Implementation guidance for the module: MRIS.bd.1 Mythos-resistant hardening of the ISMS

Version 1.3 | Status: 12 September 2026 | Author: Richard Peddi

Contents

Contents 2
1 Introduction 3
2 Measures 3
2.1 Measures for the module MRIS.bd.1 Mythos-resistant hardening of the ISMS 3
MRIS.bd.1.M1 Reassessment of the effectiveness assumptions of existing security measures (B) 3
MRIS.bd.1.M2 Use of phishing-resistant multi-factor authentication (B) 3
MRIS.bd.1.M3 Maintaining immutable data backups with regular recovery tests (B) 4
MRIS.bd.1.M4 Establishing behaviour-based attack detection with event correlation (B) 4
MRIS.bd.1.M5 Introduction of workload identities and identity-based access control (B) 4
MRIS.bd.1.M6 Producing a cryptographic inventory and a post-quantum migration strategy (B) 5
MRIS.bd.1.M7 Governed use of AI agents (B) 5
MRIS.bd.1.M8 Generating software bills of materials and build provenance evidence (S) 5
MRIS.bd.1.M9 Integration of automated security testing into the development pipeline (S) 6
MRIS.bd.1.M10 Automation of incident response with secured playbooks (S) 6
MRIS.bd.1.M11 Establishing continuous automated effectiveness checking (S) 6
MRIS.bd.1.M12 Use of signed container images and protected execution environments (H) 7
MRIS.bd.1.M13 Implementation of demonstrable tenant separation (H) 7
MRIS.bd.1.M14 Carrying out threat-led penetration tests with AI scenarios (H) 7
3 Further information 8
3.1 Useful background 8
3.2 References 8

1 Introduction

This implementation guidance concretizes the requirements of the module MRIS.bd.1 Mythos-resistant hardening of the ISMS. Its basis is the framework "Mythos-Resistant Information Security" [MRIS] and the accompanying Implementation Guide [MRISIG]. "Mythos" here denotes the class of agentic frontier AI models with autonomous vulnerability discovery and exploit chaining, irrespective of vendor.

The measures follow a common underlying principle: protective effect should derive from hard structural barriers (cryptographic impossibility, immutability, demonstrable separation, limited entitlements), not from attacker friction, which no longer holds under AI conditions. Alongside implementation, every measure therefore also names the measurement of effectiveness and typical mistakes. The products named in the measures are non-binding examples and not a recommendation.

2 Measures

Specific measures for the requirements of the module MRIS.bd.1 Mythos-resistant hardening of the ISMS are set out below.

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

2.1 Measures for the module MRIS.bd.1 Mythos-resistant hardening of the ISMS

MRIS.bd.1.M1 Reassessment of the effectiveness assumptions of existing security measures (B)

The basis of the reassessment is the MRIS four-category grid, with the guiding question whether the protective effect of a measure derives from a hard structural barrier or only 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 institution's own context; the classification is context-dependent and not a universally applicable determination [MRIS].

The approach prioritizes the measures classified as friction-only and as partially degraded for immediate reassessment. The results feed into the risk analysis — in particular the shift in likelihood of occurrence through the disappearance of the capacity barrier and the collapsed time parameters (see the module, Chapter 2) — and into the Statement of Applicability. For each degraded measure, a documented decision is taken: hardening (through measures M2 to M14), compensation or a justified acceptance of the risk. What matters is the reading "selectively degraded", not "obsolete": classic measures 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 on material changes in the threat landscape (new model capability classes, disclosure waves, changed attack patterns). Measurement: share of the Annex A measures reassessed (target 100%) and documented decisions per degraded measure. Typical mistake: the reassessment is treated as a one-off project — without defined repetition triggers it becomes out of date with the next capability level of the models.

