AISIF Companion Instrument

AISIF Maturity Self-Assessment Instrument

Companion document to the AISIF Practitioner White Paper · Supersedes the ten-question draft · Prepared July 2026

1. Purpose and Method

This instrument operationalizes the AISIF Maturity Model (white paper, Section 6) as a structured self-assessment. It comprises twenty-six questions organized by the five AISIF layers — Governance, Data, Model, Application, Operations — with the recursive feedback loop tested by the closing Operations questions. Each question is written to be binary-answerable with evidence: the honest answer is “yes, and here is the artefact” or it is “no.” A qualified answer is a no.

The question set incorporates the empirical findings of Case Studies D and E: classification of ancillary data stores, shared-responsibility assignment for third-party platforms, external attack-surface monitoring, credential lifecycle management, and third-party identity parity. It also extends the Application layer to agentic AI systems, consistent with the priority direction identified in Section 8 of the white paper.

2. Assessment Instructions

- Answer per layer and per estate. Score your own infrastructure and your SaaS/third-party data estate separately. Case Study E demonstrates that these can sit at different maturity levels simultaneously, and questions 5, 6, 20 and 23 will frequently score differently across the two. An aggregate answer conceals exactly the asymmetry that campaign exploited.
- Answer with evidence, not intent. A control that is planned, available but unmandated, or documented but unexercised is answered “no.” The operational rule from Case Study E applies: a control that is available but unmandated should be treated, for risk purposes, as absent.
- Report maturity per layer, not as a single aggregate. Case Study D demonstrates that an organization can operate at high capability in one layer while exhibiting Ad Hoc indicators in another; a single score conceals the gap that matters.

3. The Instrument

Table 1. AISIF maturity self-assessment questions by layer. Highlighted rows in the final column mark the five proposed Level 2 floor questions.

#	Self-Assessment Question	Capability Tested / Rationale	Indicative Maturity Linkage
Governance Layer — questions 1–6			

#	Self-Assessment Question	Capability Tested / Rationale	Indicative Maturity Linkage
1	Do you maintain a complete inventory of every AI system in use — including embedded SaaS AI features, AI agents, and MCP/tool integrations — with a named accountable owner recorded for each?	AI asset inventory and ownership; the shadow-AI test. You cannot govern, assess, or guardrail a system you have not recorded.	**Level 2 floor (proposed)**
2	Is there a standing AI governance body with defined membership, meeting cadence, and documented authority to accept, reject, or condition AI risk?	Institutional governance. Distinguishes a programme from a collection of individual decisions.	**Level 2 floor (proposed)**
3	Can you name the individual accountable for each AI system's security posture, and confirm they have reviewed its risk assessment in the last six months?	Live accountability. Ownership that has not exercised review within a defined period is nominal, not operating.	**Level 2 floor (proposed)**
4	Does every AI system pass a defined risk-assessment and approval gate before production, and does that gate operate at deployment cadence rather than on a periodic review calendar?	Gate velocity. Case Study D: a gate on a quarterly calendar does not govern a system that ships and scales in days.	Level 3
5	For every third-party AI or data platform, is there a documented shared-responsibility assignment, and are platform-available security controls — MFA, network policy, logging — made mandatory by your own standards rather than left to default configuration?	Shared-responsibility closure. Case Study E: a control that is available but unmandated should be treated, for risk purposes, as absent.	Level 3
6	Are contractors and other third parties accessing your AI systems or data estate held to the same identity and endpoint standards as your employees?	Third-party parity. Case Study E: the initial access vector was contractor endpoints outside the customers' own standards.	Level 2–3
Data Layer — questions 7–11			
7	Can you trace every record in your current training dataset back to its source?	Data provenance. The precondition for poisoning detection (Case Study B) and for lawful-basis and quality assurance.	Level 3
8	Are all datasets feeding AI workloads — training corpora, fine-tuning sets, feature stores, and retrieval indexes — inventoried and classified, including derived and ancillary stores such as logs, traces, and telemetry?	Estate-wide classification. Case Study D: inference telemetry is user data; ancillary stores inherit the sensitivity of their contents.	Level 2–3
9	Are chat content, personal data, and secrets redacted or minimized at the point of log and telemetry ingestion, rather than persisted in plaintext?	Minimization at ingestion. Removes the amplifier that turned a misconfiguration into a mass exposure in Case Study D.	Level 2–3
10	Would you know within 24 hours if a data distribution anomaly occurred in your inference pipeline?	Pipeline integrity monitoring. The operational control that documentation-oriented frameworks describe but do not operationalize (Case Study B).	Level 3–4

