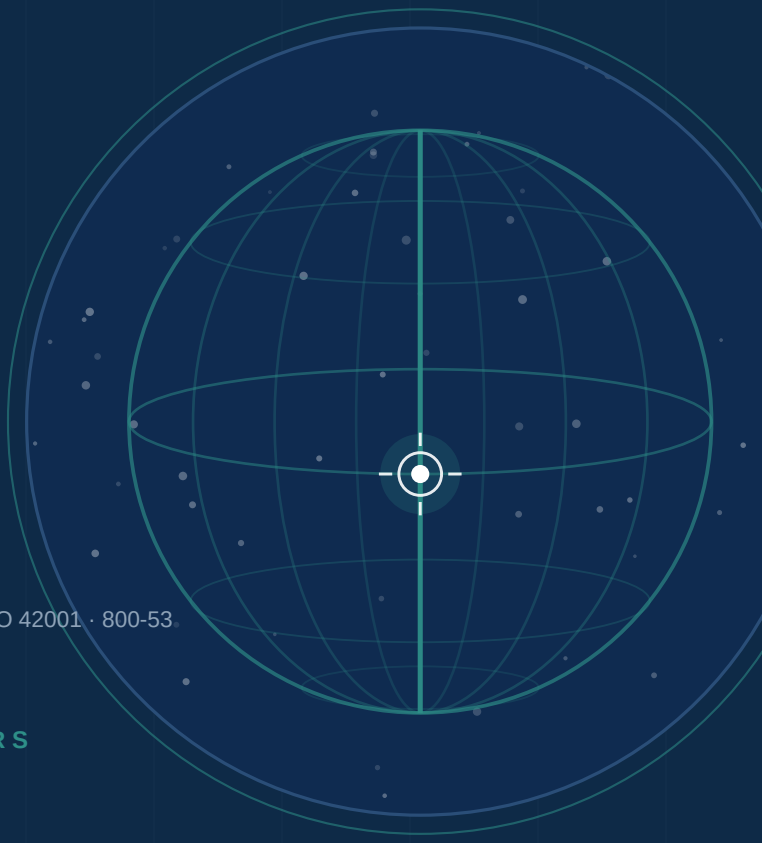
AI SECURITY NAVIGATOR

MERIDIAN

A practical control catalog for the AI your organization builds, buys, and uses.

Built on SAIF · ATLAS · NIST AI RMF · OWASP · AICM · AISMM · ISO 42001 · 800-53



THREE QUESTIONS THIS PAPER ANSWERS

Where do we start?
17 controls. 90 days. A complete first answer.

How do we prove it?
Assessor-ready evidence written into every control.

Does it fit what we already run?
Every control mapped to the frameworks you already use.

Habib Tora

Version 1.0 · June 2026

Abstract

Security teams asked to "secure AI" do not lack guidance. They are surrounded by it. MITRE ATLAS names the attacks. NIST AI RMF and ISO/IEC 42001 define governance. Google SAIF and the CSA AI Controls Matrix propose safeguards. The OWASP Top 10 for LLM Applications ranks vulnerabilities. The SANS maturity model describes progression. Every one is useful. Every one is also a deep body of work on its own. Each demands more than most teams can absorb, and together they still leave the practical questions open: what to implement on Monday, how to prove it on Friday, and how the frameworks fit together.

MERIDIAN is the synthesis: 65 controls in six functions that connect the seven frameworks to one another and to NIST 800-53. Every control says what to do in one sentence, tells an assessor how to verify it, and shows where it sits in each source framework. The catalog ships as a free reference site and as machine-readable OSCAL, with a defined federal baseline. An organization that simply uses AI can reach the baseline in about 90 days with 17 controls. An organization running high-impact systems will spend years mastering the upper tiers. Easy to start, difficult to master. That is the design.

1 The gap

The AI security framework landscape is fully populated and barely connected. When an organization adopts AI, the threat intelligence lives in ATLAS. The governance questions live in NIST AI RMF and ISO 42001. The engineering guidance lives in SAIF and AICM. Pentest findings arrive with OWASP LLM numbers. Auditors speak 800-53. These vocabularies do not talk to each other. A team that deploys prompt-injection defenses cannot easily show which ATLAS techniques it counters, which RMF outcomes it satisfies, or which 800-53 controls it inherits. The answers are knowable. No one has written them down in one place.

