

Agent Governance Maturity Model

v1.0 Reference Specification, companion to Paper III.1: The Assessment

Mohamad Amin Hasbini

Independent researcher · Paris, France

Published September 7, 2026

ABSTRACT

The reference specification of the Agent Governance Maturity Model v1.0: the agent inclusion test, two layers (estate pillars, agent surfaces), thirty-five controls with their impact ratings, the five bands with the pillar-floor gate, the three cryptographic surfaces with posture verdicts and the agility scale, the agility floor, discovery and conformance rules, outputs and versioning. Both instruments run publicly.

KEYWORDS: agent governance maturity model, AI agent assessment, agent inventory, estate pillars, agent surfaces, agility floor, pillar-floor gate, crypto-agility, Agent-CBOM, DORA, NIS2

CITE AS

Hasbini, M. A. (2026). Agent Governance Maturity Model v1.0: Reference Specification, companion to Paper III.1: The Assessment. Running Regulated Agents Series.

DOI <https://doi.org/10.5281/zenodo.22637725>

Available at <https://mahasbini.org/papers/agent-governance-maturity-model/>

PDF <https://mahasbini.org/publications/papers/agent-governance-maturity-model-v1.0.pdf>

Licence: CC BY 4.0

This document is the reference specification of the model that [Paper III.1: The Assessment](#) narrates. Readers who want the argument and the method walked end to end should read the paper first. This companion is for assessors, auditors and implementers who need the exact inclusion test, scales, weights, gates and conformance rules.

Version 1.0 corresponds to questionnaire v1.1 at assessment.mahasbini.org and Agent-CBOM schema 0.1 at agentcbom.mahasbini.org. It carries a citable DOI and an attribution licence.

1. Purpose and scope

This model measures two different things that are routinely conflated: the **governance maturity of an organisation's AI-agent estate**, and the **cryptographic and evidentiary assurance of an individual agent**. It exists so that both can be scored in one assessment pass, compared over time, and read by two audiences at once: a board deciding what to deploy, and a supervisor asking for proof.

Inclusion test. An agent is in scope when three conditions hold together. It has capability: credentials, keys or tool access that produce effects in other systems. It exercises run-time discretion: which action runs is decided at run time, by planning or dynamic action and tool selection, not fixed in advance. And it acts under authority that someone delegated, or should have: an agent running without a legitimate delegation is in scope, not out of it. A vendor-embedded copilot with tenant access passes all three. A deterministic batch job fails the second. A suggestion-only assistant with no execution rights fails the first, until someone wires it to one. When a case is

genuinely unclear, it enters the inventory: the cost of an extra line is minutes, and the cost of a missing one is the denominator.

Exclusions. The agent-assurance layer does not score model quality, alignment, task performance, prompt-injection resistance or behavioural safety. The estate governance layer contains controls that touch some of those risks, prompt-injection defences among them, but the model does not claim comprehensive coverage of them. It scores whether the estate that runs the agents is governed, and whether each agent can be identified, its traffic protected, and its actions proven.

Both instruments run publicly. Every level of every scale is defined as an observable state, so an assessment can be checked against the instruments rather than taken on trust.

2. The two layers

The model has two layers because the two questions have different objects.

Layer 1, the estate, is scored on four governance pillars through a thirty-five-control questionnaire. Its output is a weighted maturity band. Governance posture is a portfolio, and portfolios may average.

Layer 2, the agent, is scored on the three cryptographic surfaces each agent carries: the identity it presents, the channel it speaks over, the receipts it leaves. Its output is an agility floor with the three current-posture verdicts kept visible beside it, never an average. An attacker does not engage your average.

One doctrinal line connects them. Authorization has no per-agent surface, because authority is a relation between principals, not a property of one agent. What an agent carries of that relation is its evidence, and the evidence of a relation is a receipt. Authorization is therefore scored where it lives, in the estate's pillar and in the receipt surface that records its exercise.