#	Self-Assessment Question	Capability Tested / Rationale	Indicative Maturity Linkage
11	Would you detect anomalous bulk egress from the data estate that feeds your AI workloads before external parties told you about it?	Egress detection. Case Study E: intrusions surfaced through threat intelligence and extortion, not internal telemetry.	Level 3
Model Layer — questions 12–14			
12	Has an adversarial test been performed against your model(s) using ML-specific attack scenarios from MITRE ATLAS in the last 12 months — not just a standard application penetration test?	ML-specific assurance. A web-app penetration test does not exercise evasion, extraction, or inversion techniques.	Level 3
13	Is model provenance verified — models sourced from approved registries, with an AI bill of materials maintained for production systems?	Model supply chain. The model-side counterpart of question 7; unverified artefacts are un-auditable dependencies.	Level 3
14	Are model changes — fine-tunes, system prompt revisions, and version upgrades — subject to evaluation and approval before release, with the ability to roll back?	Model change control. Behavioral change without evaluation is untested deployment, whatever the release notes call it.	Level 2–3
Application Layer — questions 15–19			
15	Are prompt injection defenses implemented at the application layer?	Direct injection defense. OWASP LLM01; the baseline runtime input control.	Level 2–3
16	Is retrieved RAG content treated as untrusted input before it is passed to the model?	Indirect injection defense. Case Study A: the mechanism that input sanitization alone does not reach.	Level 2–3
17	Are model outputs consumed by downstream systems — code, queries, commands, URLs — validated or sandboxed before execution, never executed blindly?	Output handling. The control that separates a bad answer from an executed compromise.	Level 2–3
18	Do AI agents operate under dedicated least-privilege identities, invoke only allowlisted tools, and require human approval for irreversible or high-impact actions?	Agentic containment. The extension of the framework's future-work section commits to; scope defined per agent, never inherited by default.	Level 3
19	Are all credentials used by AI systems vaulted and short-lived, with secrets prohibited from prompts, logs, and agent-accessible storage?	Secrets hygiene. Case Studies D and E jointly: credentials that transit AI pipelines end up in places adversaries read.	Level 2–3
Operations Layer — questions 20–26			
20	Is every data store in the AI serving path — including analytics, observability, and telemetry stores — authenticated, network-restricted, and covered by MFA where applicable?	Infrastructure identity trust. The shared proximate cause of Case Studies D and E.	**Level 2 floor (proposed)**

#	Self-Assessment Question	Capability Tested / Rationale	Indicative Maturity Linkage
21	Do you operate continuous external attack-surface monitoring — would you discover an exposed AI-adjacent asset before an outside researcher does?	Exposure management. Case Study D: an external team found the asset within minutes; the question is who finds yours first.	Level 3–4
22	Are AI interactions — prompts, outputs, tool calls, and agent decisions — logged and integrated into your security monitoring, with alerting on anomalous behavior?	AI-specific observability. The control that makes every other control auditable; replaces a yes/no monitoring question that did not discriminate between maturity levels.	**Level 2 floor (proposed)**
23	Do credentials for AI systems and their data estate have defined maximum lifetimes, scheduled rotation, and revocation triggered by compromise intelligence?	Credential lifecycle. Case Study E: four-year-old harvested credentials remained valid because nothing forced them not to be.	Level 2–3
24	Does every production AI system have a tested disablement mechanism and an AI-specific incident response playbook?	Response capability. A kill switch that has never been exercised is a hypothesis, not a control.	Level 2–3
25	Are operational monitoring findings reviewed with appropriate business stakeholders, in your governance risk-management cycle, on a defined schedule?	Feedback-loop routing. Findings that do not reach governance on a schedule do not inform it.	Level 3
26	Does that cycle have the authority to trigger control updates?	Feedback-loop authority. Kept deliberately separate from question 25: a review without authority is the framework's definition of a feedback loop that does not exist.	Level 3–4

4. Scoring Note (Provisional)

Five questions — 1, 2, 3, 20 and 22 — are proposed as the Level 2 (Controlled) floor: an organization that cannot answer all five affirmatively remains at Level 1 (Ad Hoc) regardless of its other answers, because inventory, institutional governance, live accountability, infrastructure identity trust and basic AI observability are preconditions for a controlled program rather than features of one. Remaining questions carry indicative linkages to Levels 2–4; progression to Level 3 (Integrated) and Level 4 (Adaptive) should additionally require affirmative answers to the feedback-loop questions (25 and 26), which test whether the program learns.

These thresholds are stated as proposals, not validated cut-offs. The white paper identifies the absence of a validated quantitative scoring methodology as an open limitation; the level-gating structure above is a hypothesis to be tested against practitioner self-assessment data before the instrument is used for benchmarking or cross-industry comparison.