Fragmentation has a cost. Duplicated assessments. Governance documents disconnected from technical reality. AI security programs that are policy binders without detections. Worst of all, the quiet conclusion that the field is too immature to act on, so nothing gets done.

History shows what works. The CIS Controls won because they were prescriptive and tiered. ATT&CK won because its stable IDs became a shared vocabulary. The NIST Cybersecurity Framework won because its core fits on a napkin. MERIDIAN borrows all three lessons: a small core, a prescriptive entry tier, frozen IDs, and a crosswalk that preserves every existing framework investment.

One more lesson, learned by omission: nothing in this space ships machine-readable. Federal agencies and governance platforms ingest control catalogs as data, in a NIST format called OSCAL. MERIDIAN is OSCAL-native from day one.

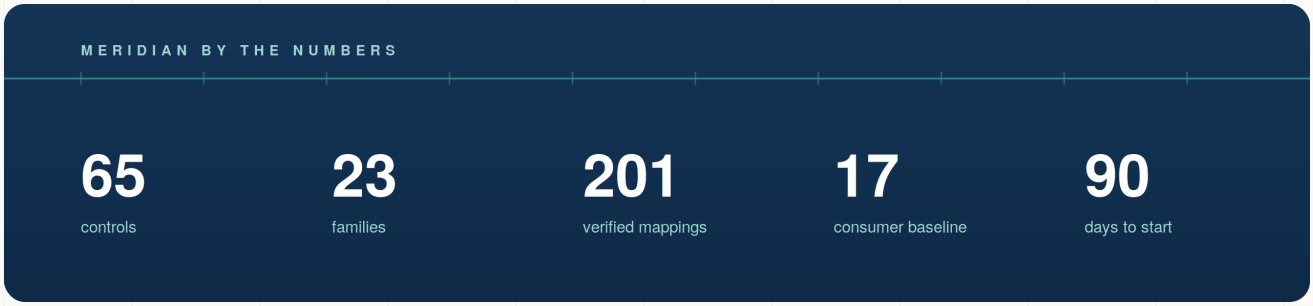


Figure 1. MERIDIAN by the numbers.

2 The model

MERIDIAN has six functions. Govern sets policy and accountability. Discover finds your AI, your data, and your exposure. Secure protects data, models, and interactions. Detect spots attacks in time to act. Respond contains and recovers. Assure proves it all works. Six words a leader can hold in one conversation.

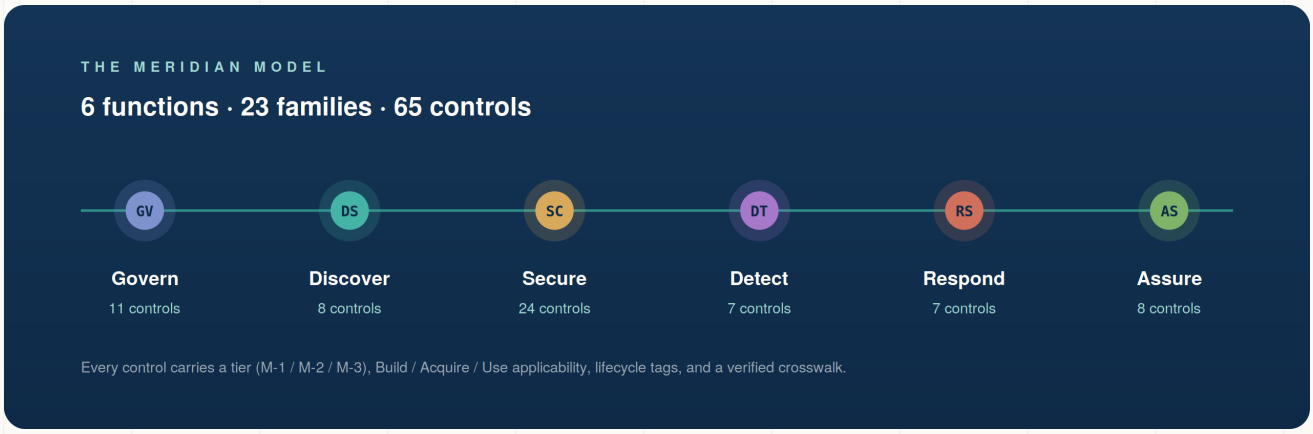


Figure 2. Six functions on one reference line.