3. Layer 1: estate maturity

3.1 Pillars and weights

PILLAR	WEIGHT	CONTROLS	FOCUS
Authenticate	0.30	10	Unique agent identities, lifecycle, credential rotation, cryptographic anchoring (PKI, post-quantum), non-repudiation, registry, human-owner linking, environment isolation, identity proofing, offboarding
Authorize	0.35	10	Least privilege, scope constraints, context-aware authorization, delegation chains, approval workflow, break-glass, just-in-time elevation, policy-as-code, usage audit, separation of duties
Communicate	0.25	10	Mutual authentication, message integrity, encryption at rest and in transit, post-quantum readiness, prompt-injection defense, outbound data-loss prevention, replay protection, state and memory integrity, audit logs, third-party API hardening
Meta (governance)	0.10	5	Governance body, agent-specific policy, role-based training, agent-aware incident response, vendor due diligence

Authorization carries the highest weight because an over-privileged agent amplifies every other failure mode. The thirty-five controls, with their identifiers and impact ratings, are listed in Appendix A.

3.2 Control scoring

Every control is scored 0 to 4 on an explicit per-control rubric in which each cell is a distinct observable state:

SCORE	LEVEL	MEANING
0	Not in place	Control absent or explicitly rejected
1	Ad hoc	Informal, personality-dependent, inconsistent
2	Managed	Partial coverage, usually production only, not standardized
3	Defined	Standardized, documented, applied across environments
4	Optimized	Measured, automated, continuously improved

3.3 Aggregation, bands and the pillar-floor gate

The pillar score is the mean of its control scores multiplied by 25, carried to two decimal places. The estate score is the weighted sum of the four pillar scores under the weights in 3.1, carried to two decimal places. Bands are assigned on that value against half-open intervals. Displayed values may be rounded to one decimal place after band assignment, never before.

INTERVAL	BAND	INDEX	READING
$0 \leq x < 20$	Initial	0	Material exposure, most controls absent
$20 \leq x < 40$	Managed	1	Foundations emerging, inconsistent coverage
$40 \leq x < 60$	Defined	2	Controls documented and applied, gaps in automation
$60 \leq x < 80$	Measured	3	Controls measured and regularly reviewed
$80 \leq x \leq 100$	Optimizing	4	Continuous improvement with quantified risk reduction

The same intervals band each pillar score. The estate band is then capped by the pillar-floor gate:

FinalBand = **min(RawBand, WeakestPillarBand + 1)**, on the band indices above.

A portfolio may average. It may not average away a collapsed pillar. An estate at authenticate 100, authorize 0, communicate 100, meta 100 computes 65.00, a raw band of Measured (3). Its weakest pillar bands Initial (0), so the gate caps the estate at index 1, Managed, with the gating pillar named in the record.

3.4 Gaps and remediation ordering

Two derived outputs make the score actionable. Every control scoring 1 or below is listed as a gap. Each control carries a normative impact rating of 1 to 5, fixed by this specification (Appendix A) and not assessor-entered. The priority of a control is its impact multiplied by (4 minus its score). The remediation backlog is ordered by priority, descending. Ties are broken by impact, descending, then by questionnaire order (pillar order authenticate, authorize, communicate, meta, then control identifier). Given the published constants, the ordering is reproducible from the control scores alone, which is what makes it defensible in front of a third party. An effort band is attached from the same product: 16 and above High, 9 to 15 Medium, 4 to 8 Low-Medium, below 4 Low.

4. Layer 2: agent assurance

4.1 The three surfaces

Each agent in the estate inventory carries exactly three cryptographic surfaces:

- **Identity.** The keys, certificates and signing material the agent presents to authenticate itself to a tool server or to another agent.
- **Channel.** The transport that carries the agent's traffic, including MCP and agent-to-agent protocols. The channel question is who set its quantum posture, and whether it can be changed.
- **Receipt.** The records that evidence an agent's actions and the authority under which they were taken, including whether and how those records are integrity-protected or signed. A receipt surface exists whether or not the records are signed. Its durability requirement: the evidence must remain independently verifiable throughout the retention horizon it serves, through the original signature or through an approved evidence-preservation mechanism (re-signing, archival timestamps, chained integrity, independently protected append-only records). Records that fail this requirement will stop verifying before the retention horizons they serve.

