2.1 Independence

Normative source requirement

Can persona, avatar, and voice be configured independently?

Score	Criteria
0	All three are hardcoded together (e.g., "the multimodal interaction service" = specific voice + specific 3D model + specific personality)
1	Limited configuration (e.g., can change voice but not avatar)
2	Can change all three but requires code changes
3	All three are configurable via settings
4	All three are file-native contracts (PERSONA.md, AVATAR.md, VOICE.md) with typed manifests
5	Perfect separation: independently swappable, version-controlled, signed packages for each; verified by configuration suite

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-0010-4.2.2

4.2.2 Persona Authority Isolation

Normative source requirement

Can the persona influence tool access, memory scope, or execution authority?

Score	Criteria
0	Persona definition includes tool access and permissions
1	Persona can request elevated permissions
2	Persona is separate but shares state with authority systems
3	Persona is strictly communication behavior; no authority coupling
4	Persona is file-native, signed, and validated against authority isolation rules
5	Perfect isolation: persona cannot grant tools, permissions, secrets, or approval authority; verified by isolation suite

---

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-0010-4.3.1

4.3.1 Generated Media Provenance

Normative source requirement

Does the system cryptographically sign all AI-generated audio, video, and images?

Score	Criteria
0	No provenance tracking
1	Basic metadata (model name, prompt) in file properties
2	Standard metadata + invisible watermark
3	C2PA (Content Authenticity Initiative) compliant manifest
4	C2PA + Ed25519 signature + Sigstore/Rekor transparency log
5	Full provenance: C2PA, cryptographic signature, transparency log, model identity, seed, prompt, persona, and tamper-evident history

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-0010-4.3.2

4.3.2 Voice Cloning Governance

Normative source requirement

Is voice cloning protected against unauthorized impersonation?

Score	Criteria
0	Voice cloning available without consent or verification
1	Basic consent prompt but no verification
2	Consent required but no biometric liveness check
3	Multi-factor verification (voice + hardware key) for voice cloning
4	Voice cloning requires explicit consent, liveness detection, and cryptographic attestation
5	Perfect voice cloning governance: consent, liveness, attestation, C2PA provenance on cloned audio, and revocation capability

---

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-0010-4.4.1

4.4.1 Lip-Sync Accuracy

Normative source requirement

Does the TTS engine produce viseme/phoneme timing data for avatar lip-sync?

Score	Criteria
0	No viseme/phoneme output; lip-sync is random or amplitude-based only
1	Basic viseme output but poorly timed
2	Standard viseme output; acceptable sync but noticeable drift
3	High-precision viseme output; seamless lip-sync
4	Phoneme-level forced alignment; perfect sync across languages
5	Perfect lip-sync: forced alignment, multi-language support, and verified by visual regression suite

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-0010-4.4.2

4.4.2 Renderer Performance

Normative source requirement

Does the avatar renderer maintain smooth frame rates without impacting model inference?

Score	Criteria
0	Avatar rendering consumes >50% of GPU, starving model inference
1	Noticeable stuttering; >30% GPU usage
2	Acceptable but drops frames during model streaming
3	Smooth 30fps; <15% GPU usage
4	Smooth 60fps; <10% GPU usage; WebGL/WebGPU accelerated
5	Perfect performance: 60fps+, <5% GPU, CPU fallback, and reduced-motion profile

---

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-0010-4.5.1

4.5.1 Reduced Motion Compliance

Normative source requirement

Does the avatar respect OS-level reduced-motion preferences?

Score	Criteria
0	No reduced motion support; animations always play
1	Basic reduced motion but avatar still animates
2	Avatar freezes in reduced motion but state indicators are unclear
3	Avatar replaced with static status indicator in reduced motion
4	Full reduced motion: static indicator, text-based state labels, and no decorative animation
5	Perfect accessibility: WCAG AAA, reduced motion, high-contrast, screen-reader semantics, and text alternative for every avatar state

---

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-0010-4.6.1

4.6.1 Avatar Footprint

Normative source requirement

How much RAM/VRAM does the avatar stack consume?

Score	Criteria
0	Requires dedicated GPU; cannot run alongside LLM
1	Requires high-end GPU (>4GB VRAM for avatar alone)
2	Runs on standard GPU (2-4GB VRAM)
3	Runs on low-end GPU (<2GB VRAM) alongside 7B model
4	Runs on CPU with WebGL acceleration
5	Runs on CPU/NPU with <500MB footprint; no GPU required

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-0010-4.7.1

4.7.1 Avatar Package Security

Normative source requirement

Are avatar packages sandboxed and verified?

