Implementation Guidance for Module MRIS.bd.1 Mythos-Resistant Hardening of the ISMS

Version 1.2 | Date: 2026-08-14 | Author: Richard Peddi

1 Introduction

This implementation guidance substantiates the requirements of module MRIS.bd.1 Mythos-Resistant Hardening of the ISMS. It is based on the "Mythos-Resistant Information Security" framework [MRIS] and the accompanying Implementation Guide [MRISIG]. "Mythos" denotes the class of agentic frontier AI models with autonomous vulnerability discovery and exploit chain building, independent of any vendor.

The safeguards follow a common principle: protective effect is to come from hard structural barriers (cryptographic impossibility, immutability, verifiable separation, limited privileges), not from attacker friction, which no longer holds under AI conditions. Each safeguard therefore states, in addition to the implementation, the measurement of effectiveness and typical mistakes. Products named in the safeguards are non-binding examples, not recommendations.

2 Safeguards

Specific safeguards for the requirements of module MRIS.bd.1 Mythos-Resistant Hardening of the ISMS are listed below.

All safeguards (marked with M) are numbered in ascending order and correspond to the respective requirements (marked with A).

2.1 Safeguards for Module MRIS.bd.1 Mythos-Resistant Hardening of the ISMS

MRIS.bd.1.M1 Reassessing the Effectiveness Assumptions of Existing Security Safeguards (B)

The reassessment is based on the MRIS four-category grid with the guiding question of whether the protective effect of a safeguard stems from a hard structural barrier or merely from attacker friction: robust (29 controls of ISO/IEC 27002:2022), partially degraded (37), friction-only (4), and not affected (23). The MRIS assessment matrix (Annex A, additionally available as a machine-readable YAML edition) can be adopted as a starting point and adapted to the organisation’s own context; the classification is context-dependent and not a universally valid determination [MRIS].

The approach prioritises the safeguards classified as friction-only and as partially degraded for immediate reassessment. The results feed into the risk analysis — in particular the shift of likelihood caused by the disappearance of the capacity barrier and the collapsed time parameters (see the module, chapter 2) — as well as into the Statement of Applicability. For each degraded safeguard, a documented decision is taken: hardening (via safeguards M2 to M14), compensation, or justified risk acceptance. The correct reading is "selectively degraded", not "obsolete": classic safeguards lose effect in individual effectiveness dimensions while retaining their full protective effect in others — the correct response is hardening, not replacement.

The reassessment is repeated on an event-driven basis, in particular upon significant changes to the threat landscape (new model capability classes, disclosure waves, changed attack patterns).

Measurement: share of reassessed Annex A safeguards (target 100 %) and documented decisions per degraded safeguard. Typical mistake: the reassessment is treated as a one-off project — without defined repetition triggers, it becomes outdated with the next capability level of the models.

MRIS.bd.1.M2 Using Phishing-Resistant Multi-Factor Authentication (B)

Organisationally, the authentication policy is framed so that phishing-resistant mechanisms based on FIDO2/WebAuthn are mandatory for privileged, externally reachable, and particularly sensitive access, and the permitted mechanisms are defined per access class. For existing legacy MFA mechanisms (SMS, simple push confirmation), a phase-out plan with deadlines is drawn up.

Technically, passkeys or hardware tokens are rolled out as the minimum form. Effectiveness rests on the origin binding of WebAuthn: an authentication credential captured on a phishing site is worthless for the genuine service, regardless of the quality of AI-generated phishing content. Particular attention is paid to the enrolment and recovery processes: these edge processes are the preferred bypass path and are protected against social engineering (identity verification, four-eyes principle for privileged accounts). It should also be noted that compromise of the endpoint after successful sign-in is not prevented by the mechanism; session protection and endpoint hardening remain separate tasks.

Measurement: share of sign-ins via phishing-resistant mechanisms, broken down by access class; the target is full coverage of privileged and externally reachable access. Typical mistakes: the mechanism is introduced, but SMS remains active as a fallback; the recovery of new authentication means runs through an unprotected helpdesk process.

MRIS.bd.1.M3 Maintaining Immutable Backups with Regular Restore Tests (B)

The backup regime is built according to the 3-2-1-1-0 model: three copies on two media types, at least one copy off-site and at least one immutable (immutability/object lock) or offline, zero errors in the periodic restore test. The backup infrastructure is given dedicated IAM paths separated from the production environment so that compromised production administrator accounts cannot delete or alter the backups.

Restore tests are performed and documented at least quarterly; what is tested is the actual restorability of defined business-critical data sets, not merely the existence of the backup. Immutable backups are one of the most effective hard barriers against ransomware and agentic data integrity attacks, because their protective effect does not depend on detection speed.