MRIS.bd.1.M2 Use of phishing-resistant multi-factor authentication (B)

Organizationally, the authentication policy is framed so that phishing-resistant methods to FIDO2/WebAuthn (passkeys or hardware tokens) are mandatory for privileged, externally reachable and particularly protection-worthy access, and so that the permitted methods are established per access class. For existing legacy MFA methods (SMS, simple push confirmation), a replacement plan with a target date is drawn up.

Technically, passkeys or hardware tokens are rolled out as the minimum form. The effectiveness rests on the origin binding of WebAuthn: an authentication proof captured on a phishing page is worthless for the genuine service, irrespective of the quality of AI-generated phishing content. Particular attention is due to the registration and recovery processes: these peripheral processes are the preferred circumvention path and are secured against social engineering (identity verification, four-eyes principle for privileged accounts). It should further be noted that compromise of the endpoint after successful sign-in is not prevented by the method; session protection and endpoint hardening remain tasks in their own right.

Measurement: share of logins using phishing-resistant methods, separated by access class; the target is full coverage of the privileged and externally reachable access. Typical mistakes: the method is introduced but SMS remains active as a fallback; the recovery of new authentication means runs through an unsecured help desk process.

MRIS.bd.1.M3 Maintaining immutable data backups with regular recovery tests (B)

Data backup is built on the 3-2-1-1-0 model: three copies on two media types, at least one copy offsite and at least one immutable (immutability/object lock) or offline, zero errors in the periodic recovery test. The backup infrastructure receives its own IAM paths, separate from the production environment, so that compromised production administrator accounts cannot delete or alter the backups.

Recovery tests are carried out and documented at least quarterly; what is tested is the actual restorability of defined business-critical data holdings, not merely the existence of the backup. If a recovery test fails, the cause is remedied without delay and the affected recovery path is tested again. Immutable backups are among the most effective hard barriers against ransomware and agentic data integrity attacks, because their protective effect does not depend on the speed of detection.

Measurement: success rate of the quarterly recovery tests (target 100%); evidence through test records, retention and lock settings, and the separate assignment of entitlements. Typical mistakes: immutability is configurable and can be switched off again with the same administrator rights; recovery tests examine only individual files instead of complete systems.

MRIS.bd.1.M4 Establishing behaviour-based attack detection with event correlation (B)

Detection is moved from signature-oriented individual alerts to behaviour-based patterns and kill chain correlation across several events and time windows. The basis is MITRE ATT&CK-based detection rules with measurable coverage; as a guide value, coverage of at least 60 per cent of the ten most important techniques of the institution's own sector applies, measured for example with DeTT&CT and confirmed through attack simulations (for example 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 and wget calls) and disguised command-and-control channels (DNS tunnelling, HTTPS with long sleep intervals, abused cloud APIs).

Threat hunting is established as a function in its own right with a documented methodology (for example PEAK or TaHiTI); as guide values, 0.5 FTE up to 500 employees and at least twelve hypothesis-based hunts per year with ATT&CK mapping apply. For institutions without a SOC of their own, managed detection and response services are a valid alternative.

Measurement: measured ATT&CK coverage confirmed by testing, together with the number and outcome of the hunts. To be noted: detection is not prevention. Low-and-slow attacks above the correlation time windows and well-disguised agents within legitimate behavioural baselines remain residual risk; the measure takes effect only 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 Introduction of workload identities and identity-based access control (B)

The migration takes place in three stages rather than as a big-bang transition. Stage 1 (privileged workloads): cryptographically verifiable, short-lived identities (for example SPIFFE/SPIRE) for service accounts with extended entitlements, together with 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 and identity-based micro-segmentation.

AI agents are a special case: coding agents and agentic workflows receive their own workload identities with time-limited entitlements scoped to the respective capability per tool call — no personal developer credentials (see M7). Long-lived shared secrets stored in cleartext 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 (guide values: stage 1 at least 80 per cent of the privileged service workloads, stage 2 mTLS mandatory for at least 95 per cent of east-west traffic), together with the time to withdrawal of an identity. Typical mistakes: mTLS is

introduced but authorization continues to take place through IP addresses; the identities are long-lived and are never rotated.