Score	Criteria
0	Avatar packages can execute arbitrary scripts
1	Basic file scanning but no sandboxing
2	Script execution disabled by default
3	Packages are declarative only (JSON/YAML); no scripts
4	Declarative packages + signed manifests + hash verification
5	Full governance: declarative, signed, sandboxed (WASM if scripts exist), Bastion-reviewed, and revocable

---

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-0010-4.8.1

4.8.1 Consent and Revocation

Normative source requirement

Can voice clones be revoked and is consent tracked?

Score	Criteria
0	No consent tracking; voice clones permanent
1	Basic consent but no revocation
2	Revocation available but not documented
3	Consent recorded with timestamp; revocation documented
4	Consent cryptographically signed; revocation immediate and auditable
5	Perfect governance: signed consent, immediate revocation, C2PA provenance, and audit trail in the evidence and history service

---

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. Trace schemas, immutable event records, integrity verification, retention policy, and operator queries.
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-0010-5.1.1

5.1.1 MABS-State

Normative source requirement

- State Integrity: 100+ tests verifying avatar state matches the evidence and history service events, not model text.
- Transparency: 50+ tests verifying avatar displays correct model, agent, and execution status.

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. Trace schemas, immutable event records, integrity verification, retention policy, and operator queries.
ASSESSMENT METHOD	MATURITY SIGNAL
Exercise crash, retry, duplicate, reorder, partition, and restore scenarios; compare authoritative state before and after replay; confirm idempotency and recovery evidence.	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-0010-5.1.2

5.1.2 MABS-Provenance

Normative source requirement

- Media Signing: 100+ generated audio/video/images verified for C2PA compliance.
- Deepfake Resistance: 50+ tests attempting to generate media without provenance.

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-0010-5.1.3

5.1.3 MABS-Accessibility

Normative source requirement

- Reduced Motion: 50+ tests verifying animation freezes when OS preference is set.
- Screen Reader: 50+ tests verifying ARIA labels for every avatar state.

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.
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-0010-6.1.1

6.1.1 Approval Pressure via Animation

Normative source requirement

Severity: High Description: Avatar uses aggressive animation, flashing colors, or urgency cues to pressure users into approving destructive actions.

Required Controls: 1. Avatar state machine must not allow "urgent" animations for approval states. 2. Approval state must use calm, amber indicators with explicit text. 3. Hold-to-click or dual-control for destructive actions.

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
Exercise crash, retry, duplicate, reorder, partition, and restore scenarios; compare authoritative state before and after replay; confirm idempotency and recovery evidence.	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-0010-6.1.2

6.1.2 Persona Authority Escalation

Normative source requirement

Severity: Critical Description: A persona definition includes tool access or permission grants, allowing the model to escalate privileges via persona modification.

Required Controls: 1. Persona files (PERSONA.md) cannot declare capabilities. 2. Persona validation must reject any file containing tool, permission, or secret references. 3. Authority is enforced by the independent policy engine, not by persona files.

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-0010-6.1.3

6.1.3 Deepfake Impersonation

Normative source requirement

Severity: Critical Description: Malicious actor uses TTS voice cloning to impersonate an authorized user and issue voice commands.

Required Controls: 1. Multi-factor authentication for voice commands (voice + hardware key). 2. Liveness detection for voice biometrics. 3. C2PA provenance for all AI-generated audio. 4. Voice command execution requires independent verification (the independent verification service).

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-0010-6.1.4

6.1.4 Avatar Package Supply Chain

Normative source requirement

Severity: High Description: A malicious avatar package from a marketplace contains executable scripts that exfiltrate data or inject prompt injection payloads.

Required Controls: 1. Avatar packages MUST be declarative (JSON/YAML). No executable scripts. 2. Packages must be signed and hash-verified. 3. Bastion behavioral sandbox review before marketplace publication. 4. Revocation capability for malicious packages.

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-0010-6.1.5

6.1.5 State Hallucination

Normative source requirement

Severity: High Description: Model output ("I'm executing now") causes the avatar to display "Executing" state before the deterministic execution service has actually started execution.