Measurement: success rate of the quarterly restore tests (target 100 %); evidence via test records, retention and lock settings, and the separated assignment of privileges. Typical mistakes: immutability is configurable and can be switched off again with the same administrator privileges; restore tests only cover individual files instead of complete systems.

MRIS.bd.1.M4 Establishing Behaviour-Based Attack Detection with Event Correlation (B)

Detection is shifted from signature-oriented single alerts to behaviour-based patterns and kill-chain correlation across multiple events and time windows. The basis is MITRE ATT&CK-based detection rules with measurable coverage; as a guideline, coverage of at least 60 % of the ten most relevant techniques of the organisation’s own sector applies, measured for example with DeTT&CT and confirmed through attack simulations (e.g. Atomic Red Team) [ATTACK]. Detection priorities against AI-typical approaches are living-off-the-land activity (PowerShell encoding, unusual process chains from office applications, conspicuous curl/wget invocations) and camouflaged command-and-control channels (DNS tunnelling, HTTPS with long sleep intervals, abused cloud APIs).

Threat hunting is established as a distinct function with a documented methodology (e.g. PEAK or TaHiTI); as guidelines, 0.5 FTE per 500 employees and at least twelve hypothesis-driven hunts per year with ATT&CK mapping apply. For organisations without an in-house SOC, managed detection and response services are a valid alternative.

Measurement: measured and test-confirmed ATT&CK coverage as well as number and outcome of the hunts. Note: detection is not prevention. Low-and-slow attacks above the correlation time windows and well-camouflaged agents within legitimate behavioural baselines remain a residual risk; the safeguard only works in combination with the structural barriers (M2, M3, M5). Typical mistake: tools are procured, but no coverage is measured — the detection gaps remain unknown.

MRIS.bd.1.M5 Introducing Workload Identities and Identity-Based Access Control (B)

The migration proceeds in three stages instead of a big-bang migration. Stage 1 (privileged workloads): cryptographically verifiable, short-lived identities (e.g. SPIFFE/SPIRE) for service accounts with extended privileges, plus mTLS on the three to five most critical service-to-service paths. Stage 2 (production mainstream): workload identity for all production microservices and Zero Trust Network Access (ZTNA) for privileged remote access. Stage 3 (full coverage): extension to development and test environments plus identity-based micro-segmentation.

AI agents are a special case: coding agents and agentic workflows receive their own workload identities with time-limited privileges tailored to the respective capability per tool invocation — no personal developer credentials (see M7). Long-lived shared secrets stored in plain text are eliminated on critical paths; they are the classic fuel for lateral movement.

Measurement: share of production service-to-service connections with workload identity and mTLS (guidelines: stage 1 at least 80 % of privileged service workloads, stage 2 mandatory mTLS for at least 95 % of east-west traffic) as well as the time to revoke an identity. Typical mistakes: mTLS is introduced, but authorisation continues to rely on IP addresses; the identities are long-lived and never rotated.

MRIS.bd.1.M6 Creating a Cryptographic Inventory and a Post-Quantum Migration Strategy (B)

First, a central inventory of all algorithms, key lengths, protocols, certificates, and libraries in use is created and prioritised according to the protection needs and confidentiality lifetime of the data. On this basis, a migration strategy towards quantum-safe mechanisms is documented, with defined triggers, resources, transition paths, and success criteria. The threat reference is the "harvest now, decrypt later" pattern: today’s data exfiltration determines tomorrow’s decryption risk, which is why data with a long confidentiality lifetime is migrated first.

Technically, migration targets the finalised NIST standards FIPS 203 (ML-KEM) for key establishment and FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA) for signatures [FIPS]; mechanism and key length recommendations are provided by BSI TR-02102 [TR02102], and the migration approach by NIST SP 1800-38 [SP180038]. During the transition, hybrid mechanisms combining classical and quantum-safe cryptography are used. Crypto agility is anchored structurally: encryption is encapsulated so that mechanisms can be replaced without deep changes to the applications. Organisations without in-house development govern the topic via inventory overview and procurement by querying their suppliers’ migration roadmaps and contractually requiring quantum-safe mechanisms.

Measurement: share of mechanisms captured and classified by migration priority; tolerance limit: no long-lived confidential data without a migration plan. Typical mistakes: a strategy without a complete inventory; blanket hybrid roll-out without prioritisation by confidentiality lifetime; encryption hard-wired into code, turning every mechanism change into an application project.

