AI Clarity Framework

An Evidence Based Approach to Enterprise AI Risk and Assurance

Ascendra | White paper | Version 0.2 for public consultation

Executive perspective

AI risk assessments are useful only when their findings lead to decisions. Leaders need to understand where AI is used, who is accountable, which protections operate, whether its data is fit for purpose, and whether its behavior remains acceptable. A single maturity number can conceal important differences between these questions.

Ascendra proposes the AI Clarity Framework as a repeatable method for converting assessment evidence into five independent capability ratings, explicit risk findings and a measurable improvement plan. Its signature AI Clarity Profile gives executives a consistent view while preserving the evidence and scope behind each result.

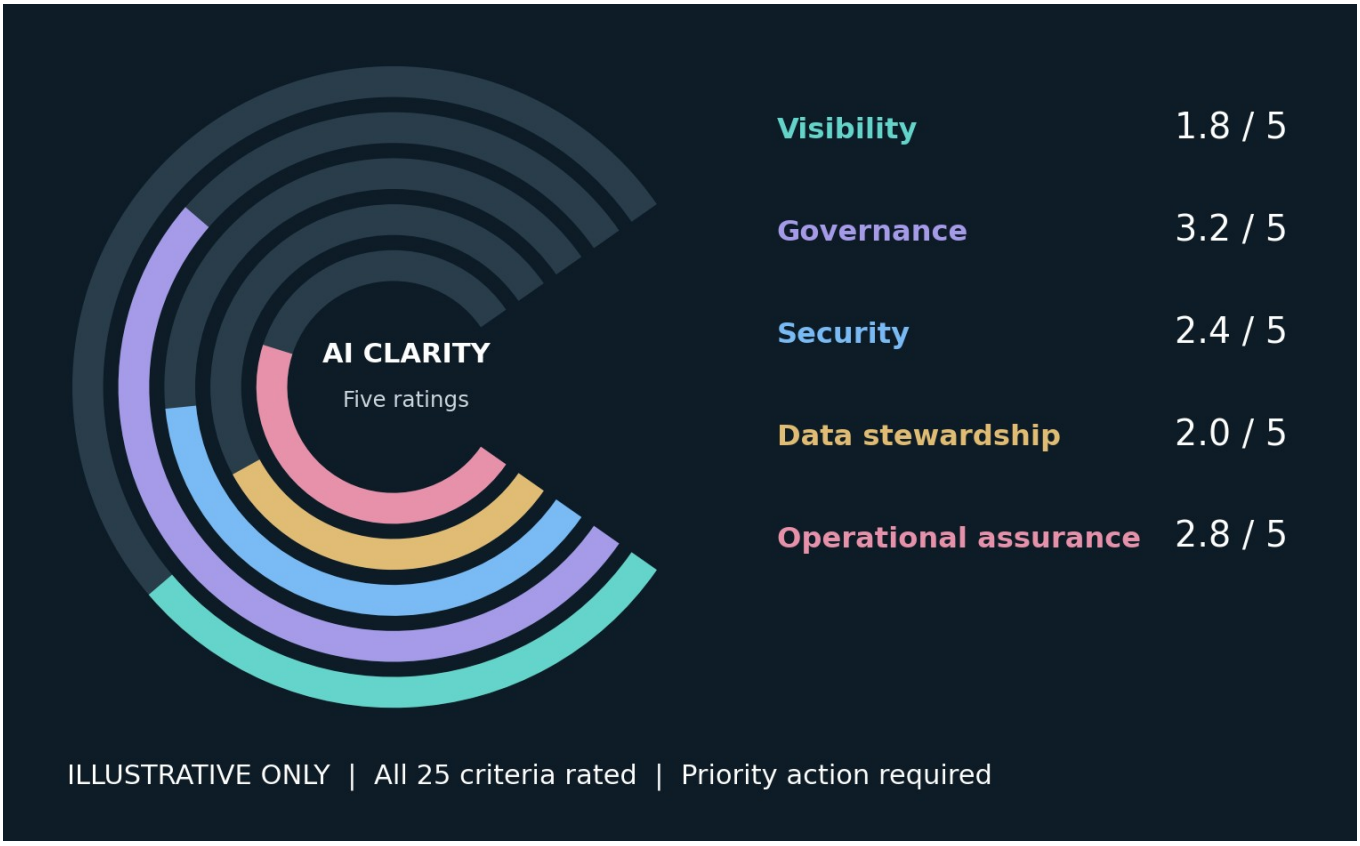


Figure 1. Proposed five-ring profile. Values are fictional, not customer results or industry benchmarks. Each ring uses the same 0 to 5 scale; ring length and area must not be interpreted as weights. Gray segments show distance to level 5, not missing evidence.