Required Controls: 1. Avatar state MUST be driven exclusively by the evidence and history service/the human control plane authoritative events. 2. Model text output MUST NOT trigger avatar state transitions. 3. State transitions must be typed and validated against the deterministic state machine.

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. Trace schemas, immutable event records, integrity verification, retention policy, and operator queries.
ASSESSMENT METHOD	MATURITY SIGNAL
Exercise crash, retry, duplicate, reorder, partition, and restore scenarios; compare authoritative state before and after replay; confirm idempotency and recovery evidence.	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 evaluation of AI avatars and personas has failed.

Current benchmarks measure whether a 3D model can blink or whether a TTS engine sounds natural. They do not measure whether the avatar's state is driven by authoritative system events or hallucinated by the model. They do not measure whether the persona can be weaponized via prompt injection to pressure users into approving destructive actions. They do not measure whether generated media carries cryptographic provenance to prevent deepfake impersonation.

This standard exists because:

1. Avatars are Not Authority: An avatar is a presentation layer. It is not the model, not the agent, and not the authorization system. An avatar that implies it executed an action when it has not, or that pressures a user into approval through animation, is a manipulation vector.

2. Persona is Not Permission: A persona defines communication behavior (tone, formality, verbosity). It does not define tool access, memory scope, or execution authority. Conflating persona with capability creates privilege escalation paths.

3. State Must Be Deterministic: Avatar state (Listening, Planning, WaitingForApproval, Executing) MUST be driven by authoritative the evidence and history service events, not inferred from conversational text. A model that says "I'm executing now" must not cause the avatar to change state unless the deterministic execution service has actually started execution.

4. Generated Media Requires Provenance: As AI generates audio, video, and images, cryptographic provenance (C2PA) and transparency logs are mandatory to distinguish AI-generated media from human reality. Without provenance, AI-generated evidence is inadmissible and AI-generated communication is indistinguishable from deepfakes.

5. Voice Cloning is a Security Risk: Voice cloning technology can be used to impersonate authorized users. It MUST be governed by explicit consent, multi-factor verification, and liveness detection.

This document establishes the Mythos Avatar, Persona, & Provenance Standard (MAPPS), a comprehensive, evidence-based methodology for evaluating the visual, behavioral, and cryptographic systems that govern AI presentation layers.

Part I. Scope and Applicability

1.1 In Scope

This standard covers the evaluation of:

- Avatar rendering architectures (VRM, Live2D, 3D GLB, 2D canvas, minimal orb)
- Avatar state machines and deterministic event-driven state transitions
- Persona definitions and communication behavior governance
- Voice and avatar independence (configurable separation)
- Lip-sync and viseme generation standards
- Avatar package manifests and marketplace distribution
- AI-generated media provenance (C2PA, Sigstore, watermarking)
- Voice cloning governance and deepfake prevention
- Accessibility for avatar-driven interfaces (WCAG, reduced motion)
- Performance profiles and rendering tiers
- Transparency requirements (distinguishing presentation from execution)

1.2 Out of Scope

- STT/TTS engine evaluation (covered in MYTHOS-AI-0008)
- Agentic framework authority separation (covered in MYTHOS-AI-0001)
- General UI accessibility (covered in MYTHOS-UX-0001)

1.3 Audience

- UX designers and frontend engineers building avatar interfaces
- Principal engineers selecting avatar rendering and persona systems

- Security teams evaluating deepfake and impersonation risks
- AI governance, risk, and compliance teams
- Standards bodies seeking a reference avatar and provenance methodology

Part II. Definitions and Taxonomy

2.1 The Presentation Trinity

Concept	Definition	Grants Authority?
Persona	How the system communicates: tone, formality, verbosity, explanation depth, disagreement behavior, uncertainty language	No
Avatar	How the system looks and moves: renderer, visual identity, animation, expressions, gestures, lip-sync, gaze	No
Voice	How speech is captured and rendered: STT placement, TTS placement, pronunciation, speaking rate, interruption	No

> Governing Rule: Human-readable files (PERSONA.md, AVATAR.md, VOICE.md) describe intent and behavior. They do not grant authority. Authority remains machine-readable, scoped, temporary, revocable, and enforced by policy engines and deterministic executors.

2.2 The Avatar Doctrine

> The avatar makes the governed reference platform understandable and present. > The persona makes it coherent and appropriate. > The voice makes it natural and accessible. > Presentation may be expressive. Authority must remain explicit.