MRIS.bd.1.M7 Regulated Use of AI Agents (B)

The governance of production AI agents covers five dimensions:

- Harness audit: prompts, tool definitions, retrieval pipelines, and escalation logic are treated with the same code review discipline as production code — versioning in the source repository, signed releases, tests against prompt injection and confused deputy attacks before go-live.
- Blast radius limitation: privileges per tool invocation with an explicitly defined scope, rate limits per agent identity, circuit breakers on unusual action sequences, and binding human approval thresholds for irreversible actions (production deployment, data deletion, spending above defined limits).
- Human control: every agent session has a named responsible person with the ability to shut it down; all actions are logged with correlation to the initiating human identity, retained for at least twelve months.
- Supply chain inventory of agentic components: MCP servers (source, update channel), IDE extensions with AI functionality (allow list, controlled updates), skills and rules files (versioning, code review), and external agent frameworks.
- Pre-production review: documented scope definition, threat model for the use case, rollback plan, and monitoring plan with defined anomaly thresholds.

In the first phase, inventorying, logging, and privilege limitation are prioritised — these three elements already address the largest share of the risk; full harness penetration tests follow depending on tool availability. Measurement: share of production-adjacent agents with a documented threat model and defined blast radius (target 100 %), logging coverage for actions with write or network privileges (target 100 %), time to revoke compromised agent identities (target under five minutes). Typical mistake: agents run under the personal user accounts of developers — no action can then be limited or cleanly attributed.

MRIS.bd.1.M8 Generating Software Bills of Materials and Build Provenance Evidence (S)

SBOM generation is integrated automatically into the build/CI pipeline, in an established machine-readable format — CycloneDX (ECMA-424) or SPDX (ISO/IEC 5962) — including transitive dependencies. Each software version receives its own SBOM, retained under version control. Components are correlated automatically against vulnerability databases (CVE/NVD, OSV, EUVD); vulnerability information is communicated separately via VEX/CSAF and not mixed into the static SBOM. For critical artefacts, tamper-proof build provenance following the SLSA framework is additionally maintained [SLSA]; quality requirements for SBOMs are provided by BSI TR-03183-2 [TR03183], and the legal framework by the Cyber Resilience Act [CRA].

The effectiveness test is concrete: upon a newly disclosed vulnerability in a widespread library, it must be determinable within minutes, on the basis of the SBOMs alone, which of the organisation's own artefacts are affected. The provision of up-to-date SBOMs is contractually required from critical software suppliers; organisations without in-house development apply the principle entirely through procurement.

Measurement: SBOM coverage — share of shipped artefacts with a current, automatically generated SBOM (target: full coverage). Typical mistakes: SBOMs are generated but never evaluated; only top-level dependencies are captured; the SBOM is not maintained per version and does not match the shipped state.

MRIS.bd.1.M9 Integrating Automated Security Testing into the Development Pipeline (S)

The pipeline combines three test classes: static code analysis (SAST, e.g. Semgrep, SonarQube), dynamic testing of the running application (DAST, e.g. OWASP ZAP), and dependency analysis (SCA, e.g. Dependency-Track, Snyk), complemented by AI-assisted code scanning. Findings rated High or Critical with high confidence block the merge into the main branch (pre-merge block); vulnerabilities from the catalogue of known exploited vulnerabilities (KEV) block immediately. Running systems are scanned daily.

AI-generated code is treated as a distinct risk category: AI coding assistants regularly produce insecure patterns such as hard-coded credentials, missing input validation, insecure defaults, and outdated dependencies. Pre-commit checks with rule sets against these AI-typical anti-patterns and regular audit reporting with trend tracking address this. The shift-left principle is the structural answer to the collapse of the patch-exploit window (see the module, chapter 2.1): it removes the attacker's time window between exploit availability and production patch.

Measurement: time from disclosure of an actively exploited vulnerability to the rolled-out patch (guideline for internet-exposed systems: under 24 hours) and share of pipelines with security testing and pre-merge block. Typical mistake: the tests run, but blocks are routinely bypassed via exceptions — the control exists only on paper.

MRIS.bd.1.M10 Automating Incident Response with Hardened Playbooks (S)

For unambiguous incident types, predefined playbooks are stored on a SOAR platform that initiate containment actions automatically; human escalation takes place in a targeted manner for ambiguous cases. AI-specific playbooks cover at least: mass data exfiltration, parallel multi-vector attacks, supply chain compromise, credential stuffing waves, and MFA fatigue campaigns. Exercises for three to five parallel incidents take place at least every six months, since agentic attackers deliberately use parallelism as an overload tactic.