4.2 Posture verdict and agility rating

Each surface receives two things.

A current-posture verdict. The algorithms and key material the surface runs on today, recorded per surface as observed, with whether the posture is classical, hybrid or post-quantum, read against the retention horizon the surface serves. The verdict is reported. It is not scored, and it is not averaged with anything.

An agility rating on a 0 to 4 scale:

LEVEL	NAME	OBSERVABLE STATE
0	Blind	Inherited from infrastructure, unseen by the operator
1	Inventoried	Known and recorded, not changeable in practice
2	Abstracted	Behind an interface that admits substitution
3	Replaceable and owned	Substitution possible and under the operator's control
4	Drilled	The swap has been rehearsed in production

The depth reference for this scale is the crypto-agility maturity ladder of Series II, Paper 2 ("The Agility Question"). Within this model it is the measure of one dimension, not the whole.

4.3 The agility floor and the assurance result

The agent's **agility floor** is the minimum of its three agility ratings. Surface floors never average: an agent with a Drilled identity, a Blind channel and an Abstracted receipt has an agility floor of Blind.

The agent's **assurance result** is the agility floor together with the three current-posture verdicts. Version 1.0 does not combine them into a single number, because they measure different properties. An agent whose three surfaces run obsolete cryptography behind abstracted, operator-owned, drilled interfaces has an agility floor of 4 and three failing posture verdicts. An agent whose surfaces run hybrid post-quantum cryptography inherited from a provider has passing verdicts and an agility floor of 0. Both facts belong on the scorecard, and neither hides the other.

The two aggregations differ on purpose. Estates average because governance is a portfolio question. Agents floor because compromise is a weakest-link question.

5. The post-quantum dimension

Post-quantum readiness enters the model in both layers rather than as a separate track. Estate-side, it is scored by the communicate pillar's readiness control (COMM-04) and by the top rung of the cryptographic-anchoring control in the authenticate pillar (AUTH-04). Agent-side, it is the substance of the three posture verdicts and the agility scale. The model treats post-quantum migration as one expression of the same capability the agility scale measures: the ability to change cryptography faster than the threat changes.

6. Outputs: one pass, two artifacts

A complete assessment produces both layers from one pass and renders them twice:

- **The scorecard**, board-legible: the estate band with its gate, the pillar profile, the believed and discovered agent counts, the distribution of agent agility floors with the posture verdicts beside them, and the top five remediation moves in priority order.
- **The evidence map**, supervisor-legible: per control and per agent, what can be proved, from which record, on which regulatory clock. Every questionnaire control carries regulatory, standards and guidance mappings (EU AI Act, NIS2, DORA, NIST AI RMF, ISO/IEC 42001, ISO/IEC 27001, ANSSI guidance, NIST post-quantum migration guidance, OWASP LLM Top 10, ENISA, GDPR). The mapping gives the assessment traceability to existing duties and lets the underlying evidence be organised against them. The mapping is traceability. The evidence is the artifact itself: the configuration, the signed record, the policy, the inventory, the approval trail, the retained log.

7. Discovery, instruments, record format, conformance

Discovery precedes scoring. An assessment records its discovery scope (the platform consoles, pipelines, SaaS administration panels, workspaces and credential stores searched), the believed agent count before discovery, the discovered count after discovery, and the inclusion test applied to candidate agents. An assessment that begins from a supplied register without a discovery pass must identify itself as a **partial assessment** and cannot claim estate-level conformance.

Instruments. The model's two instruments are public and running: the estate questionnaire (v1.1, thirty-five controls with full rubrics) at assessment.mahasbini.org, and the per-agent inventory (Agent-CBOM, schema 0.1, CycloneDX-style references) at agentcbom.mahasbini.org.

Record. Results serialize to the Agent Governance Maturity Record, a versioned schema designed to support longitudinal comparison within an organisation and, once assessor and field calibration mature, cross-organisation benchmarking.

Conformance. An assessment conforms to this model when all of the following hold: the discovery scope, the believed and discovered counts and the inclusion test are recorded. Every control score maps to a rubric cell. Every inventoried agent carries three posture verdicts and three agility ratings. The agent's agility floor is the minimum of its three agility ratings. The estate score uses the published weights, the half-open band intervals and the pillar-floor gate. The remediation ordering is reproducible from the control scores and the published impact ratings.