MRIS.bd.1.M6 Producing 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 produced and prioritized by protection requirement and the confidentiality period of the data. On that basis, a migration strategy towards quantum-safe methods is documented, with defined triggers, resources, transition paths and success criteria. The threat relevance is the "harvest now, decrypt later" pattern: today's data outflow determines tomorrow's decryption risk, which is why data with a long confidentiality period are migrated first.

Technically, the migration is to the finalized NIST standards FIPS 203 (ML-KEM) for key exchange and FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA) for signatures [FIPS]; method and key length recommendations are contained in BSI TR-02102 [TR02102], and the migration approach in NIST SP 1800-38 [SP180038]. During the transition, hybrid methods are used that combine classical and quantum-safe cryptography. Crypto agility is anchored structurally: encryption is encapsulated so that methods are exchangeable without deep intervention in the applications. Institutions without development of their own manage the topic through inventory overview and procurement, by asking their providers for their migration roadmap and requiring quantum-safe methods contractually.

Measurement: share of the methods recorded 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 introduction of hybrid methods without prioritization by confidentiality period; encryption hard-wired into the code, which turns every change of method into an application project.

MRIS.bd.1.M7 Governed use of AI agents (B)

The governance of production AI agents covers five dimensions:

- Harness audit for components under the organization's own control: prompts, tool definitions, retrieval pipelines and escalation logic are treated with the same code review discipline as production code. For externally operated agents, vendor evidence, allow lists for connectors and extensions, and configuration controls take their place — versioning in the source repository, signed releases, tests against prompt injection and confused deputy attacks before production deployment.
- Blast radius limitation: entitlements per tool call with an explicitly defined scope, rate limits per agent identity, a circuit breaker on unusual action sequences, and binding human approval thresholds for irreversible actions (production deployment, data deletion, expenditure 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 triggering human identity. The retention period is established and documented on a risk basis and having regard to legal, contractual and forensic requirements; agent logs may contain personal data, prompts, credentials or source code.
- 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.
- Checking before production deployment: a documented scope definition, a threat model for the use case, a rollback plan and a monitoring plan with defined anomaly thresholds.

In the first phase, inventory, logging and entitlement limitation are prioritized — these three elements already address the bulk of the risk; full harness penetration tests follow depending on tool availability. Measurement: share of the production AI agents and agentic workflows — including SaaS, low-code and no-code agents — with a complete inventory entry (inventory coverage, target 100%) and, collected separately, their share with a documented threat model, a named responsible person, a limited identity and permission boundaries (governance coverage, target 100%); logging coverage for actions with write or network entitlements (target 100%); time to withdrawal of compromised agent identities (target below five minutes). Typical mistake: agents run under the personal user accounts of the developers — every action is then neither capable of limitation nor cleanly attributable.

MRIS.bd.1.M8 Generating software bills of materials and build provenance evidence (S)

Bill of materials generation is integrated automatically into the build and CI pipeline, in an established machine-readable format — CycloneDX (ECMA-424) or SPDX (ISO/IEC 5962) — including the transitive

dependencies. A separate bill of materials, retained under version control, arises for each software version. The components are correlated in automated form against vulnerability databases (CVE/NVD, OSV, EUVD); vulnerability information is communicated separately via VEX/CSAF and is not mixed into the static bill of materials. For critical artefacts, tamper-resistant build provenance to the SLSA framework is additionally maintained [SLSA]; quality requirements for bills of materials are contained in BSI TR-03183-2 [TR03183], and the legal framework in the Cyber Resilience Act [CRA].

The effectiveness test is concrete: when a new vulnerability becomes known in a widely used library, it must be determinable within minutes, solely from the bills of materials, which of the organization's own artefacts are affected. The provision of current bills of materials is required contractually from critical software suppliers; institutions without development of their own use the principle entirely through procurement.