Three naming rules cover the whole catalog:

- A function is a two-letter code: GV, DS, SC, DT, RS, AS.
- A family is a code plus a number. SC.7 is the seventh family inside Secure.
- A control is a code plus a two-digit ID. SC-13 reads as Secure, control 13.

Every control then carries three labels, answering three questions:

- **Tier** answers how soon. M-1 is the 20-control entry tier. M-2 adds 31 controls for organizations building or deeply integrating AI. M-3 adds 14 for high-impact systems.
- **Applicability** answers who it binds: Build (you train or host models), Acquire (you buy products with AI inside), or Use (your people use AI services). Most organizations never train a model; MERIDIAN tells a pure consumer exactly which controls apply, in one query.
- **Lifecycle** answers when it applies, from data through decommission.

Figure 3. The adoption ladder: three tiers, and the maturity scale behind them.

The labels combine like filters, and the combinations are where the catalog gets practical.

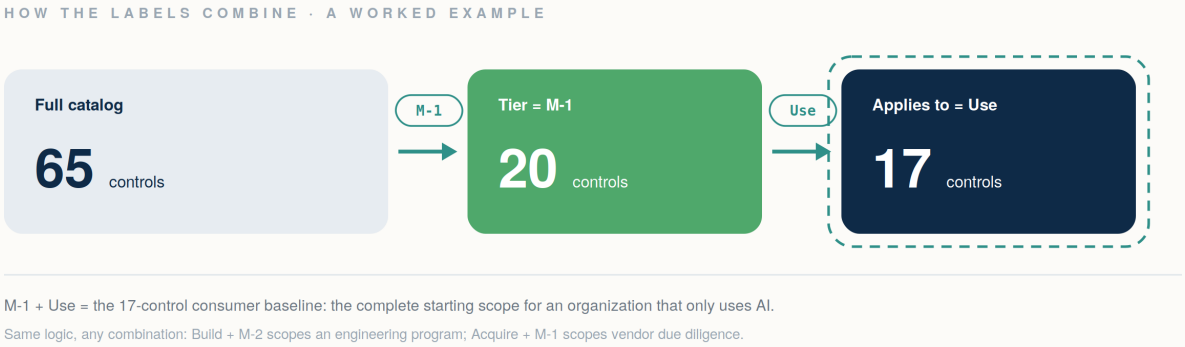


Figure 4. From 65 to 17: tier and applicability filters derive the consumer baseline.

The applicability axis also draws a shared-responsibility line, the one cloud security took a decade to formalize. Consume a frontier model through an API and the provider owns the model layer: weight security, training data, pre-training safeguards, most abuse monitoring. The deployer owns the rest: data classification, retrieval permissions, output handling, agent permissions, logging, and incident response. MERIDIAN's Acquire controls exist to verify the provider's half. Assess, contract, confirm. Never assume.