This consultation paper proposes a method for testing and refinement. It does not establish an accredited certification, demonstrate industry adoption or imply endorsement by MITRE or NIST. The five dimensions and scoring rules are Ascendra proposals.

Why five dimensions

The earlier three-dimension concept covered visibility, governance and security controls. Adding data stewardship and operational assurance makes the model more useful for real AI decisions: a well-governed, access-controlled system can still rely on obsolete information or perform poorly in its intended task.

Dimension	Question answered	Assessment boundary
AI Visibility	What AI exists and what is it doing?	Inventory, discovery, attribution and observable data or tool activity.
AI Governance	Who decides and who is accountable?	Policy, approval, oversight, supplier decisions and risk acceptance.
AI Security	Can threats or misuse compromise the system?	Access, protection, secure change, runtime defenses and incident response.
AI Data Stewardship	Is the data appropriate for its intended use?	Ownership, provenance, quality, permitted use, retention and source upkeep.
Operational Assurance	Does AI perform acceptably over time?	Evaluation, harmful failure tests, monitoring, intervention and recovery.

Established references and Ascendra contributions

MITRE SAFE-AI provides a security assessment foundation based on NIST SP 800-53 Rev. 5 and MITRE ATLAS. It considers controls for AI-enabled systems and the division of responsibilities across system elements. Its stated scope excludes several broader trustworthiness objectives. It therefore supports the security dimension without serving as a complete basis for every dimension. [1]

NIST AI RMF organizes risk management through Govern, Map, Measure and Manage. Its trustworthiness characteristics include validity and reliability, safety, security and resilience, accountability and transparency, explainability, privacy and fairness. These references inform the breadth of the proposal; the five dimensions are not an official NIST mapping. [2, 3]

Keep scope and evidence separate from capability

Record data sensitivity, decision impact, system criticality, autonomy and affected populations as assessment context. These factors determine applicable tests and required rigor. Report evidence confidence separately. Neither context nor confidence should become an extra maturity dimension that offsets a weak capability.

Agentic AI and Model Context Protocol connections require a contextual module when present. Map agent identities, delegated authority, tool permissions, traces, approval boundaries and revocation tests into the same five dimensions. Do not penalize an organization for having no applicable agent deployment.

The proposed assessment register

Version 0.2 defines 25 pilot criteria, five per dimension. Each criterion receives one rating. The evidence below illustrates what to inspect; assessors must define required tests, sampling and level-specific anchors before pilot scoring. This is a proposed register, not a completed control-by-control standards crosswalk.

ID	Criterion	Evidence or test
V1	AI inventory	Reconciled AI services, models, embedded features and agents.
V2	Discovery coverage	Measured telemetry coverage across known environments.
V3	Attribution	Records linking AI activity to identities and responsible owners.
V4	Data and action visibility	Traces of relevant data flows, connectors and tool actions.
V5	Change reconciliation	New or changed AI compared with approved records.
G1	Accountability	Named decision owners, policy and responsibility assignments.
G2	Risk decisions	Intake, impact assessment, approval and exception records.
G3	Human oversight	Defined review, escalation and affected-party processes.
G4	Supplier and lifecycle oversight	Supplier obligations, change approvals and retirement decisions.
G5	Management assurance	Training, reporting, challenge and tracked issue closure.
S1	Identity and authorization	Least privilege, delegation and denied-action tests.
S2	Data and model protection	Access boundaries, leakage tests and integrity protections.
S3	Secure lifecycle	Component provenance, security testing and controlled releases.
S4	Runtime safeguards	Adversarial-input and prohibited tool-action tests.
S5	Security response	Detection, investigation, containment and response exercises.
D1	Ownership and provenance	Named data owners and traceable source lineage.
D2	Quality and suitability	Accuracy, completeness and relevance checks for intended use.
D3	Permitted use and privacy	Documented use restrictions and data minimization practices.
D4	Retention and deletion	Applied retention rules and verifiable deletion workflows.
D5	Knowledge maintenance	Refresh, versioning and removal of obsolete source material.
O1	Task performance	Predefined acceptance measures and representative evaluations.
O2	Safety and impact testing	Use-specific harmful failure and fairness evaluations.
O3	Behavior monitoring	Quality changes, drift and failure signals with response owners.
O4	Intervention and fallback	Human escalation, suspension and alternative-process tests.
O5	Operational recovery	Recovery exercises and revalidation after material changes.