Part III. Evaluation Methodology

3.1 Scoring Model

Each capability is scored from 0 to 5.

Score	Meaning
0	Fails basic task; no governance or deterministic state
1	Basic avatar exists but state is model-generated, not event-driven
2	Functional but lacks persona/voice separation or provenance
3	Solid production performance; deterministic state, separation enforced
4	Strong performance; full provenance, accessibility, and transparency
5	Best-in-class; sets the standard for avatar and persona architecture

Weighting

Category	Weight	Rationale
Deterministic State Machine & Transparency	20%	Model-generated state is a manipulation vector
Persona/Avatar/Voice Separation	15%	Conflation creates privilege escalation and lock-in
Provenance & Deepfake Prevention	15%	Unsigned AI media is a security and compliance liability
Rendering & Lip-Sync Quality	15%	Broken lip-sync or 3D stuttering destroys immersion
Accessibility & Reduced Motion	10%	WCAG compliance is mandatory for production
Performance & Resource Efficiency	10%	Avatar must not consume VRAM needed for model inference

Category	Weight	Rationale
Package Governance & Security	10%	Malicious avatar packages are an attack surface
Voice Cloning Governance	5%	Voice cloning is a biometric security risk

Total: 100%

A system MUST score at least 3.0 in Deterministic State Machine & Transparency to be considered for production use.

Part IV. Evaluation Categories

4.1 Deterministic State Machine and Transparency (Weight: 20%)

4.1.1 Event-Driven State

Is the avatar state driven by authoritative system events, not model-generated text?

Score	Criteria
0	Avatar state is inferred from conversational text (model says "executing" so avatar shows "Executing")
1	Some events drive state, but model output can override
2	Basic event subscription exists, but state transitions are not typed
3	Typed state machine subscribed to authoritative events (the evidence and history service/the human control plane)
4	Full deterministic state: avatar cannot transition state without an authoritative event; model output is ignored for state purposes
5	Perfect deterministic governance: typed state machine, cryptographic event signing, and zero model-inferred state transitions

4.1.2 Transparency

Does the avatar explicitly distinguish presentation from execution?

Score	Criteria
0	Avatar implies it is the intelligence and the authority
1	Basic disclaimers in documentation
2	Avatar displays active model and agent name
3	Avatar explicitly states: "the multimodal interaction service is presenting; a model is reasoning; the deterministic execution service is executing"
4	Avatar displays: active model, agent, execution state, verification status, and local/remote processing indicator
5	Perfect transparency: real-time display of model identity, agent role, execution boundary, evidence status, and cost; verified by transparency evaluation suite

4.1.3 Approval Pressure

Does the avatar use animation or urgency to pressure users into approving actions?

Score	Criteria
0	Avatar uses aggressive animation, flashing, or urgency to pressure approval
1	Avatar displays standard approval prompts but with subtle pressure cues
2	Avatar is neutral during approval but does not explicitly calm the user
3	Avatar explicitly states risk, target, and blast radius without pressure

Score	Criteria
4	Avatar shifts to amber/warning state, explains risk in plain language, and waits patiently
5	Perfect approval governance: no pressure cues, explicit risk communication, hold-to-click approval, and dual-control for destructive actions

4.2 Persona/Avatar/Voice Separation (Weight: 15%)

4.2.1 Independence

Can persona, avatar, and voice be configured independently?

Score	Criteria
0	All three are hardcoded together (e.g., "the multimodal interaction service" = specific voice + specific 3D model + specific personality)
1	Limited configuration (e.g., can change voice but not avatar)
2	Can change all three but requires code changes
3	All three are configurable via settings
4	All three are file-native contracts (PERSONA.md, AVATAR.md, VOICE.md) with typed manifests
5	Perfect separation: independently swappable, version-controlled, signed packages for each; verified by configuration suite

4.2.2 Persona Authority Isolation

Can the persona influence tool access, memory scope, or execution authority?

Score	Criteria
0	Persona definition includes tool access and permissions
1	Persona can request elevated permissions
2	Persona is separate but shares state with authority systems
3	Persona is strictly communication behavior; no authority coupling
4	Persona is file-native, signed, and validated against authority isolation rules
5	Perfect isolation: persona cannot grant tools, permissions, secrets, or approval authority; verified by isolation suite

4.3 Provenance and Deepfake Prevention (Weight: 15%)

4.3.1 Generated Media Provenance

Does the system cryptographically sign all AI-generated audio, video, and images?