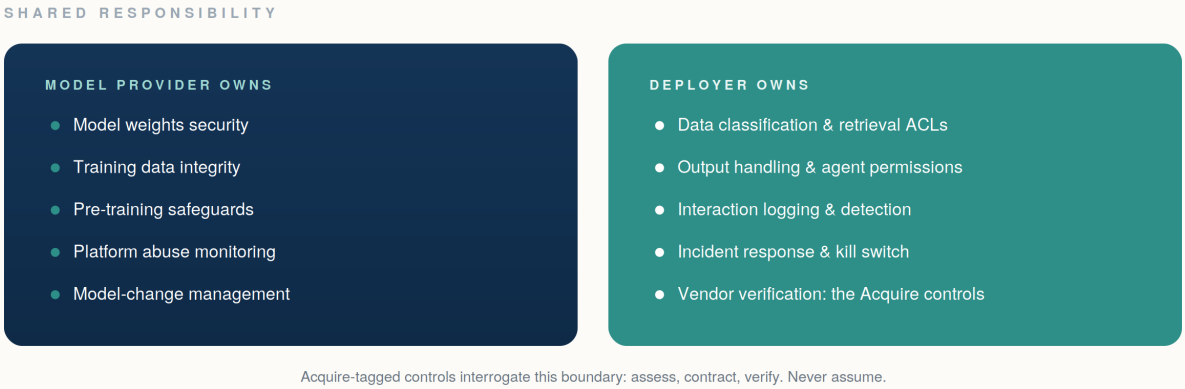


Figure 5. The shared-responsibility boundary. Acquire controls exist to verify the provider's half.

The control count is deliberately skewed. Secure holds 24 of 65 controls, because that is where practitioners work. Govern holds 11, enough to anchor accountability without drowning the catalog in policy. Agentic AI gets its own family, SC.7, the seventh family in Secure: agent permissions, vetting of tools and MCP servers (the Model Context Protocol connectors that give agents their capabilities), sandboxing, non-human identity, inter-agent traffic, and agent memory. The threat catalog already moved there. Section 6 shows how.

Control IDs are frozen as of this release. New controls append. Nothing is renumbered or reused. ATT&CK proved why: an ID that might change is an ID nobody cites, and citation is how a vocabulary spreads.

3 Anatomy of a control

Every control stands alone. Take SC-13, System Prompt Protection. The statement is one testable sentence: system prompts contain no secrets and are protected against disclosure, and disclosure is treated as expected rather than catastrophic. That framing corrects a common error. Teams that treat the system prompt as a security boundary build systems that fail badly when it leaks. The evidence statement tells the assessor what to check: prompts pass secret scans, and disclosure tests show no change in privilege or data access. The tier is M-1. The applicability is Build and Use. The crosswalk places it in every vocabulary at once: OWASP LLM07, ATLAS AML.T0051.000, and ATLAS AML.T0084.

The evidence statements are the difference between a framework you implement and a framework you admire. All 65 follow NIST 800-53A style: name the artifact, state the property, use the sampling language assessors already know. At the advanced tier, evidence means demonstration. Simulated extraction traffic must trigger the rate limits. Agent deviation tests must fire alerts. Kill switches must show activation records. A control whose evidence cannot be falsified is a press release. MERIDIAN aims to contain none.

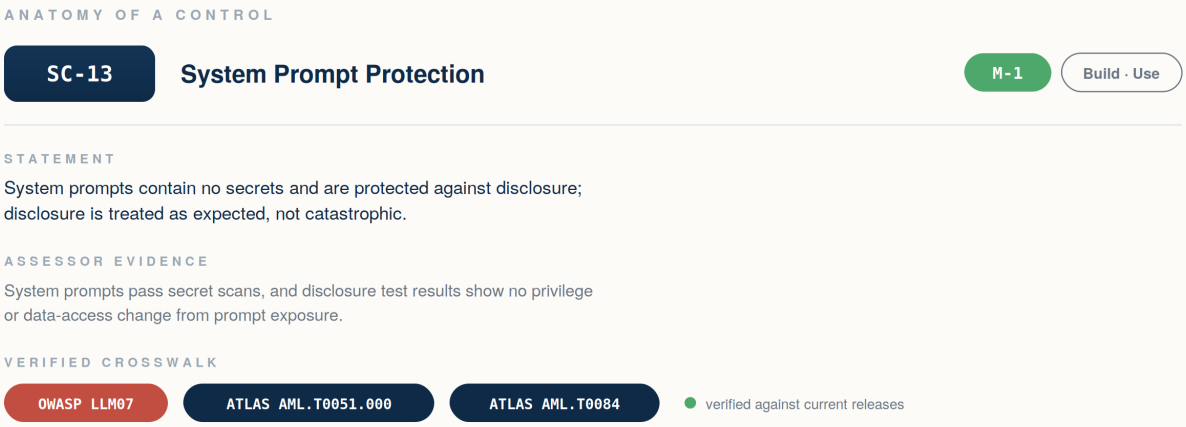


Figure 6. SC-13 as a complete record: statement, assessor evidence, verified crosswalk.