Assign each tested outcome one primary criterion. Reusing an artifact is allowed when it supports different outcomes; duplicating credit for the same outcome is not. Governance sets oversight requirements; operational assurance tests whether intervention works.

How to evaluate and score

Assess a declared organization, business unit or AI system over a stated period. Select the applicable profile and agree the evidence plan before examining scores. A system assessment cannot be presented as an enterprise-wide result.

Level	Meaning	Required basis
0	Absent	Evidence confirms the required practice is absent.
1	Ad hoc	Isolated or reactive practices are demonstrated.
2	Defined	An approved process or design exists.
3	Implemented	Evidence demonstrates operation across declared scope.
4	Tested effective	Representative tests meet predeclared acceptance criteria.
5	Sustained	Repeated tests and monitoring show effectiveness over a stated period and an operating improvement process.
U	Unrated	Evidence is insufficient to support a rating.
N/A	Not applicable	A scope-based exclusion is documented and independently reviewed.

Levels are cumulative. Award the highest level for which all required anchors are met. When a criterion contains several required outcomes, use the lowest demonstrated level across them. Do not translate questionnaire answers such as Yes or Partial directly into ratings. A policy is evidence of design, not proof of effectiveness.

Aggregation and publication rules

A dimension score is the arithmetic mean of its applicable criterion ratings, rounded to one decimal. Equal weighting is a transparent pilot convention, not an empirically validated measure of risk. Publish a dimension score only when every applicable criterion in that dimension is rateable. Otherwise display Unrated and disclose the evidence gaps. Never treat U as zero or silently exclude it.

Do not publish an overall average across the five dimensions. Report the criterion values alongside the profile so a weak criterion remains visible. A critical exposure receives a Priority action flag regardless of its dimension average; a serious unknown receives Further evidence required. Assess residual risk separately using documented impact and likelihood judgments.

Evidence and coverage

For every rating, retain the criterion ID, applicable requirements, artifact identifiers, dates, environment, sample, test procedure, result, assessor rationale and limitations. Separate provider controls from customer responsibilities. Describe freshness, corroboration and representativeness; avoid unsupported numerical confidence claims.

Report evidence completeness as rateable applicable criteria divided by all applicable criteria, and operational coverage using known populations. For example, 180 of 200 managed endpoints monitored is 90% endpoint coverage; it is not proof that 90% of all AI use has been discovered. Record exclusions and blind spots explicitly.

Turning findings into a defensible profile

A repeatable assessment sequence

1 Define context. List the system boundary, AI uses, affected people, data sensitivity, autonomy, criticality, review period and inherited responsibilities.

2 Fix the assessment plan. Select criteria, contextual modules, exact reference versions, test anchors and sampling. Document exclusions and approval before scoring.

3 Inspect and test. Use documents, configuration, telemetry, interviews and controlled exercises. Test both expected operation and relevant failure paths.

4 Review judgments. Have a second assessor challenge material ratings, exclusions, unknowns and critical findings. Preserve disagreements and their resolution.

5 Report and improve. Issue the profile, criterion register, evidence limitations and remediation plan. Assign owners, target dates and closure tests. Reassess after material change.

Worked example of the score calculation

The fictional profile on page 1 assumes that all 25 criteria have sufficient evidence and that no criterion is excluded. The numbers illustrate the arithmetic only; they do not represent completed tests or an observed organization.

Dimension	Five criterion ratings	Mean
Visibility	1, 2, 2, 2, 2	1.8
Governance	3, 3, 4, 3, 3	3.2
Security	2, 2, 3, 2, 3	2.4
Data stewardship	2, 2, 2, 2, 2	2.0
Operational assurance	3, 3, 2, 3, 3	2.8

Example anchor for S1 authorization

At level 2, an approved identity and permission model names permitted tool actions. Level 3 additionally requires deployed enforcement across the declared scope. Level 4 requires representative attempts at unauthorized actions, cross-user access and revoked credentials to be denied. Level 5 adds repeated evidence over a predeclared review period and demonstrated handling of permission changes. Any unmet required outcome limits the rating.