8. Versioning

This specification is versioned with its instruments. Changes to controls, weights, impact ratings, bands, surfaces or the agility scale require a version bump of the model, the questionnaire, and the record schema together. Version 1.0 corresponds to questionnaire v1.1 and Agent-CBOM schema 0.1.

Weights and thresholds. The pillar weights, impact ratings and band thresholds are v1.0 normative choices, argued from failure-mode reasoning rather than calibrated from field data. The public instruments accumulate the distributions that will calibrate them. The pillar-floor gate bounds what a mis-weight can hide.

Version history. 1.0, September 2026: first published specification, aligned to the running instruments, with the inclusion test, the pillar-floor gate, the agility floor and the discovery requirement.

Appendix A. The thirty-five controls of v1.0

Control identifiers, names and impact ratings are normative for v1.0. The full rubrics and the regulatory, standards and guidance mappings live in questionnaire v1.1. This table fixes what the model contained at publication, independently of the live instrument.

ID	CONTROL	PILLAR	IMPACT
AUTH-01	Unique, enumerable identity per agent	Authenticate	5
AUTH-02	Identity lifecycle: provisioning, renewal, revocation	Authenticate	4
AUTH-03	Credential rotation on a defined schedule	Authenticate	5
AUTH-04	Cryptographic identity anchoring, with a post-quantum path	Authenticate	5
AUTH-05	Non-repudiable attribution of actions to an agent identity	Authenticate	4
AUTH-06	Current registry and inventory of all agents	Authenticate	4
AUTH-07	Named, accountable human owner per agent	Authenticate	4
AUTH-08	Environment isolation of agent identities	Authenticate	4
AUTH-09	Identity proofing at creation	Authenticate	4
AUTH-10	Tested offboarding of retired agents	Authenticate	4
AUTHZ-01	Least privilege on tools, APIs and data scopes	Authorize	5
AUTHZ-02	Explicit scope constraints on tool and API calls	Authorize	5
AUTHZ-03	Dynamic, context-aware authorization	Authorize	4
AUTHZ-04	Explicit, bounded delegation chains	Authorize	5
AUTHZ-05	Approval workflow before production access	Authorize	4
AUTHZ-06	Break-glass and override procedures	Authorize	3
AUTHZ-07	Just-in-time elevation of sensitive permissions	Authorize	4
AUTHZ-08	Policy-as-code, versioned, reviewed and tested	Authorize	4
AUTHZ-09	Usage audit against granted permissions	Authorize	4
AUTHZ-10	Separation of duties between execution and approval	Authorize	5
COMM-01	Mutual authentication of agent channels	Communicate	5
COMM-02	Message integrity proofs validated at the receiver	Communicate	4
COMM-03	Encryption at rest and in transit with managed keys	Communicate	4
COMM-04	Post-quantum migration path for agent channels	Communicate	5
COMM-05	Prompt-injection and instruction-smuggling defences	Communicate	5
COMM-06	Outbound data-loss prevention and content filtering	Communicate	4
COMM-07	Replay protection on agent protocols	Communicate	3
COMM-08	Integrity of prompts, context and memory	Communicate	4
COMM-09	Tamper-evident, queryable logs of communications and tool calls	Communicate	4
COMM-10	Hardened calls to third-party APIs and model providers	Communicate	4
META-01	Governance body with explicit authority over agents	Meta	4
META-02	Written, approved agent-specific policy	Meta	3

ID	CONTROL	PILLAR	IMPACT
META-03	Role-based training for developers, operators and approvers	Meta	4
META-04	Agent-aware incident response	Meta	4
META-05	Vendor and third-party due diligence on model and platform providers	Meta	4

About the author

Amin Hasbini is an AI and cybersecurity executive and former director of Kaspersky's Global Research and Analysis Team for the META region. He works on post-quantum maturity and AI agent security, and carries the post-quantum and crypto-agility roadmap of the European Champions Alliance's cybersecurity group. mahasbini.org