4 Easy to implement: the 90-day baseline

The tiers answer one question: which controls, in what order. M-1 is the entry tier: 20 controls, the minimum for an organization to use AI without flying blind. A competent team can stand it up in roughly 90 days. M-2 adds 31 controls for organizations building or deeply integrating AI. M-3 adds the final 14, for high-impact systems. That is where the hard problems live.

The baseline gets sharper still when intersected with applicability. An organization that only uses AI: no training, no fine-tuning, no model hosting. It owes exactly 17 controls. A policy. Acceptable use. A named risk owner. A risk assessment process. Vendor assessment. Training. An AI inventory. Data classification. Sensitive-data minimization. Model access control. Prompt-injection defense. Input validation. System prompt protection. Output handling. Secrets hygiene. Interaction logging. An incident playbook. Count them: 17, and on the reference site they are a one-tap preset, each with implementation guidance attached. Every other framework answers "where do we start?" with "it depends." MERIDIAN answers with a list you can finish.

> *Every other framework answers "where do we start?" with "it depends." MERIDIAN answers with a list you can finish.*

5 Difficult to master: the maturity ladder

Tiers answer which controls. Maturity answers how well. Each of the 23 families carries a five-level rubric, L1 through L5, and the upper levels are demanding on purpose.

The climb has a shape. At L1 a control exists somewhere. At L3 it is enforced everywhere, which is where strong programs tend to plateau. L4 is different in kind, not just in degree: now you measure. Detection efficacy proven by purple-team detonation, meaning real attack techniques run against your own defenses to see what actually

fires. Guardrail bypass rates tracked against adversarial test sets. Agent blast radius quantified. L5 anticipates: defenses retune from observed attacks, inventories maintain themselves, attestation is always current.

Why so much weight on measurement? Because of what testing can and cannot tell you. A deployed model's behavior is a distribution, not a fixed property, and an eval samples that distribution once. Benchmarks saturate. Evals can be gamed. A model that passes a leakage suite today can fail after a provider update or a corpus change. MERIDIAN therefore treats pre-deployment evals as necessary and weak evidence, and makes continuous validation a family of its own: eval regression in production, recurring red teams, re-tests after material change. Testing raises confidence and bounds risk. It does not certify safety.

6 The agentic family

Agentic systems moved the threat catalog faster than anything before them. MITRE ATLAS v5.6.0 now documents agent-specific techniques: context poisoning of agent memory and threads, poisoned and trojanized tools, supply-chain rug pulls where a component turns malicious by update, credential harvesting from agent configurations, exfiltration and data destruction through an agent's own tool calls, sandbox escape, and machine compromise through local agents.

SC.7 maps to this list one to one. Agency limits counter tool-call exfiltration and destruction. Tool and MCP vetting counters poisoned tools, with re-review on update against rug pulls. Sandboxing counters escape to host. Non-human identity counters credential harvesting. Inter-agent security counters triggered and delayed injection. Memory protection counters context poisoning. Every mapping cites the exact ATLAS technique ID, so a red team can test it and a detection engineer can build on it.

6 controls, 1:1

Every SC.7 agentic control cites the exact MITRE ATLAS v5.6.0 technique it counters.

Two commitments shape the family. First, assume prompt injection sometimes succeeds. No input filter stops it reliably, and any framework that claims otherwise is selling something. The layers exist so the inner ones hold when the outer one fails. Second, oversight must scale with autonomy. An approval gate sized for a chatbot is theater for a multi-agent system running long action chains. As models gain capability and autonomy, yesterday's measurements stop binding. The controls force re-measurement instead of assuming stability.

> *Assume prompt injection sometimes succeeds. The layers exist so the inner ones hold when the outer one fails.*

SC.7 AGENTIC AI SECURITY × ATLAS v5.6.0

Control to technique, one to one

CONTROL		ATLAS TECHNIQUES COUNTERED
SC-18	Agency limitation	T0086 · T0101
SC-19	Tool & MCP vetting	T0110 · T0104 · T0109 Rug Pull
SC-20	Execution sandboxing	T0105 Escape to Host · T0112.000
SC-21	Non-human identity	T0083 · T0098 Credential Harvesting
SC-22	Inter-agent comms	T0051.001 · T0051.002 · T0094
SC-23	Memory & state	T0080 Context Poisoning · T0081
DT-05	Behavior detection	T0053 · T0034.002

Design stance: injection will sometimes succeed. Inner layers hold when the outer fails. Oversight scales with autonomy.