The automation layer itself is treated as an attack surface and hardened: signed, authenticated triggers, a complete audit trail of every playbook execution with the executing identity, rate limits per playbook, human-in-the-loop approval for actions with a large blast radius (mass account lockout, isolation of larger network segments, certificate revocation), and protection of the playbook definitions like source code. Without this hardening, the response automation can be turned against the organisation, for example through provoked false alarms.

For organisations without an in-house security team, managed detection and response services with 24/7 operation and a contractually agreed containment time are the more viable alternative to operating SOAR in-house. Measurement: time to containment of unambiguous incidents (guideline: under ten minutes at high unambiguity). Typical mistake: automation is introduced before detection (M4) is stable — the playbooks then react to unreliable signals.

MRIS.bd.1.M11 Establishing Continuous Automated Effectiveness Verification (S)

Security requirements are, where possible, formalised as executable rules (policy as code, e.g. with OPA or comparable engines) and integrated twice: as checks before deployment (pre-deploy checks in CI) and as enforcement during operation. Compliance with the target states is verified continuously and at least daily; deviations automatically trigger workflows (e.g. incident tickets) and become visible in dashboards with quantified deviation indicators.

The core of the safeguard is reducing the detection latency for control deviations from months (periodic audit) to minutes. Under AI conditions this is decisive, because a configuration that has drifted unnoticed can already be exploited in the time until the next audit. The safeguard does not replace periodic audits, but shifts their focus to verifying the quality of the rules.

Measurement: continuous monitoring coverage — share of the safeguards listed in the Statement of Applicability with active automated verification; guidelines: at least 60 % after twelve months, at

least 80 % after 24 months. Essential safeguards without automated verification are only permissible temporarily and with justification. Typical mistake: dashboards are built without deviations triggering a binding response.

MRIS.bd.1.M12 Using Signed Container Images and Protected Execution Environments (H)

For containers, the following applies: base images originate exclusively from controlled registries, are signed (e.g. Sigstore/cosign), and are scanned automatically before deployment. Enforcement takes place technically via admission controller policies of the platform (e.g. OPA Gatekeeper, Kyverno), so that unsigned or unverified images do not start in the first place. The signature creates a cryptographic hard barrier against manipulated images in the supply chain.

For workloads with particularly high protection needs, protected execution environments are used (trusted execution environments, e.g. Intel TDX, AMD SEV-SNP, Arm CCA), whose authenticity is demonstrated via remote attestation. Confidential computing is currently the hardest available barrier against attackers with access to the operator's infrastructure and is therefore particularly relevant for regulated workloads and highly sensitive data processing.

Measurement: share of production containers from signed, verified images (target: full coverage) and the confidential computing coverage of workloads classified as highly sensitive; tolerance limit: no production container without a valid signature and verification. Typical mistake: signature verification in the admission controller is configured in audit mode and does not block.

MRIS.bd.1.M13 Implementing Verifiable Tenant Separation (H)

Separation is documented and implemented on three levels: for data (tenant-specific keys, logical or physical separation), for networks (tenant-specific VPCs, namespaces, software-defined networking rules), and for compute resources (tenant-specific scheduling policies). The decisive step is the transition from conceptual assertion to evidence: the effectiveness of the separation is demonstrated through regular, preferably automated separation tests.

As guidelines for the test tiers, the following apply: annually internal, automated per major release, complemented by an assessment by independent third parties. The safeguard structurally limits the blast radius: an attacker who gains a foothold in one tenant is demonstrably prevented from accessing other tenants — regardless of how fast the attacker operates.

Measurement: outcome of the separation tests (no cross-tenant access demonstrable) as well as frequency and coverage of the tests; evidence via separation concepts, test records, and remediation records. Typical mistake: separation is implemented only at the application level (tenant ID in code) — a single application error then removes it entirely.

MRIS.bd.1.M14 Conducting Threat-Led Penetration Tests with AI Scenarios (H)

Threat-led penetration tests are conducted with external red teams that replicate realistic scenarios tailored to the current threat landscape. AI-specific scenarios include at least: AI-assisted spear phishing including deepfake voice, fragmented attack chains via micro-steps, parallel multi-vector attacks, supply chain compromise, and cloud API abuse via compromised service accounts. The scenario pool is designed multi-systemically (at least two independent agentic systems) and not tied to a single model. Critical functions are tested at least annually; between the cycles, purple team exercises with ATT&CK coverage mapping take place.