The pilot protocol must define sample selection, expected results and the observation period before testing. The same discipline must be extended to every criterion. This paper does not claim that the full criterion-specific test catalog is complete.

Applying the framework to real decisions

Employee knowledge assistant

Consider an internal assistant answering employee policy questions from an approved repository. The assessment should establish which users and sources are in scope, whether restricted documents can be exposed, who approves the service, how obsolete policies are removed and whether answers are accurate enough for the intended use.

A stale policy document demonstrates why the new dimensions matter. Data stewardship assesses source currency and ownership. Operational assurance tests whether the assistant gives acceptable answers and escalates uncertainty. Security assesses unauthorized access or manipulation. Governance establishes the decision owner and acceptable-use boundaries. Visibility supplies the inventory and traceability needed to investigate the result.

Agent connected to customer systems

Consider an agent using MCP-connected tools to retrieve customer records and propose account changes. Evaluate the full authorization path, including the initiating user, agent identity, tool server and target system. Record what is observable and which actions can alter persistent records. The module applies because of the architecture and authority granted, not because a particular vendor is used.

Dimension	Illustrative module test
Visibility	Trace an action to the initiating user, agent, tool and affected record.
Governance	Inspect approval for delegated actions and escalation responsibilities.
Security	Attempt an action outside delegated authority and test revocation.
Data stewardship	Verify permitted data use, source provenance and retention.
Operational assurance	Test approval handling, safe interruption and recovery from partial failure.

An actionable remediation record

Suppose testing finds that an agent can perform a sensitive update without the required approval. Record the affected systems and evidence, assign an accountable owner, restrict the action pending remediation and define a retest. Closure requires successful denial of unauthorized actions and successful execution of approved actions under the intended identity.

A profile may carry both a moderate security score and a Priority action flag. The flag must remain prominent until the finding is remediated or an authorized decision is recorded. Risk acceptance does not improve capability scores. A vendor purchase, policy signature or planned remediation also does not demonstrate implementation.

Building a method the industry can adopt

Recognition must come from repeatable judgments, transparent evidence rules and practical use by independent practitioners. The publishable asset is the complete assessment method and reporting language. The visual is its recognizable summary.

Release as an open consultation

Publish this white paper as proposed version 0.2, with an accessible feedback route and a version history. Invite independent practitioners and affected-user representatives to review scope and missing outcomes. Select reuse and attribution terms before distribution.

Complete the assessment pack and run pilots

Before offering benchmark ratings, complete the criterion-specific anchors, applicability profiles, assessor guide, evidence templates and exact SAFE-AI and NIST reference crosswalk. For every mapping, record source version, control or outcome identifier, relationship, test coverage and any remaining gap. A thematic relationship is not equivalent compliance.

Use five to ten varied, consenting pilot organizations as an initial calibration sample, not a representative industry sample. Have assessors independently rate overlapping evidence. Track exact agreement, agreement within one level, disputed exclusions, evidence burden and sensitivity to aggregation. Set acceptance thresholds before collecting results; investigate disagreements and publish the resulting method changes.

Release version one with accountable stewardship

Release a stable methodology only after pilot results and external review support consistent interpretation. Establish independent review, assessor training, corrections, appeals and conflict-of-interest disclosures. Keep commercial sales decisions separate from scoring. Evidence may come from any suitable technology; no vendor or product determines the result.

A future State of AI Clarity report should compare compatible scopes, use cases and method versions, with explicit consent and anonymized aggregates. Disclose sample sizes, selection bias and assessment periods. Do not publish industry percentiles or leader labels until the sampling supports those claims. Set suppression rules to protect small or identifiable cohorts.

Questions for consultation

Are the five dimensions sufficiently distinct? Which outcomes require noncompensable minimum requirements? What evidence and observation periods justify sustained effectiveness? Can independent assessors reproduce ratings at reasonable cost? Use these questions to shape the pilot protocol.

Primary references

[1] MITRE. SAFE-AI A Framework for Securing AI-Enabled Systems. April 2025. https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

[2] NIST. AI RMF Core. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[3] NIST. AI Risks and Trustworthiness. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/

References accessed 13 October 2026. Groupings, examples and scoring conventions are Ascendra proposals.