Figure 7. SC.7 mapped one to one onto the agentic techniques of MITRE ATLAS v5.6.0.

Figure 8. Defense in depth. The outer layer will be bypassed; the inner layers contain it.

7 Federal by design

A note on scope before the details: if your organization never touches government work, skim this section; if it does, this is the part that makes MERIDIAN submittable rather than merely likable. Federal interoperability is an engineering requirement here, with acceptance criteria. The catalog publishes as OSCAL, validated against NIST's official 1.2.1 models. Statements sit in statement parts. Evidence sits in assessment-objective parts, where 800-53A tooling expects it. The 201 mappings travel as properties. Document UUIDs are deterministic per version, so re-exports are byte-stable. Anyone who diffs attestation artifacts knows why that matters.

MERIDIAN-FED is the named federal profile: every M-1 and M-2 control plus dataset provenance, OSCAL attestation, and independent assessment. It aligns with the current OMB AI memoranda, M-25-21 on use and

M-25-22 on acquisition, and points high-impact AI systems to the full M-3 set. Because every control maps to 800-53 Rev 5, agencies extend the baseline they already operate. The catalog also practices what it requires: maintaining OSCAL attestation is itself control AS-07.

54 controls

MERIDIAN-FED: the OSCAL profile a federal program can ingest as data, not prose.

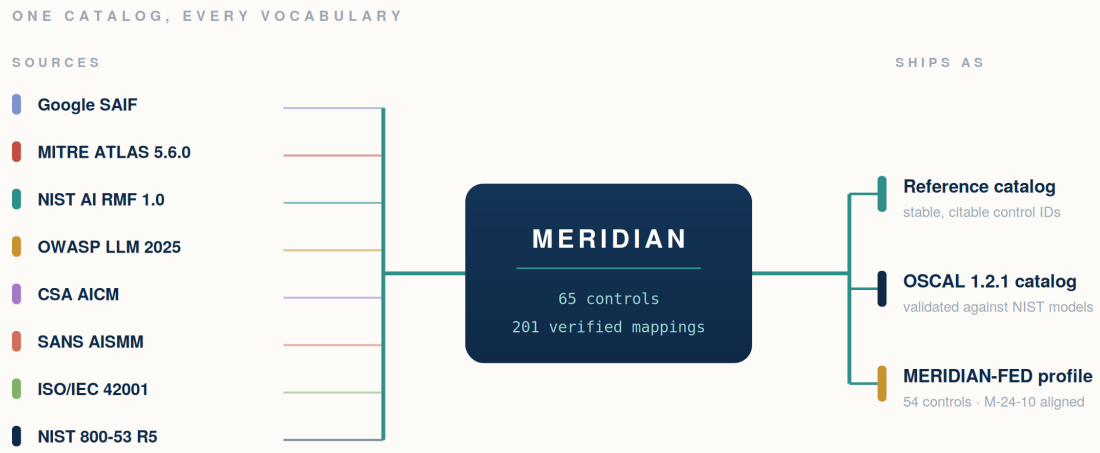


Figure 9. Eight frameworks in, one verified crosswalk, OSCAL and the federal profile out.

8 Rigor as a feature

Two decisions define the credibility posture. First, every mapping carries a verification flag, and nothing ships verified without a check against the current source release. The gate has already paid for itself. It caught a draft mapping that pointed model-weight protection at an ATLAS impact technique instead of the exfiltration techniques it actually counters, and it migrated every OWASP reference to 2025 numbering. A framework that shows its corrections makes a stronger promise than one that claims it never needed any.

Second, the crosswalk is version-pinned: ATLAS 5.6.0 and OWASP LLM 2025 by name, with every ATLAS reference validated programmatically against ATLAS's published data. Source frameworks move fast. The agentic expansion proves it. When they move, MERIDIAN re-verifies and re-pins. A crosswalk is a rule library. It is only as good as its last validation run.