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

MRIS.bd.1.M9 Integration of automated security testing into the development pipeline (S)

Three classes of check are combined in the pipeline: static code analysis (SAST, for example Semgrep, SonarQube), dynamic testing of the running application (DAST, for example OWASP ZAP) and dependency analysis (SCA, for example Dependency-Track, Snyk), supplemented by AI-supported code scanning. Findings at high or critical level with high confidence block the merge into the main branch (pre-merge block); vulnerabilities from the catalogue of actively exploited flaws (KEV) block immediately. The associated metric is remediation latency: the time from the availability of a robust remediation or effective compensating option to its demonstrated implementation — not the time from becoming known, since a patch is not always available immediately for newly listed vulnerabilities. The running systems are scanned daily.

AI-generated code is treated as a risk category of its own: 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, together with regular audit reporting with trend tracking, address that. The shift-left principle is the structural answer to the collapse of the patch gap window (see the module, Chapter 2.1): it removes from the attacker the window between exploit availability and the production patch.

Measurement: remediation latency for actively exploited vulnerabilities (guide value for internet-exposed systems: below 24 hours), together with the share of pipelines with security checking and a pre-merge block. Typical mistake: the checks run but blocks are routinely circumvented by exception — the control then exists only on paper.

MRIS.bd.1.M10 Automation of incident response with secured playbooks (S)

For unambiguous incident types, predefined playbooks are held on a SOAR platform that initiate containment actions automatically; human escalation takes place deliberately for ambiguous cases. AI-specific playbooks cover at minimum: 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 half-yearly, because 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 approval (human in the loop) for actions with a large blast radius (mass account lockout, network isolation of larger segments, certificate revocation), and protection of the playbook definitions like source code. Without this hardening, the response automation can be turned against the organization itself, for example through provoked false alerts.

For institutions without a security team of their own, managed detection and response services with 24/7 operation and a contractually agreed containment time are the more workable alternative to operating a SOAR platform in-house. Measurement: time to containment of unambiguous incidents (guide value: below 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 checking (S)

Security requirements are, where possible, formalized as executable rules (policy as code, for example with OPA or comparable engines) and integrated at two points: as a check before deployment (pre-deploy checks in CI) and as enforcement in running operations. Observance of the target states is checked continuously and at least daily; deviations automatically trigger work items (for example incident tickets) and become visible on dashboards with quantified deviation indicators.

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

Measurement: two separate metrics. Automation eligibility — share of the measures carried in the Statement of Applicability that are technically capable of automated checking; this classification is documented and reviewed annually. Continuous monitoring coverage — share of the measures classified as automatable with active automated checking. Referring to all measures would be misleading, because personnel, contractual and structural measures do not lend themselves to daily automated checking; guide values: at least 60 per cent after twelve months, at least 80 per cent after 24 months. Essential measures without automated checking are permissible only on a time-limited and justified basis. Typical mistake: dashboards are built without deviations triggering a binding response.

MRIS.bd.1.M12 Use of signed container images and protected execution environments (H)

For containers the following applies: base images come exclusively from controlled registries, are signed (for example with Sigstore/cosign) and are scanned automatically before deployment. Enforcement takes place technically through the admission controller policies of the platform (for example 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 a particularly high protection requirement, protected execution environments are deployed (trusted execution environments, for example Intel TDX, AMD SEV-SNP, ARM CCA), whose authenticity is evidenced through remote attestation. Confidential computing is currently the hardest available barrier against attackers with access to the operator's infrastructure and is therefore relevant in particular for regulated workloads and highly sensitive data processing.

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

MRIS.bd.1.M13 Implementation of demonstrable tenant separation (H)

The separation is documented and implemented at 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). What is decisive is the transition from conceptual assertion to evidence: the effectiveness of the separation is evidenced through regular, ideally automated separation tests.