Score	Criteria
0	No provenance tracking
1	Basic metadata (model name, prompt) in file properties
2	Standard metadata + invisible watermark
3	C2PA (Content Authenticity Initiative) compliant manifest
4	C2PA + Ed25519 signature + Sigstore/Rekor transparency log
5	Full provenance: C2PA, cryptographic signature, transparency log, model identity, seed, prompt, persona, and tamper-evident history

4.3.2 Voice Cloning Governance

Is voice cloning protected against unauthorized impersonation?

Score	Criteria
0	Voice cloning available without consent or verification
1	Basic consent prompt but no verification
2	Consent required but no biometric liveness check
3	Multi-factor verification (voice + hardware key) for voice cloning
4	Voice cloning requires explicit consent, liveness detection, and cryptographic attestation
5	Perfect voice cloning governance: consent, liveness, attestation, C2PA provenance on cloned audio, and revocation capability

4.4 Rendering and Lip-Sync Quality (Weight: 15%)

4.4.1 Lip-Sync Accuracy

Does the TTS engine produce viseme/phoneme timing data for avatar lip-sync?

Score	Criteria
0	No viseme/phoneme output; lip-sync is random or amplitude-based only
1	Basic viseme output but poorly timed
2	Standard viseme output; acceptable sync but noticeable drift
3	High-precision viseme output; seamless lip-sync
4	Phoneme-level forced alignment; perfect sync across languages
5	Perfect lip-sync: forced alignment, multi-language support, and verified by visual regression suite

4.4.2 Renderer Performance

Does the avatar renderer maintain smooth frame rates without impacting model inference?

Score	Criteria
0	Avatar rendering consumes >50% of GPU, starving model inference
1	Noticeable stuttering; >30% GPU usage
2	Acceptable but drops frames during model streaming
3	Smooth 30fps; <15% GPU usage
4	Smooth 60fps; <10% GPU usage; WebGL/WebGPU accelerated
5	Perfect performance: 60fps+, <5% GPU, CPU fallback, and reduced-motion profile

4.5 Accessibility and Reduced Motion (Weight: 10%)

4.5.1 Reduced Motion Compliance

Does the avatar respect OS-level reduced-motion preferences?

Score	Criteria
0	No reduced motion support; animations always play
1	Basic reduced motion but avatar still animates
2	Avatar freezes in reduced motion but state indicators are unclear
3	Avatar replaced with static status indicator in reduced motion
4	Full reduced motion: static indicator, text-based state labels, and no decorative animation

Score	Criteria
5	Perfect accessibility: WCAG AAA, reduced motion, high-contrast, screen-reader semantics, and text alternative for every avatar state

4.6 Performance and Resource Efficiency (Weight: 10%)

4.6.1 Avatar Footprint

How much RAM/VRAM does the avatar stack consume?

Score	Criteria
0	Requires dedicated GPU; cannot run alongside LLM
1	Requires high-end GPU (>4GB VRAM for avatar alone)
2	Runs on standard GPU (2-4GB VRAM)
3	Runs on low-end GPU (<2GB VRAM) alongside 7B model
4	Runs on CPU with WebGL acceleration
5	Runs on CPU/NPU with <500MB footprint; no GPU required

4.7 Package Governance and Security (Weight: 10%)

4.7.1 Avatar Package Security

Are avatar packages sandboxed and verified?

Score	Criteria
0	Avatar packages can execute arbitrary scripts
1	Basic file scanning but no sandboxing
2	Script execution disabled by default
3	Packages are declarative only (JSON/YAML); no scripts
4	Declarative packages + signed manifests + hash verification
5	Full governance: declarative, signed, sandboxed (WASM if scripts exist), Bastion-reviewed, and revocable

4.8 Voice Cloning Governance (Weight: 5%)

4.8.1 Consent and Revocation

Can voice clones be revoked and is consent tracked?

Score	Criteria
0	No consent tracking; voice clones permanent
1	Basic consent but no revocation
2	Revocation available but not documented
3	Consent recorded with timestamp; revocation documented
4	Consent cryptographically signed; revocation immediate and auditable
5	Perfect governance: signed consent, immediate revocation, C2PA provenance, and audit trail in the evidence and history service

Part V. The Mythos Avatar Benchmark Suite (MABS)

Traditional avatar benchmarks measure visual fidelity. The MABS measures deterministic state, security, and accessibility.