9 Scope, limits, and non-goals

MERIDIAN replaces nothing. ATLAS stays the threat encyclopedia; MERIDIAN says which controls counter which techniques. RMF and ISO 42001 stay the governance systems of record; MERIDIAN supplies the control layer they presuppose. OWASP stays the vulnerability lens. SAIF and AICM stay architectural references. SANS

stays a maturity philosophy. Meridians make a map usable. They do not redraw the territory.

The boundary on the other side matters just as much. MERIDIAN secures organizational AI: the enterprise that builds on models, buys products containing them, or puts them in employees' hands. It does not address frontier model development. Alignment research, catastrophic-misuse evaluation, and defending frontier weights against state actors are problems of a different scale, governed by the safety frameworks of the labs that train such models. An organization facing nation-state pursuit of its weights needs more than any general-purpose catalog. Drawing the line plainly is a feature. Frameworks that blur enterprise security and frontier safety tend to serve neither.

> A note on confidence. Claims here sit at three levels. The crosswalk is empirical: every mapping names a specific reference in a specific version, and anyone can check it. The tier and maturity assignments are judgment calls, published so practitioners can challenge them. The sufficiency of any control set against systems that improve every year is genuinely uncertain. We expect some judgments to be wrong, and the framework is instrumented to find out which.

The last limit is time. Capability moves faster than catalogs. A control sized for today's agents may not bind next year's. The answer is the same as for crosswalk drift: scheduled re-verification, version pinning, and append-only IDs. MERIDIAN carries an expiration assumption, not a permanence claim. Its maintenance discipline is part of its definition.

MERIDIAN is not a certification scheme or a product. The catalog and the OSCAL artifacts are free. The intended adopter is a security team that wants a defensible answer to two questions: where do we start, and how do we prove it?

10 Roadmap and invitation

Version 1.0 is the complete release: 65 frozen-ID controls, evidence throughout, 201 verified mappings, maturity rubrics, the reference site, and validated OSCAL with a federal profile. The path forward runs through the community: review of statements and tiers, extension of the implementation guidance that ships with every entry- and managed-tier control, detection content for the Detect controls, and above all crosswalk challenges. Crosswalks are version-pinned and re-verified against each source framework release; corrections ship as point releases, and control IDs stay frozen. Send challenges to ai@htora.dev. Verified corrections are credited in the release notes. "This mapping is wrong, and here is why" is the most valuable sentence a practitioner can send. Every control has a stable ID. Every mapping is published for scrutiny.

The field has enough frameworks. The work is making them answer to each other.

A Appendix A. Functions and families

Function	Families	Controls
GV Govern	Accountability & Policy · Risk Management · Third-Party & Supply Chain Governance · Workforce & Culture	11
DS Discover	AI Asset Inventory · Data Provenance & Classification · Exposure Mapping	8
SC Secure	Data Security · Model Protection · Supply Chain Security · Input & Interaction Hardening · Output & Action Safety · Infrastructure & Pipeline · Agentic AI Security	24
DT Detect	Telemetry & Logging · Threat Detection · Model & Pipeline Monitoring	7
RS Respond	AI Incident Response · Containment & Recovery · Disclosure & Learning	7
AS Assure	Pre-Deployment Assurance · Continuous Validation · Audit & Conformance	8

B Appendix B. At a glance

Property	Value
Controls / families / functions	65 / 23 / 6
Crosswalk mappings (verified)	201 / 201
Source frameworks	SAIF · ATLAS 5.6.0 · NIST AI RMF 1.0 · OWASP LLM 2025 · CSA AICM · SANS AISMM · ISO/IEC 42001 · 800-53 Rev 5
Tiers	M-1 = 20 · M-2 = 31 · M-3 = 14
Consumer 90-day baseline	17 controls (M-1 ∩ Use)
Federal profile (MERIDIAN-FED)	54 controls, OSCAL profile
Machine-readable	OSCAL 1.2.1, NIST-model validated, deterministic UUIDs
ID policy	Frozen at 1.0; append-only

C Appendix C. Control index

The full catalog, one line per control. Statements, evidence, crosswalks, and implementation guidance live in the reference catalog, the workbook, and the OSCAL artifacts. This index is generated from the same database that powers them.