As guide values for the test levels, the following apply: annually internally, automated per major release, supplemented by an examination by independent third parties. If cross-tenant access is identified during the separation tests or in ongoing operation, it is stopped without delay and the cause is analysed as part of incident handling. The measure limits the blast radius structurally: an attacker who gains a foothold in one tenant is demonstrably prevented from accessing other tenants — irrespective of how fast they operate.

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

MRIS.bd.1.M14 Carrying out threat-led penetration tests with AI scenarios (H)

Threat-led penetration tests are carried out with external red teams that reproduce real scenarios tailored to the current threat landscape. AI-specific scenarios comprise at minimum: AI-supported spear phishing including deepfake voice, attack chains fragmented into micro-steps, parallel multi-vector attacks, supply chain

compromise and cloud API abuse through compromised service accounts. The scenario pool is built multi-systemically (at least two independent agentic systems) and is 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 tests under the TIBER-EU methodology typically incur costs in the mid six-figure range per exercise. For institutions outside the scope of DORA, lower-cost formats are sufficiently effective: purple team exercises with MITRE ATT&CK scenarios, open source adversary emulation (for example Caldera, Atomic Red Team) and breach-and-attack simulation platforms.

The core of the measure: what is tested is whether the protective measures work against real AI approaches — not whether they are documented. That reduces the risk of false-positive compliance assurances. Measurement: closure rate — share of the vulnerabilities identified that are remedied within the agreed time window (guide value: at least 90 per cent). Typical mistake: test reports are filed without remediation being tracked.

3 Further information

3.1 Useful background

Origin and mapping of the requirements. The requirements of this module operationalize the thirteen Mythos Hardening Controls (MHC) of the MRIS framework; A1 additionally implements the MRIS recommendation for 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, the whole of 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 supplements 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 supplements A.5.35 and A.8.29.

Implementation sequence. The assignment of the requirements to basic, standard and elevated protection requirements follows the MRIS standard prioritization P0/P1/P2 for typical enterprise environments. It is a starting point, not a universal plan: where data have a long confidentiality period, A6 moves forward; where AI agents are deployed in production, A7 does; on container or multi-tenant platforms, A12 and A13 gain urgency. The dependency rules that apply in particular are: reassessment (A1) before investment in new measures; detection (A4) before response automation (A10); bills of materials (A8) as the basis for pipeline testing (A9); validation through testing (A14) after baseline implementation [MRISIG].

Maturity levels. The Implementation Guide describes, for each measure, a cumulative maturity path in three levels (Initial, Defined, Managed) with stage gates and metrics. Binding reading: the Initial level documents an implementation path begun and does not yet evidence fulfilment of the requirement; only from Defined onwards is the associated requirement deemed implemented in principle, and Managed describes the advanced and continuously effective implementation. The levels are suitable for internal progress measurement and for reporting to the leadership [MRISIG].

Integration into the ISMS. Every requirement of this module is worded so that it can be adopted into the Statement of Applicability of an existing ISO 27001-certified ISMS. Requirements that are not applicable are excluded there with a rationale — typical cases: A8 and A9 for institutions without software development of their own (the principle then takes effect through procurement, see M8), and A13 without multi-tenant operation. Statutory notification obligations under NIS2 and DORA are deliberately not carried as requirements of their own: they are direct legal obligations and are implemented in the ISMS through incident handling; A10 supplies the response capability necessary for that.

3.2 References

[MRIS] Mythos-Resistant Information Security (MRIS) — framework, Version 2.0 (EN), Richard Peddi, August 2026, CC BY-SA 4.0.

[MRISIG] MRIS Implementation Guide, Version 2.0 (EN), Richard Peddi, August 2026, CC BY-SA 4.0.

[ATTACK] MITRE ATT&CK Knowledge Base (content version v19.2), The MITRE Corporation, continuously updated, https://attack.mitre.org, last accessed 27 August 2026.

[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 14 August 2026.

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