5.1 Suite Composition

5.1.1 MABS-State

- State Integrity: 100+ tests verifying avatar state matches the evidence and history service events, not model text.
- Transparency: 50+ tests verifying avatar displays correct model, agent, and execution status.

5.1.2 MABS-Provenance

- Media Signing: 100+ generated audio/video/images verified for C2PA compliance.
- Deepfake Resistance: 50+ tests attempting to generate media without provenance.

5.1.3 MABS-Accessibility

- Reduced Motion: 50+ tests verifying animation freezes when OS preference is set.
- Screen Reader: 50+ tests verifying ARIA labels for every avatar state.

5.2 Execution Protocol

```
1. Deploy avatar system in isolated environment
2. Run MABS suite (automated)
3. Score each category 0-5
4. Check for state integrity violations (instant disqualification if model can drive avatar state)
5. If all thresholds met: approve for production
```

Part VI. Security Threat Model for Avatars & Personas

6.1 Threat Catalog

6.1.1 Approval Pressure via Animation

Severity: High Description: Avatar uses aggressive animation, flashing colors, or urgency cues to pressure users into approving destructive actions.

Required Controls:

1. Avatar state machine must not allow "urgent" animations for approval states.

2. Approval state must use calm, amber indicators with explicit text.

3. Hold-to-click or dual-control for destructive actions.

6.1.2 Persona Authority Escalation

Severity: Critical Description: A persona definition includes tool access or permission grants, allowing the model to escalate privileges via persona modification.

Required Controls:

1. Persona files (PERSONA.md) cannot declare capabilities.

2. Persona validation must reject any file containing tool, permission, or secret references.

3. Authority is enforced by the independent policy engine, not by persona files.

6.1.3 Deepfake Impersonation

Severity: Critical Description: Malicious actor uses TTS voice cloning to impersonate an authorized user and issue voice commands.

Required Controls:

1. Multi-factor authentication for voice commands (voice + hardware key).

2. Liveness detection for voice biometrics.

3. C2PA provenance for all AI-generated audio.

4. Voice command execution requires independent verification (the independent verification service).

6.1.4 Avatar Package Supply Chain

Severity: High Description: A malicious avatar package from a marketplace contains executable scripts that exfiltrate data or inject prompt injection payloads.

Required Controls:

1. Avatar packages MUST be declarative (JSON/YAML). No executable scripts.

2. Packages must be signed and hash-verified.

3. Bastion behavioral sandbox review before marketplace publication.

4. Revocation capability for malicious packages.

6.1.5 State Hallucination

Severity: High Description: Model output ("I'm executing now") causes the avatar to display "Executing" state before the deterministic execution service has actually started execution.

Required Controls:

1. Avatar state MUST be driven exclusively by the evidence and history service/the human control plane authoritative events.
2. Model text output MUST NOT trigger avatar state transitions.
3. State transitions must be typed and validated against the deterministic state machine.

Part VII. The Mythos Avatar & Persona Standard

7.1 The Mythos Avatar Doctrine

```
The avatar makes the governed reference platform understandable and present.
The persona makes it coherent and appropriate.
The voice makes it natural and accessible.

Presentation may be expressive.
Authority must remain explicit.

An avatar that pressures is a manipulation vector.
A persona that grants authority is a privilege escalation path.
A generated media without provenance is a deepfake.
```

7.2 Selection Rules

1. No avatar system enters production without passing MABS.
2. No avatar is approved if its state can be driven by model text output.
3. No persona is approved if it can influence tool access or permissions.
4. No generated media is distributed without C2PA provenance.
5. No voice cloning is approved without consent, liveness detection, and revocation capability.

7.3 The Prime Directive for Avatars & Personas

> Drive the state with events, not text. > Separate the persona from the authority. > Sign the media, not just the message. > Test the accessibility, not just the fidelity.

```
Academic benchmarks measure visual fidelity.
Applied benchmarks measure state integrity.
Security benchmarks measure deepfake prevention.
Operational benchmarks measure accessibility.

> Presentation may be expressive.
> Authority must remain explicit.
```

End of Standard MYTHOS-AI-0010

End of controlled publication

MYTHOS-AI-0010 remains a Candidate Standard until technical review, security and privacy review, accessibility review, current-source validation, and formal approval are recorded.