ID	Control	Tier	Applies to
GV-01	AI Security Policy	M-1	A B U
GV-02	Named AI Risk Ownership	M-1	A B U
GV-03	AI Acceptable Use Policy	M-1	U
GV-04	AI Risk Assessment Process	M-1	A B U
GV-05	Impact Thresholds & Risk Tolerance	M-2	A B U
GV-06	Regulatory & Obligation Mapping	M-2	A B U
GV-07	Vendor AI Risk Assessment	M-1	A U
GV-08	Contractual AI Security Requirements	M-2	A
GV-09	Embedded-AI Disclosure	M-2	A
GV-10	AI Security Awareness & Training	M-1	A B U
GV-11	Secure AI Development Standard	M-2	B
DS-01	AI System Inventory	M-1	A B U
DS-02	Model Registry & Model Cards	M-2	A B
DS-03	Shadow AI Detection	M-2	U
DS-04	Training Data Inventory & Lineage	M-2	B
DS-05	Data Classification for AI Use	M-1	B U
DS-06	Dataset Provenance Attestation	M-3	A B
DS-07	AI Attack Surface Mapping	M-2	A B U
DS-08	AI Dependency Mapping (AI-BOM)	M-3	A B
SC-01	Training Data Integrity	M-2	B
SC-02	Sensitive Data Minimization	M-1	B U
SC-03	RAG & Knowledge-Base Access Control	M-2	B U
SC-04	Model Weight Protection	M-2	A B
SC-05	Model Access Control	M-1	A B U
SC-06	Extraction & Theft Resistance	M-3	B
SC-07	Model Provenance Verification	M-2	A
SC-08	Third-Party Model Scanning	M-2	A
SC-09	ML Package & Dependency Security	M-1	B
SC-10	Safe Model Serialization	M-2	A B
SC-11	Prompt Injection Defense	M-1	B U
SC-12	Input Validation & Filtering	M-1	B U
SC-13	System Prompt Protection	M-1	B U
SC-14	Output Handling & Encoding	M-1	B U

ID	Control	Tier	Applies to
SC-15	Guardrails & Content Controls	M-2	A B U
SC-16	AI Infrastructure Hardening	M-2	B
SC-17	Secrets Hygiene in AI Pipelines	M-1	B U
SC-24	Secure AI Decommissioning	M-2	A B U
SC-18	Agency Limitation	M-2	B U
SC-19	Plugin, Tool & MCP Server Security	M-2	B U
SC-20	Agent Execution Sandboxing	M-3	B
SC-21	Agent Identity & Credentialing	M-2	B U
SC-22	Inter-Agent Communication Security	M-3	B
SC-23	Agent Memory & State Protection	M-3	B U
DT-01	AI Interaction Logging Standard	M-1	A B U
DT-02	AI Log Protection & Privacy	M-2	A B U
DT-03	Injection & Jailbreak Detection	M-2	A B U
DT-04	Abuse & Extraction Detection	M-3	B
DT-05	Anomalous Agent Behavior Detection	M-3	B U
DT-06	Model Drift & Behavior Monitoring	M-2	A B
DT-07	Pipeline Integrity Monitoring	M-2	B
RS-01	AI Incident Response Playbooks	M-1	A B U
RS-02	AI Incident Classification	M-2	A B U
RS-03	Model Kill Switch & Isolation	M-1	A B
RS-04	Model Rollback & Version Recovery	M-2	A B
RS-05	Poisoning Remediation	M-3	B
RS-06	AI Incident Reporting & Disclosure	M-2	A B U
RS-07	Post-Incident Eval Feedback	M-3	A B U
AS-01	Pre-Deployment Security Evaluation	M-1	A B
AS-02	AI Red Teaming	M-2	A B
AS-03	Security Gates in AI CI/CD	M-2	B
AS-04	Eval Regression in Production	M-3	A B
AS-05	Periodic Adversarial Re-Testing	M-3	A B
AS-06	AI Control Audit Trail	M-2	A B U
AS-07	Machine-Readable Attestation (OSCAL)	M-3	A B U
AS-08	Independent Assessment	M-3	A B