Full-scope tests following the TIBER-EU methodology typically cost in the mid six-figure range per exercise. For organisations outside the DORA scope, more affordable formats are sufficiently effective: purple team exercises with MITRE ATT&CK scenarios, open-source adversary emulation (e.g. Caldera, Atomic Red Team), and breach-and-attack simulation platforms.

The core of the safeguard: what is tested is whether the protective safeguards work against real AI attack approaches — not whether they are documented. This reduces the risk of false-positive compliance assurances. Measurement: closure rate — share of identified weaknesses remediated within the agreed time frame (guideline: at least 90 %). Typical mistake: test reports are filed without remediation being followed up.

3 Additional Information

3.1 Useful Information

Origin and mapping of the requirements. The requirements of this module operationalise the thirteen Mythos Hardening Controls (MHC) of the MRIS framework; A1 additionally implements the MRIS recommendation on immediate reassessment [MRIS]. For adoption into the Statement of Applicability and reconciliation with ISO/IEC 27001 Annex A, the following mapping applies: A1 corresponds to the MRIS reassessment (chapter 10.2, entire Annex A); A2 corresponds to MHC-03 and flanks A.5.17 and A.8.5; A3 corresponds to MHC-08 and deepens A.8.13; A4 corresponds to MHC-05 and flanks A.8.16 and A.8.12; A5 corresponds to MHC-04 and flanks A.5.16, A.8.20, and A.8.21; A6 corresponds to MHC-01 and flanks A.8.24; A7 corresponds to MHC-13 and extends A.5.9, A.5.16, and A.8.27; A8 corresponds to MHC-02 and flanks A.5.21 and A.8.30; A9 corresponds to MHC-09 and flanks A.8.29 and A.8.25; A10 corresponds to MHC-11 and complements A.5.24 to A.5.26; A11 corresponds to MHC-10 and supports A.5.35 and A.5.36; A12 corresponds to MHC-06 and flanks A.5.23 and A.8.31; A13 corresponds to MHC-07 and flanks A.5.23 and A.8.22; A14 corresponds to MHC-12 and complements A.5.35 and A.8.29.

Implementation sequence. The assignment of the requirements to basic, standard, and increased protection needs follows the MRIS standard prioritisation P0/P1/P2 for typical enterprise environments. It is a starting point, not a universal plan: with long-lived confidential data, A6 moves forward; with production use of AI agents, A7; on container or multi-tenant platforms, A12 and A13 gain urgency. The following dependency rules apply in particular: reassessment (A1) before investing in new safeguards; detection (A4) before response automation (A10); SBOMs (A8) as the basis for pipeline testing (A9); validation through tests (A14) after the basic implementation [MRISIG].

Maturity levels. The Implementation Guide describes for each safeguard a cumulative maturity path in three levels (Initial, Defined, Managed) with stage gates and key figures. These levels are suitable for internal progress measurement and for reporting to top management [MRISIG].

Integration into the ISMS. Every requirement of this module is formulated so that it can be adopted into the Statement of Applicability of an existing ISO/IEC 27001-certified ISMS. Requirements that are not applicable are excluded there with justification — typical cases: A8 and A9 for organisations without in-house software development (the principle then works through procurement, see M8), A13 without multi-tenant operation. Statutory reporting obligations under NIS2 and DORA are deliberately not included as separate requirements: they are direct legal obligations and are implemented in the ISMS via incident handling; A10 provides the necessary response capability.

3.2 References

[MRIS] Mythos-Resistant Information Security (MRIS) — Framework, Version 1.9 (German edition), Richard Peddi, August 2026, CC BY-SA 4.0.

[MRISIG] MRIS Implementation Guide, Version 1.5 (German edition), Richard Peddi, August 2026, CC BY-SA 4.0.

[ATTACK] MITRE ATT&CK Knowledge Base, The MITRE Corporation, continuously updated, https://attack.mitre.org, last accessed on 2026-08-14.

[FIPS] FIPS 203 (ML-KEM), FIPS 204 (ML-DSA), FIPS 205 (SLH-DSA), NIST, August 2024.

[TR02102] Technical Guideline TR-02102: Cryptographic Mechanisms — Recommendations and Key Lengths, BSI, current edition.

[SP180038] NIST SP 1800-38: Migration to Post-Quantum Cryptography, NIST NCCoE.

[TR03183] Technical Guideline TR-03183-2: Cyber Resilience Requirements — Software Bill of Materials, BSI.

[SLSA] Supply-chain Levels for Software Artifacts (SLSA), OpenSSF, https://slsa.dev, last accessed on 2026-08-14.

[CRA] Regulation (EU) 2024/2847 on horizontal cybersecurity requirements (Cyber Resilience Act), Annex I.